SPRINGER
REFERENCE

Claude Diebolt
Michael Haupert
Editors

Handbook of Cliometrics

Second Edition

Springer

Handbook of Cliometrics

Claude Diebolt • Michael Haupert
Editors

Handbook of Cliometrics

Second Edition

With 131 Figures and 73 Tables

Springer

Editors
Claude Diebolt
BETA/CNRS
University of Strasbourg, Institute for Advanced Study
Strasbourg, France

Michael Haupert
University of Wisconsin – La Crosse
La Crosse, WI, USA

ISBN 978-3-030-00180-3 ISBN 978-3-030-00181-0 (eBook)
ISBN 978-3-030-00182-7 (print and electronic bundle)
https://doi.org/10.1007/978-3-030-00181-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Introduction

Aims and Scope

The New Economic History (a term proposed by Jonathan Hughes) or Cliometrics (coined by Stan Reiter), meaning literally the measurement of history, is of very recent origin. Its first practitioners are considered to be Alfred Conrad and John Meyer, who published "Economic Theory, Statistical Inference, and Economic History" in the *Journal of Economic History* in 1957 after its presentation earlier that year at the joint meetings of the Economic History Association and the NBER Conference on Research in Income and Wealth. They followed that up in 1958 with a paper demonstrating the cliometric methodology as it applied to slavery in antebellum America. Robert Fogel's seminal research work on the impact of the railroad on American economic growth is in extension a true revolution in the history of economics, even a complete break with the tradition. It reestablished a role for history in economics by expressing it in the language of the discipline. Today, one can even say that it is an expanding domain in economics, contributing to new debates and challenging conventional wisdom. The use of econometric techniques and economic theory has contributed to the rejuvenation of economic history debates, made quantitative arguments unavoidable, and contributed to the emergence of a new historical awareness among economists.

Cliometrics does not concern economic history in the limited, technical meaning of the term. It modifies historical research in general. It represents the quantitative projection of social sciences in the past. The question of knowing whether slavery benefited the United States before the Civil War or if railways had substantial effects on the development of the US economy is as important for general history as for economic history and will necessarily weigh on any interpretation or appraisal (anthropological, legal, political, sociological, psychological, etc.) of the course of American history.

Furthermore, cliometrics challenges one of the basic hypotheses of the idealistic school: that history can never provide scientific proof because it is impossible to subject unique historical events to experimental analysis. On the contrary, cliometricians have shown that such experimentation is possible by construction of a

counterfactual that can be used to measure the deviation between what actually happened and what could have happened under different circumstances.

Robert Fogel famously used a counterfactual to measure the impact of the railroad on American economic growth. This methodological principle is perhaps, along with historical time series econometrics, the most important contribution of cliometrics for researchers in social science in general, and historians in particular.

The Methodological Features

Fogel defined the methodological features of cliometrics. He considered it fundamental that cliometrics should stress measurement while recognizing the existence of close links between measurement and theory. Indeed, unless it is accompanied by statistical and/or econometric processing and systematic quantitative analysis, measurement is just another form of narrative history. It is true that it replaces words with figures, but it does not bring in any new factors. In contrast, cliometrics is innovative when it is used to attempt to model all the explanations of past economic development. In other words, the main characteristic of cliometrics is the use of hypotheticodeductive models that call on the closest econometric techniques with the aim of establishing the interaction between variables in a given situation in mathematical form.

This generally consists of constructing a model – of general or partial equilibrium – that represents the various components of the economic evolution in question and showing the way in which they interact. Correlations and/or causalities can thus be established to measure the relative importance of each over a given period of time.

The final ingredient of the cliometric approach concerns the concepts of a market and price. Even in areas where there is no explicit market, the cliometric approach will often study the subject by analogy with the market concepts of supply, demand, and price.

So far, hypotheticodeductive models have mainly been used to determine the effects of innovations, institutions, and industrial processes on growth and economic development. As there are no records saying what would have happened if the innovations in question had not occurred or if the factors involved had not been present, this can only be found out by drawing up a hypothetical model used for deducing a hypothesized alternative situation – i.e., the counterfactual. It is true that the use of propositions contrasting with the facts is not new in itself. Such propositions are implicitly involved in a whole series of judgments, some economic and others not.

The use of such counterfactual analysis has not escaped criticism. Many researchers still believe that the use of hypotheses that cannot be verified generates quasihistory rather than history proper. Furthermore, the results obtained by the most elaborate cliometric applications have been less decisive than many cliometricians had hoped for. Critics are doubtless right to conclude that economic analysis in itself, with the use of econometric tools, is unable to provide causal explanations for the process and structure of change and development. There appear to be nonsystematic

breaks in normal economic life (wars, bad harvests, collective hysteria during market crashes, etc.) that require overall analysis but that are too frequently considered as extrinsic and abandoned to the benefit of an a priori formulation of theoretical suppositions.

Nevertheless, in spite of the disappointments resulting from some of its more extreme demonstrations, cliometrics also has its successes, together with continuous theoretical progress. The risk would obviously be that of allowing economic theory to neglect a whole body of empirical documentation that can enrich our knowledge about the reality of economic life. Conversely, theory can help to bring out certain constants, and only mastery of theory makes it possible to distinguish between the regular and the irregular and the foreseeable and the unforeseeable.

The Main Achievements

To date, the main achievements of cliometrics have been to slowly but surely establish, in the Fogel tradition, a solid set of economic analyses of historical evolution by means of measurement and theory and, following the path blazed by Douglass North, to recognize the limits of neoclassical theory and bring into economic models the important role of institutions. Indeed, this latter focus ultimately spawned a new branch of economics altogether, the new institutional economics. Nothing can now replace rigorous statistical and econometric analysis based on systematically ordered data. Impressionistic judgments supported by doubtful figures and fallacious methods and whose inadequacies are padded by subjective impressions have now lost all credibility. Economic history in particular should cease to be a "simple" story, illustrating with facts the material life during different periods, and become a systematic attempt to provide answers to specific questions. The ambition should be to move from the *Verstehen*, or understanding, to the *Erklären*, or explanation epistemology.

By extension, the more the quest for facts is dominated by the conception of the problems, the more research will address what forms the true function of economic history in the social sciences. This change of intellectual orientation, of cliometric reformulation, can thus reach other human and social science disciplines (law, sociology, political science, geography, etc.) and engender similar changes.

Indeed, the most vigorous new trend in the social sciences is without a doubt the preoccupation with quantitative and theoretical aspects. It is the feature that best distinguishes the concepts of the current generation of scholars from its forbears. Even the most literary of our colleagues is ready to agree to this. There is nothing surprising about this interest. One of the characteristic features of today's younger generation of scholars is most certainly that their intellectual training is much more deeply marked by science and the scientific spirit than that of the generations that preceded them. It is, therefore, not surprising that young scientists should have lost patience with regard to the tentative approach of traditional historiography and have sought to build their work on foundations that are less "artisanal."

Human and social sciences are thus becoming much more elaborate in the technical respect, and it is difficult to believe that a reversal of the trend is likely to occur. However, it is also clear that a significant proportion of human and social scientists have not yet accepted the new trends aimed at using more elaborate methodology and clear concepts conforming to new norms in order to develop, in a Fogelian tradition, a truly scientific human and social science.

A Branch of History?

For many authors – and many of its protagonists – cliometrics appears to be first of all a branch of history. Using economic tools, techniques, and theories, it provides answers to historical, rather than economic, debates per se.

The meaning of the word "empirical" for (American) economic historians has varied considerably with the passing of time. One can observe a shift from a concept of empirical fact as understood by the "classical historian" (for whom anything, as opposed to only quantitative data, retrieved from archives can be used in his demonstration) to one as understood by (applied) economists (the empirical aspect consists of analyzing numerical time series) and a convergence of theoretical viewpoints of historians and economists thanks to a common interest in the building of theories of development.

Here, Simon Kuznets seems to have played a key role by emphasizing the importance of performing at the onset a serious macroeconomic analysis of the major quantitative macrochanges in the past economic history before possibly identifying certain sectors that are deemed central for economic development. One should note that even in his concern to combine history with economic analysis, he thought of a theory of development that remained inductively based upon the observation of the major past evolution enlightened by the analysis of long-run time series patiently accumulated by the economic historian.

This (inductive) view is therefore intimately linked with the historical current in economics, the German Historical School, despite the use of more sophisticated techniques. It could be said that the two disciplines became closer but probably within the frame of "inductive" economics. On top of that, despite those early interests in building a kind of historically (i.e., inductively) grounded development economics, cliometrics mainly tried to provide answers to historiographical questions – and therefore spoke more to the historian than to the standard economist. Econometric techniques may be used with the reconstitution of time series and identification of missing figures by interpolation or extrapolation – something, by the way, that annoys professional historians. But these cliometric procedures have nonetheless a historical vocation – that of shedding light on historical questions – considering economic theory or econometrics as auxiliary disciplines of history. And when the cliometric approach was mobilized to build a development theory based upon clearly measured facts, it developed an economics more akin to the objectives of the German Historical School than one participating in the movement toward

highly abstract and deductive theory that characterized the development of the neoclassical school of the time.

The conflict between Kuznets and Walt Rostow regarding the stages in economic development was actually based upon the empirical foundations of Rostow's theory and not at all on a debate concerning the shortcomings of a very inductive and aggregate perspective lacking formal rigor (no use of growth theories) or micro-foundations, which would doubtless be the main subject of criticism today. In short, either cliometrics is still a (modernized) branch of (economic) history – in the same way as the modernization of methods in archaeology (from carbon-14 measurement to the use of statistical techniques such as discriminant analysis) does not turn the discipline into a branch of natural science – or the cliometric approach is mobilized to obtain theoretical results grounded more on induction from collected time series than from a deductive explicit modeling exercise, i.e., economic theory that must be primarily founded on facts and a generalization of empirical evidence. In this way, it contributes to an economic science that is more related to the German Historical School than to the neoclassical perspective.

An Auxiliary Discipline of Economics?

But this is not the end of the story. Some recent work in cliometrics performed by economists (stricto sensu) reveals the possibility of a cliometrics that could also be an auxiliary discipline of economics per se. As such, it should be part of the toolkit and competencies of all economists. However, as the term auxiliary discipline indicates, it could only fulfill its proper role for economics if it remains slightly (not too much) outside the realm of standard neoclassical economics. It must be a compound of the application of the newest econometric techniques and economic theory with the old institutional and factual culture characterizing the old economic history.

History is indeed always a discipline of synthesis. It should also be the case for cliometrics. If not, if cliometrics were to be deprived of all its "historical dimensions," it would simply cease to exist (it would only be economics applied to the past or mere retrospective econometric exercises). To be helpful for the economics profession at large, its main job should be to mobilize all the relevant information that can be gathered from history to enrich or even challenge economic theory (or theories). And this relevant information should also include cultural or institutional development, provided that they can be properly presented as useful for the profession.

A conventional belief among economists (in fact that of Lord Kelvin) is that "qualitative is poor quantitative." But could it not be possible that "quantitative is poor qualitative" might also sometimes be true? A big difference between economists and historians is the sense of so-called historical criticism and the desire to avoid any anachronism. In addition to close examination of the historical sources, this involves the close examination of the institutional, social, and cultural context that forms the framework constraining the players' behavior. It is true that the (new)

economic history will not build a general theory – it shares too strongly the belief in the necessity of examining economic phenomena in their context – but it could suggest a few useful ideas and insights, based upon solid investigations and correctly estimated stylized facts, to economists who are attempting to develop laws of economic behavior (unlike history, economics is still a nomological science). Economists and cliometricians can also cooperate and jointly author research. This is a view shared by Daron Acemoglu, Simon Johnson, James Robinson, and Oded Galor, among others, trying to use the material derived from traditional history to build new ideas useful for economic theorists.

In summary, it could be contended that a good cliometric practice is not an easy exercise. Becoming too narrowly "economic," it would not be possible for cliometrics to answer certain questions that would require, for example, more information about the microstructure of financial markets or the actual functioning of stock exchanges during the period under scrutiny – it would only measure phenomena that it cannot explain. It would require the specific approach (and extraneous information) of the historian to describe the reasons for the lack of relevance (or understand the shortcoming) of such an economic theory in a given context (precise place and period). It is perhaps only in this regard that cliometrics can provide something for economists by suggesting lines of research. However, if it became too "historical," cliometrics would cease to appeal to the economics profession. Economists need new economic historians aware of their debates and their interests.

A Full-Fledged Field of Economic Theory?

Last, but not least, cliometrics could one day be more than just an ancillary discipline of economics and instead become a full-fledged field of economic theory. There is indeed another possibility: viewing cliometrics as the science of the emergence of institutional and organizational structures, and that of path dependence. Economic history would use the old techniques of the discipline coupled with the state-of-the-art arsenal of econometrics in order to reveal stylized facts about the efficiency of various institutional arrangements as well as the causes and consequences of institutional change. It would help the theorist in developing a true theory of institutional change, i.e., one that at the same time would be general (serving the needs of policymakers today, for example) and theoretically solid (grounded on economic principles) while solidly grounded on empirical regularities as put forward by a joint economic and historical analysis. This analysis of institutional morphogenesis would be the true theoretical part of a cliometric science that would emancipate itself from its apparently purely empirical fate – being the playing ground of long-run econometricians. It is clear that economists' desire for generality and their fascination for the mathematical science do not encourage them to pay too much attention to contextualization. However, neoinstitutionalist economists like North warn us to seriously consider institutional (including cultural) contexts.

Our ambition for the *Handbook of Cliometrics* was thus also aimed at encouraging economists to examine more systematically these theories grounded upon history and nevertheless aiming at the determining general laws on the creation of institutions or of institutional changes. Beyond the study of long-run quantitative data sets, a branch of cliometrics is more and more focused on the role and evolution of institutions by aiming at combining the economist's desire for generality with the concern for the precise context in which economic players act that characterizes historians and other social scientists. This middle road between pure empiricism and disincarnate theory might perhaps open the door to a better economic theory. This will enable economists to interpret current economic issues in the light of the past and, in so doing, understand more deeply the historical working of economies and societies. This is the path to offering better policy advice for today.

The Contents

When putting together the first edition of this handbook, the most difficult question we faced was what to include. The possibilities were endless, but the space was limited. For this second edition we were allocated much more space, and the results are obvious: we have tripled the size of the original handbook, adding 43 new chapters to the original 22, of which several were revised and updated by their original authors. Even with this expanded coverage, there are still techniques and topics of importance that are not included. Their absence should not by any means be considered as an indictment of their importance or historical significance. In some cases, we had chapters committed, but for a variety of reasons the author was unable to meet our publication deadline. In this situation the chapters will be added to the online version of the handbook, which can be found at: https://link.springer.com/referencework/10.1007/978-3-642-40458-0

The expanded selection of chapters in this edition still represents only a sampling of the topics that cliometrics has helped to transform over the past half century. It includes topics that established cliometrics as the "new" economic history in the 1960s, including Richard Sutch's chapter on slavery and Jeremy Atack's contribution on railroads. It features chapters that have long been at the center of cliometric analysis, such as Greg Clark's chapter on the industrial revolution, Larry Neal's chapter on financial markets, and the contribution by Chris Hanes on the Great Depression. And we offer chapters on narrower topics that have been developed largely as a result of the cliometric approach, such as the age heaping work discussed by Franziska Tollnek and Joerg Baten, Douglas Puffert's chapter on path dependence, Thomas Rahlf's contribution on statistical inference, and Florian Ploeckl's chapter on spatial modeling. In between, we have included articles by Stanley Engerman, Deirdre McCloskey, Roger Ransom, and Peter Temin, who began plying their trade when cliometrics really was the "new" way of studying economic history, and young scholars who represent the next generation of cliometricians, including Matthew Jaremski and Chris Vickers. The common link running throughout the Handbook is the focus on the contributions of cliometrics.

The *Handbook of Cliometrics* is a milestone in the field of historical economics and econometric history through its emphasis on the concrete contribution of cliometrics to our knowledge in the fields of economics and history. It is a work of tertiary literature. As such, it contains digested knowledge in an easily accessible format. The articles are not original research or review articles, but rather an overview of the contributions of cliometrics to the topic of discussion. The articles stress the usefulness of cliometrics for economists, historians, and social scientists in general. The Handbook offers a wide range of topical coverage, with each article providing an overview of the contributions of cliometrics to a particular topic.

The book is organized into eight parts, grouping the 65 contributions by general topic, starting with six chapters on the history of economic history and cliometrics, and the contributions of its two most prominent practitioners, Robert Fogel and Douglass North. The second part focuses on human capital, with nine chapters ranging from broad topical coverage of labor markets, education and gender, to two specific surveys of the role of clio in age heaping and church book registry.

Part III takes the big picture into consideration with nine papers on economic growth. We feature chapters on industrial growth, the industrial revolution, antebellum growth, trade, market integration, and economic-demographic interactions, to name but a handful. Part IV covers institutions, both broadly defined (institutions, political economy, property rights, merchant empires) and more narrowly focused (slavery, colonial America, water rights).

The largest section is Part V, which features a dozen articles on money, banking and finance in a variety of formats. Early capital markets, origins of the U.S. financial system, and antebellum banking lead off the coverage, followed by big picture coverage of financial markets and systems, panics, and interest rates. Also included are chapters on the Great Depression, central banking, sovereign debt, and corporate governance.

In Part VI we offer eight chapters on the topics of government, health, and welfare. Clio offspring are featured here, including anthropometrics and agricliometrics. Chapters on income inequality, nutrition, health care, war, and the role of the government in the Great Depression can also be found. Part VII covers innovation both mechanical and creative, railroads, transportation, and tourism.

The Handbook concludes with a section on technique and measurement, two hallmarks of cliometrics. Here you will find chapters on analytic narratives, path dependence, spatial modeling, and statistical inference alongside offerings on African economic history, output measurement, and the census of manufactures.

We enjoyed the process of putting this second edition together. What began nearly a decade ago as an innocent query (Why isn't there a handbook of cliometrics?) has now grown to two editions with more than five dozen entries. The process was a labor of love and the result is an assemblage of top scholars analyzing the role that cliometrics has played in the advancement of knowledge across a wide array of topics. We present it to you, and dedicate it to all cliometricians, past, present, and future.

Claude Diebolt
Michael Haupert

References

Acemoglu D, Johnson S, Robinson J (2005) Institutions as a fundamental cause of long-run growth, Chapter 6. In: Aghion P, Durlauf S (eds) Handbook of economic growth, 1st edn, vol 1. North-Holland, Amsterdam, pp 385–472. ISBN 978-0-444-52041-8

Conrad A, Meyer J (1957) Economic theory, statistical inference and economic history. J Econ Hist 17:524–544

Conrad A, Meyer J (1958) The economics of slavery in the Ante Bellum South. J Polit Econ 66:95–130

Carlos A (2010) Reflection on reflections: review essay on reflections on the cliometric revolution: conversations with economic historians. Cliometrica 4:97–111

Costa D, Demeulemeester J-L, Diebolt C (2007) What is 'Cliometrica'. Cliometrica 1:1–6

Crafts N (1987) Cliometrics, 1971–1986: a survey. J Appl Econ 2:171–192

Demeulemeester J-L, Diebolt C (2007) How much could economics gain from history: the contribution of cliometrics. Cliometrica 1:7–17

Diebolt C (2012) The cliometric voice. Hist Econ Ideas 20:51–61

Diebolt C (2016) Cliometrica after 10 years: definition and principles of cliometric research. Cliometrica, 10:1–4

Diebolt C, Haupert M (2018) A cliometric counterfactual: what if there had been neither Fogel nor North? Cliometrica, 12:407–434

Fogel R (1964) Railroads and American economic growth: essays in econometric history. The Johns Hopkins University Press, Baltimore

Fogel R (1994) Economic growth, population theory, and physiology: the bearing of long-term processes on the making of economic policy. Am Econ Rev 84:369–395

Fogel R, Engerman S (1974) Time on the cross: the economics of American Negro Slavery. Little, Brown, Boston

Galor O (2012) The demographic transition: causes and consequences. Cliometrica 6:1–28

Goldin C (1995) Cliometrics and the Nobel. J Econ Perspect 9:191–208

Kuznets S (1966) Modern economic growth: rate, structure and spread. Yale University Press, New Haven

Lyons JS, Cain LP, Williamson SH (2008) Reflections on the cliometrics revolution. Conversations with economic historians. Routledge, London

McCloskey D (1976) Does the past have useful economics? J Econ Lit 14:434–461

McCloskey D (1987) Econometric history. Macmillan, London

Meyer J (1997) Notes on cliometrics' fortieth. Am Econ Rev 87:409–411

North D (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge

North D (1994) Economic performance through time. Am Econ Rev 84 (1994):359–368

Piketty T (2014) Capital in the twenty-first century. The Belknap Press of Harvard University Press, Cambridge, MA
Rostow WW (1960) The stages of economic growth: a non-communist manifesto. Cambridge University Press, Cambridge
Temin P (ed) (1973) New economic history. Penguin Books, Harmondsworth
Williamson J (1974) Late nineteenth-century American development: a general equilibrium history. Cambridge University Press, London
Wright G (1971) Econometric studies of history. In: Intriligator M (ed) Frontiers of quantitative economics. North-Holland, Amsterdam, pp 412–459

Preface to Second Edition

Welcome to the second edition of the *Handbook of Cliometrics*, a part of the Springer Reference Library. This second edition builds on the first, published in 2016, by including the original 22 chapters and adding 43 new chapters. In order to foster world-class research, this edition, like its predecessor, includes economists and economic historians of the highest caliber from around the world. The handbook impacts historical economics and econometric history through its emphasis on the concrete contribution of cliometrics to our knowledge of economics and history.

Cliometrics dates formally to the joint meeting of the Economic History Association and the Conference on Research in Income and Wealth (under the purview of the NBER) in 1957. The concept of cliometrics, the application of economic theory and quantitative techniques to the study of history, is somewhat older. Regardless of its precise origin, this focus on the use of theory and formal modeling that distinguishes cliometrics from "old" economic history has redefined the discipline and made an indelible mark on economics. The works in this handbook recognize these contributions and highlight them in a variety of subdisciplines.

The handbook is a work of tertiary literature. As such, it contains digested knowledge in an easily accessible format. The chapters provide an overview of the contributions of cliometrics to various subdisciplines in the field of economic history. Each one stresses the usefulness of cliometrics for economists, historians, and social scientists in general.

A project of this size and scope does not come to a successful conclusion without the contributions of many people. We want to thank all those who helped to bring this idea to fruition and see it through to its conclusion. First and foremost, we thank our authors, who have produced articles of the highest quality under demanding deadlines and through numerous drafts. Their time and expertise are what elevates this handbook to the highest level. We also need to thank the editorial and production team, who turned this work from concept to final printed and online product: Martina Bihn, who nourished our idea from the beginning, and Shruti Datt and Rebecca Urban, our editors, who kept us on task and provided copious and valuable advice at every turn. Many thanks also to Michael Hermann for his unconditional support. We would also like to thank the Board of Trustees of the Cliometric Society, who inspired us to carry on with our initial proposal to create a handbook and cheered us on when we added a second edition.

Finally, we would be remiss if we did not thank our spouses, Valérie and Mary Ellen, who put up with late nights over computers, long days at the office, and our general demeanor as we stared down deadlines, all while working at their careers while tolerating our obsession.

May 2019

Claude Diebolt
Michael Haupert

Preface to First Edition

Welcome to the *Handbook of Cliometrics*, a part of the Springer Reference Library. In order to foster world-class research, the handbook includes economists and economic historians of the highest caliber from around the world. It is a milestone in the field of historical economics and econometric history through its emphasis on the concrete contribution of cliometrics to our knowledge of economics and history. Cliometrics dates formally to the joint meeting of the Economic History Association and the Conference on Research in Income and Wealth (under the purview of the NBER) in 1957. The concept of cliometrics, the application of economic theory and quantitative techniques to the study of history, is somewhat older. Regardless of its precise origin, this focus on the use of theory and formal modeling that distinguishes cliometrics from "old" economic history has redefined the discipline and made an indelible mark on economics. The works in this handbook recognize these contributions and highlight them in a variety of subdisciplines. The handbook is a work of tertiary literature. As such, it contains digested knowledge in an easily accessible format. The chapters provide an overview of the contributions of cliometrics to various subdisciplines in the field of economic history. Each one stresses the usefulness of cliometrics for economists, historians, and social scientists in general.
A project of this size and scope does not come to a successful conclusion without the contributions of many people. We want to thank all those who helped to bring this idea to fruition and see it through to its conclusion. First and foremost, we thank our authors, who have produced articles of the highest quality under demanding deadlines and through numerous drafts. Their time and expertise are what elevates this handbook to the highest level. We also need to thank the editorial and production team, who turned this work from concept to final printed and online product: Martina Bihn, who nourished our idea from the beginning and guided us through the long process from idea to output, and Karin Bartsch, our primary editor, who kept us on task and provided copious and valuable advice at every turn. Many thanks also to Michael Hermann and Nicholas Philipson for their unconditional support. We would also like to thank the Board of Trustees of the Cliometric Society, who inspired us to carry on with our initial proposal to create a handbook. Finally, we would be remiss if we did not thank our spouses, Valérie and Mary Ellen, who put up with late nights

over computers, long days at the office, and our general demeanor as we stared down deadlines, all while working at their careers while tolerating our obsession.

January 2015

Claude Diebolt
Michael Haupert

Contents

Volume 1

About the Editors

Claude Diebolt is CNRS research professor of economics at the University of Strasbourg. He is the founder and the editor in chief of the journal *Cliometrica*. He is also the coeditor of the *Handbook of Cliometrics* and was the organizer of the 8th World Congress of Cliometrics. Claude Diebolt is the current president of the *Comité National de la Recherche Scientifique* (CNRS Section 37, Economics and Management Science) and the president of the *Association Française de Science Economique*. He is the founding president of the *Association Française de Cliométrie* and a former chair of the board of trustees of the *Cliometric Society* in the United States. On March 6, 2019, Claude Diebolt received the 2018–2019 Sarton Medal for his research in Cliometrics.

Michael Haupert is professor of economics at the University of Wisconsin-La Crosse and executive director of the Economic History Association. Previously, he served in the same position for 8 years for the Cliometric Society. His research interests are the economic history of the sports and entertainment industries and the history of the economic history discipline. He has authored two books on the history of the entertainment industry in America and coauthored a book on the economics of baseball in the 1920s. His research has also appeared in *Cliometrica*, *The Journal of Economic History*, and *The Journal of Money, Credit and Banking*, among others.

Contributors

Guido Alfani Dondena Centre and IGIER, Bocconi University, Milan, Italy

Jeremy Atack Vanderbilt University, Nashville, TN, USA
NBER, Cambridge, MA, USA

Joerg Baten University of Tuebingen and CESifo, Tuebingen, Germany

Sascha O. Becker University of Warwick, Coventry, UK
CAGE, Coventry, UK
CEPR, London, UK
CESifo, Munich, Germany
IZA, Bonn, Germany
ifo, Munich, Germany
ROA, Maastricht, Netherlands

Dan Bogart Department of Economics, University of California, Irvine, CA, USA

Karol Jan Borowiecki Department of Business and Economics, University of Southern Denmark, Odense, Denmark

Carsten Burhop University of Bonn, Bonn, Germany

Joyce Burnette Department of Economics, Wabash College, Crawfordsville, IN, USA

Ann M. Carlos Department of Economics, University of Colorado Boulder, Boulder, CO, USA

Gregory Clark University of California, Davis, CA, USA

Nicholas Crafts CAGE, University of Warwick, Coventry, UK

Lee A. Craig Department of Economics, North Carolina State University, Raleigh, NC, USA

Alexandra M. de Pleijt University of Oxford, Oxford, UK

Claude Diebolt BETA/CNRS, University of Strasbourg Institute for Advanced Study, Strasbourg, France

Mauricio Drelichman The University of British Columbia, Vancouver, Canada

Brandon Dupont Western Washington University, Bellingham, WA, USA

Jari Eloranta University of Helsinki, Helsinki, Finland

Stanley L. Engerman Department of Economics, University of Rochester, Rochester, NY, USA

Giovanni Federico Department of Economy and Management, University of Pisa, Pisa, Italy

CEPR, London, UK

Emanuele Felice Dipartimento di Scienze Filosofiche, Pedagogiche ed Economico-Quantitative, Università "G. D'Annunzio" Chieti-Pescara, Pescara, Italy

Alexander J. Field Department of Economics, Santa Clara University, Santa Clara, CA, USA

Price Fishback Economics Department, University of Arizona, Tucson, AZ, USA

Caroline Fohlin Johns Hopkins University, Baltimore, MD, USA

Emory University, Atlanta, GA, USA

James Foreman-Peck Cardiff University, Cardiff, UK

Roger Fouquet Grantham Research Institute of Climate Change and the Environment, London School of Economics and Political Science (LSE), London, UK

Johan Fourie LEAP, Department of Economics, Stellenbosch University, Stellenbosch, South Africa

Stellenbosch University, Stellenbosch, South Africa

Claudia Goldin Department of Economics, Harvard University and National Bureau of Economic Research, Cambridge, MA, USA

Diana Seave Greenwald National Gallery of Art, Washington, DC, USA

Christopher Hanes Department of Economics, State University of New York at Binghamton, Binghamton, NY, USA

Zeynep K. Hansen Department of Econimics, Boise State University, Boise, ID, USA

Timothy J. Hatton Department of Economics, University of Essex, Colchester, UK

Research School of Economics, Australian National University, Canberra, Australia

Michael Haupert University of Wisconsin – La Crosse, La Crosse, WI, USA

Ralph Hippe European Commission, Joint Research Centre (JRC), Seville, Spain

Philip T. Hoffman California Institute of Technology (CalTech), Pasadena, CA, USA

John A. James Department of Economics, University of Virginia, Charlottesville, VA, USA

Matthew Jaremski Colgate University and NBER, New York, USA

Ian Keay Department of Economics, Queen's University, Kingston, ON, Canada

Mark Koyama Department of Economics, George Mason University, Fairfax, VA, USA

Sumner La Croix Department of Economics, University of Hawai'i-Mānoa, Honolulu, HI, USA

Markus Lampe Vienna University of Economics and Business, Vienna, Austria

Charlotte Le Chapelain Centre Lyonnais d'Histoire du Droit et de la Pensée Politique, Université de Lyon, Lyon, France

Frank D. Lewis Department of Economics, Queen's University, Kingston, ON, Canada

Gary D. Libecap National Bureau of Economic Research, University of California, Santa Barbara, CA, USA

Hoover Institution, Stanford University, Stanford, CA, USA

Scott E. Lowe Department of Econimics, Boise State University, Boise, ID, USA

Robert A. Margo Boston University and National Bureau of Economic Research, Boston, MA, USA

Deirdre Nansen McCloskey University of Illinois at Chicago, Chicago, IL, USA

Terence C. Mills School of Business and Economics, Loughborough University, Loughborough, UK

David Mitch Department of Economics, University of Maryland, Baltimore County, Baltimore, MD, USA

Jon Moen Department of Economics, The University of Mississippi, Oxford, MS, USA

Philippe Mongin GREGHEC, Economics and Decision Sciences, CNRS & HEC Paris, Jouy-en-Josas, France

Eric Monnet Banque de France, Paris School of Economics and CEPR, Paris, France

Larry Neal Department of Economics, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Gregory T. Niemesh Department of Economics, Miami University, Oxford, OH, USA

NBER, Cambridge, MA, USA

Nonso Obikili LEAP, Department of Economics, Stellenbosch University, Stellenbosch, South Africa

Faustine Perrin BETA/CNRS, University of Strasbourg Institute for Advanced Study, Strasbourg, France

Vicente Pinilla Department of Applied Economics, Faculty of Economics and Business Studies, Universidad de Zaragoza and Instituto Agroalimentario de Aragon -IA2- (Universidad de Zaragoza-CITA), Zaragoza, Spain

Florian Ploeckl School of Economics, The University of Adelaide, Adelaide, SA, Australia

Douglas J. Puffert LCC International University, Klaipeda, Lithuania

Stephen Quinn Department of Economics, Texas Christian University, Fort Worth, TX, USA

Thomas Rahlf German Research Foundation, Bonn, Germany

Roger Ransom University of California, Riverside, CA, USA

Claudia Rei University of Warwick, Coventry, UK

Hugh Rockoff Department of Economics, Rutgers University, New Brunswick, NJ, USA

Nathan Rosenberg Department of Economics, Stanford University, Emeritus, Stanford, CA, USA

Joshua L. Rosenbloom Department of Economics, Iowa State University, Ames, IA, USA

NBER, Cambridge, MA, USA

Paul Sharp University of Southern Denmark, Odense M, Denmark

Richard H. Steckel Ohio State University, Columbus, OH, USA

Jochen Streb Abteilung Volkswirtschaftslehre, Lehrstuhl für Wirtschaftsgeschichte, Universität Mannheim, Mannheim, Germany

Richard Sutch Economics Department, University of California, Riverside, CA, USA

National Bureau of Economic Research, Cambridge, MA, USA

Richard Sylla Stern School of Business, New York University, New York, NY, USA

Peter Temin Department of Economics, Massachusetts Institute of Technology, Cambridge, MA, USA

Melissa A. Thomasson Department of Economics, Miami University, Oxford, OH, USA

NBER, Cambridge, MA, USA

Franziska Tollnek University of Tuebingen, Tuebingen, Germany

Jan Luiten van Zanden Department of History and Art History – Economic and Social History, Utrecht University, Utrecht, The Netherlands

Chris Vickers Department of Economics, Auburn University, Auburn, AL, USA

Zachary Ward Department of Economics, Hankamer School of Business, Baylor University, Waco, TX, USA

Jacob Weisdorf University of Southern Denmark and CEPR, Odense M, Denmark

Thomas Weiss University of Kansas, Lawrence, KS, USA

Ludger Woessmann University of Munich and ifo Institute, Munich, Germany

CESifo, Munich, Germany

IZA, Bonn, Germany

CAGE, Coventry, UK

ROA, Maastricht, Netherlands

Gavin Wright Stanford University, Stanford, CA, USA

Robert E. Wright Social Science, Augustana University, Sioux Falls, SD, USA

Nicolas L. Ziebarth Department of Economics, Auburn University and NBER, Auburn, AL, USA

Part I

History

History of Cliometrics

Michael Haupert

Contents

Abstract

Economic historians have contributed to the development of economics by combining theory with quantitative methods, constructing and revising databases, discovering and creating new ones entirely, and adding the variable of time to traditional economic theories. This has made it possible to question and reassess earlier findings, thus increasing our knowledge, refining earlier conclusions, and correcting mistakes. It has contributed greatly to our understanding of economic growth and development. The use of history as a crucible to examine economic theory has deepened our knowledge of how, why, and when economic change occurs. The focus of this essay is to detail the history of the discipline of cliometrics, the quantitative study of economic history, and outline its evolution within the discipline of economic history.

M. Haupert (✉)
University of Wisconsin – La Crosse, La Crosse, WI, USA
e-mail: mhaupert@uwlax.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_2

Keywords
Business history · Cliometrics · Economic history · Economic thought · New economic history

Introduction

Economic historians have contributed to the development of economics by combining theory with quantitative methods, constructing and revising databases, discovering and creating new ones entirely, and adding the variable of time to traditional economic theories. This has made it possible to question and reassess earlier findings, thus increasing our knowledge, refining earlier conclusions, and correcting mistakes. It has contributed greatly to our understanding of economic growth and development.[1] The use of history as a crucible to examine economic theory has deepened our knowledge of how, why, and when economic change occurs.

In December of 1960, the "Purdue Conference on the Application of Economic Theory and Quantitative Techniques to Problems of History" was held on the campus of Purdue University.[2] It is recognized as the first meeting of what is now known as the Cliometric Society.[3] While it was the first formal meeting of a group of like-minded applicants of economic theory and quantitative methods to the study of economic history, it was not the first time such a concept had been broached, practiced, or even mentioned in the literature.[4] Cliometrics was a long time in coming, but when it arrived, it eventually overran the approach to the discipline of economic history, leading to a bifurcation of the economists and historians who practice the art and the blurring of the distinction between cliometricians (i.e., economic historians) and theorists who use historical data.

Before there was a Cliometric Society, there was the *Economic History Association* (*EHA*). And before the *EHA*, there were a number of societies that American economic historians could join, but none that they could really call their own. The closest thing they had was the *Economic History Society*, founded in 1926 and headquartered in the UK. In the USA, economic historians spread themselves out among a variety of associations according to their primary historical interests, such as the *Agricultural History Society* (founded in 1916), the *American Historical Association* (1884), the *Business Historical Society* (1926), and the *American Economic Association* (1885). None of these precisely fit the bill, however. As a

[1] See Drukker (2006), for example.

[2] A selection of the papers presented in these early meetings was published by Purdue University in 1967.

[3] The Cliometric Society was formally organized in 1983 by Sam Williamson and Deirdre (nee Donald) McCloskey.

[4] The first use of the term in print: "the logical structure necessary to make historical reconstructions from the surviving debris of past economic life essentially involves ideas of history, economics and statistics . . . has been labeled "Cliometrics" (Davis et al. 1960, p. 540).

result, a movement began in early 1937 to establish an American organization that was dedicated to the study and teaching of economic history. Actually, two different organizations were formed to meet these goals: the *Industrial History Society*, organized in 1939, followed by the *EHA* 1 year later.

What makes economic historians unique is not their use of historical data or their focus on the past but that they study the growth and evolution of economies over the long term. In this way, economic history's closest kin is development economics. In addition, the attention that economic historians give to noneconomic factors, such as legal and political systems, distinguishes them from economic theorists. Given the longer time span economic historians consider, doing so gives fuller attention to changes in institutions.[5]

Clio's roots are historical in nature, and its focus on theory has actually come full circle over the last century and a half. A mathematical movement in the economics discipline, advanced computing technology, and a shift in the focus of the role of history within economics all contributed to the proliferation of the "new" economic history that rewrote the landscape of the discipline. The emphasis on theory and formal modeling that distinguishes cliometrics from the "old" economic history now blurs the distinction between economic history and economic theory, to the extent that the need for economic historians is questioned and indeed no longer considered necessary in many economics departments.[6] The focus of this essay is to detail the history of the discipline of cliometrics, the quantitative study of economic history, and outline its evolution within the discipline of economic history.

Cliometrics

Cliometrics has been defined and summarized in numerous scholarly articles.[7] They all pretty much start with the obvious, that cliometrics is the application of economic theory and quantitative techniques to study history, and then move on to the origin of the name, the joining of *Clio* (the muse of history), with *metrics* ("to measure," or "the art of measurement"), allegedly coined by economist Stanley Reiter while collaborating with economic historians Lance Davis and Jonathan Hughes.[8] From there they recount the evolution of the discipline, highlight its major contributions, and mention its detractors. While some of that ground will be retread here, the focus of this essay is less about adding another history of cliometrics[9] and more about

[5]See Goldin (1995), Mitch (2011), and Tawney (1933) for discussions of the role of economic historians.

[6]Temin (2014)

[7]See, for example, Engerman (1996), Floud (1991), Lyons et al. (2008), Williamson (1991, 1994), and Williamson and Whaples (2003).

[8]Williamson and Whaples (2003), p. 446

[9]See Carlos (2010), Coats (1980), Crafts (1987), Fenoaltea (1973), Greif (1997), Lamoreaux (1998), Libecap (1997), Meyer (1997), and North (1997) for an overview of the evolution of cliometrics.

highlighting the literature in the history of the discipline that is cliometrics and from whence clio came.

de Rouvray (2004a, b, 2014), in her research on US economic history, describes the discipline as one aimed at understanding the origin, dynamics, and consequences of past economic events. She categorizes cliometrics as a movement that transformed that study from a narrative to a mathematical format. Her definition is not unique, but her attention to the historical detail which begat the cliometric revolution is without equal.

The origin of cliometrics can be found in the origin of economic history, which evolved as a separate discipline in Germany and England in the late nineteenth century. It migrated to the USA in 1892 in the person of W. J. Ashley and ultimately flourished. It was neither a rapid nor accepted emergence, however.

Cliometrics today is closely related to, but not necessarily the same thing as its progenitor, economic history. While there is considerable overlap between the membership of the Cliometric Society and its American brethren, the *Economic History Association*, the latter has many more members who reside in history departments than does the Cliometric Society. Indeed, one of the great criticisms of the cliometric movement is the wedge that it has driven between the practitioners of economic history in history and economics departments (Boldizzoni 2011)[10] due to its focus on quantitative measures and neoclassical theory.[11]

Despite the current strains, cliometrics does owe its very existence to economic history, having grown out of that discipline in the last half of the twentieth century. The skills of a cliometrician include those of any other economic historian. In his inaugural presidential address to the *EHA*, Edwin Gay (1941) noted that economic historians required two sets of skills, which they needed to wed in order to accomplish their task. He believed the molding of the skills of the economist and historian was essential, but not easy to accomplish.[12] That has not changed over the past three quarters of a century. What has changed is the degree to which those economic skills have become more formalized and technically demanding.

The clash between cliometricians and historians today is not all that different from the clash between economists and historians that began in the nineteenth century. Carl Menger (1884) compared historians to foreign conquerors, complaining that they were forcing their terminology and methods on economists. Half a century later, Ashton (1946) accused those who objected to the idea that economic theory should be applied to history of not truly understanding the nature of economics.

[10]For earlier laments about the encroachment of theory and mathematics on the study of history, see Braudel (1949) and Polanyi (1944).

[11]Perhaps more than anyone, D.N. McCloskey has been responsible for holding all economists, not just economic historians, accountable for moving the frontiers of knowledge forward and not simply using the latest techniques to measure something because it can be measured. For example, see McCloskey (1978, 1985, 1987, 2006).

[12]See also Ashley (1927), Ashton (1946), Gallman (1965), McCloskey (1986), and Nef (1941) for viewpoints of the melding of the skills of historians and economists.

While economic history had been dominated by qualitative studies, cliometrics was not the first application of quantitative methods to the discipline. As early as the seventeenth century, scholars attempted to infer an explanation of some aspects of economic history by examining data (D'Avenant 1699; Graunt 1662). In 1707, Bishop William Fleetwood wrote *Chronicon Preciosum*, a precursor of what a good cliometric article would become. He used archival records of prices and wages to measure the decline in the value of money over time. Uncharacteristic of a typical cliometric argument, however, his research was conducted in an effort to protect his Cambridge fellowship.

In fact, the discipline originated largely as a revolt against classical theory, and in its early years, it shunned the use of statistical techniques. By the 1920s, the attitude toward theory and statistics began to soften. Cliometrics is the continuation of this theoretical-quantitative tradition now nearly a century old and fortified by advances in economic theory, the melding of economics with approaches from other disciplines, and the growth of computing power. The latter has had profound impacts on the ability to analyze and disseminate data.

The Economic History Discipline

Economic history as a formal discipline dates only to the late nineteenth century, though books on topics considered to be economic history existed well before this. Harte (1971) noted the existence of historical treatments of economic problems as early as the seventeenth century, regarding macroeconomic issues created by the fashion for "political arithmetic." Among the earliest works recognized today as economic history were by Sir William Temple (1672) and John Evelyn (1674). Both were written to address concerns over contemporary international political and economic rivalries.

In the UK, there were antecedents to the English historical economists. As early as the 1850s, Richard Jones, who taught political economy at Haileybury, was calling for greater attention to historical context in which economic activity took place. And in the following generation, John Kells Ingram and T. E. Cliffe Leslie, both in Ireland, were distinguished advocates of a more historical approach to economics.

Before economic history, there were political economics departments and history departments, and neither was a natural home for economic history. Political economics departments tended not to focus on history. And as Cole (1968) discusses in his overview of economic history in America, the general approach by scholars trained in history departments in the nineteenth century was to consider economic factors as only one cause of change and not always necessarily the most important one.

The first formal organization of economic history as an academic discipline appeared in Germany in the mid-nineteenth century. In part, this was the result of German interest in establishing the most appropriate economic policies to be followed by the developing states of that time. In turn, it became an academic

discipline in the UK at the end of that century largely as a result of social concern over the poverty of the urban industrial working class.[13]

In Germany, the approach to economics was altered by the publication of Wilhelm Roscher's *Grundriss* (1843). Roscher was a historical economist. Along with Friedrich List, Bruno Hildebrand, and Karl Knies, and later followed by Gustav Schmoller, they focused on economic activities and institutions in the past as well as in the present. Before the end of the century, they published much of their economic history research relating to England, though little of it was ever translated into English.[14]

The earliest form of economic history was narration fortified with the occasional bit of quantitative data. When formal economic history began to evolve in Germany and England in the late nineteenth century, however, leading scholars such as Schmoller in Germany and Sir John Clapham in England sought to develop it independent of standard economic theory. Clapham (1929) argued that the central problems of economic theory, though stated in terms of a particular historical phase, were in essence independent of history. With few exceptions, this general view permeated the writing of economic history for more than half a century. Data were only occasionally collected, and when they were, they were seldom manipulated or used to test mathematical propositions, and economic models were practically unknown.

By the 1870s, political economy had devolved into a methodological debate (Methodenstreit) about whether economics should be inductive (develop theories providing evidence of the truth) or deductive (gather facts leading to a certain conclusion). Three developments coming out of this debate helped pave the way for historical economics: political economy begot economics, which was less simplistic; the thorough investigation of social problems and their origins fostered an interest in the origins of economic-based issues as well; and the ideas of evolution generated by the explosion of "historical" natural sciences (think Darwinism).

Economic history emerged as a distinct discipline during the course of the revolt against the deductive theories of classical economics. Led by Roscher, Knies, Hildebrand, List, and Schmoller in Germany and by Leslie, Ingram, William Ashley, and Clapham in England, the original aim of the historical school was to replace what they believed to be the unrealistic theories of deductive economics with theories developed inductively through the study of history. They held that history was the key source of knowledge about humans and human organizations, and because it was culture and time specific, it could not be generalized over time or space; hence, general theories were useless. Their view was that economics was best approached from the vantage point of empirical and historical analysis, not abstract theory and deduction.

[13]Ashley (1893, 1927), Cameron (1976), Clapham (1931), Harte (1971), Kadish (1989), Maloney (1976), and Mitch (2010, 2011), all wrote about the evolution of the economic history discipline.

[14]See Reinert and Carpenter (2014) for an overview of German language economics texts written before 1850.

The historical school was a reaction against abstract theory, and it was highly critical of the method, fundamental assumptions, and results thereof. List (1877) was the mouthpiece for the rising nationalistic rivalry of Germany with England, where modern political economy, founded on the practices of abstract theory, ruled. He accused theorists of failing to recognize historical relativity in the stages of national economic development and the use of the productive forces of a nation.

Knies (1853) weighed in against the "absolutism of theory" and the economists, such as Ricardo and Smith, who he claimed based their entire deductive system upon the operation of self-interest for the greater good. Like the other historical economists, he demanded that the whole complex of motives and interests, varying among themselves in intensity at different occasions and times should always be taken into account when considering any type of human behavior. All members of the historical school, but chiefly Roscher, stressed the importance of the comparative method as essential to the understanding of any people or institutions.

Before Schmoller, the historical economists had focused their work more on the field of history than economics. The distinguishing characteristic of Schmoller's work was that it aimed to account for the origin, growth, persistence, and variation of institutions in so far as they affected the economic aspect of life. While he was trained in the historical school, he differed in his emphasis on economics, making him perhaps the first true economic historian.

Schmoller, who studied with Roscher, did not believe the social sciences were suited for any but the simplest mathematical treatment due to the plethora of social interactions that need be considered. He considered statistics, for those variables that could be measured, an invaluable auxiliary to historical research, but always questioned the source and interpretation of the data in relation to other cognate facts and theories. However, the fact that he was willing to go this far is what distinguished him from his mentor and the elder historical scholars.

In the 1880s, the historical school of economics began to diverge. The more conservative branch, the historical economists who followed in the line of the original historical school (the elder branch), abandoned the use of theory altogether. This line was headed by Adolph Wagner. It was an important and valuable work, but Veblen (1901) argued that this conservative historical economics was bereft of theory and hence not economics at all. The other branch was represented by Schmoller and was the wellspring of the first generation of American economic historians.

In the UK, Alfred Marshall and Francis Edgeworth represented the antithesis of the "elder branch" and were on the forefront of a movement to incorporate formal, mathematical models into economics. It was the publication of Edgeworth's book, *New and Old Methods of Ethics*, in 1877 that prompted Marshall to write him and say "There seems to be a very close agreement between us as to the promise of mathematics in the sciences that relate to man's action."[15] Marshall (1897) viewed mathematics as a method of constructing absolutely true arguments. Whereas the

[15]Weintraub (2002), p. 21

historians called for facts and figures, Marshall stressed the danger of committing oneself to them before the theoretical foundations had been established.

The interest in economic history began to grow in the late nineteenth century. This led to the creation of exams, which necessitated teachers. The adoption of economic history for examination in the History Tripos at Cambridge in 1875 led to the publication of the first English language textbook in the subject by William Cunningham in 1882. The History Tripos produced its first fully fledged economic historian, John Harold Clapham, in 1898.[16]

Cunningham's two seminal contributions to economic history were his lifelong efforts to advance the subject through the further work on his textbook (1882), which had five editions and grew to three volumes, and his vigorous campaign to achieve public and scholarly recognition for the approach of economic history (1892). He was a candidate for the Chair in Political Economy at Cambridge that went to Marshall in 1885. The two were antagonists for the remainder of their lives, which did nothing to promote the discipline of economic history.

In Cunningham's view, the election of Marshall to the chair signified that Cambridge was favoring an antihistorical approach to economics. Marshall's triumph within his own field represented the nearly complete victory of the deductive over the inductive approach to economics on the eve of the twentieth century.

Economic history set its first serious footings in 1895 when the London School of Economics (LSE) opened its doors. It was founded in opposition to the tenets of orthodox economics. As a result, economic history was an important presence from the beginning. The first Director of LSE was a young economic historian named W.A.S. Hewins. In 1901, it became the first British university to offer a degree in economics, and economic history became a possible specialty. The first teachers of the subject were Hewins and Cunningham.

In France, the Annales School, focusing primarily on late medieval and early modern Europe, prevailed. It was developed by French historians to stress long-term social history. The school has been highly influential in the use of social scientific methods by historians, emphasizing social rather than political themes. Stoianovich (1976) and Forster (1978) credit the functional and structural approaches to history of the *Annales* School with moving the study of history from storytelling to problem solving.

At the dawn of the twentieth century, it appeared that the attempt of the historical school to replace deductive theory with inductive theory had failed. In fact, the economics discipline was moving toward a more deductive approach. The movement to turn economics into a science, which grew out of the rising stature of the natural sciences, gave way to a new understanding that for economics to take its place at the pinnacle of the social sciences, it needed to formalize and rely more on mathematical models.[17] This set in a period of waning of the historical movement and a historical low point in the discipline.

[16]See Tribe (2000).

[17]For a history of the mathematical movement in economics, see Weintraub (2002).

After WWI, economists became less theoretical and more statistical in their approach. The creation of the National Bureau of Economic Research, which is discussed below, is an example. This movement brought economists and historians a bit closer together. As an added benefit, it forced historians of all stripes to be less tolerant of loose, unsupported generalizations. The culmination was the creation of the first dedicated economic history society, the *Economic History Society*, organized in the UK in 1926, followed in 1927 by the first dedicated economic history journal, the *Economic History Review*.[18]

Economic History in America

American academics were from an early day interested in data. The *American Statistical Association* was launched in 1839, its membership consisting of individuals who paid serious attention to compiling time series data. By the late nineteenth century, numerous state and local historical societies as well as the *American Antiquarian Society* (founded in 1884) could boast of vigorous data accumulation efforts. The federal censuses had flourished from 1790 onward, with attention to economic measurements increasing after 1850. Among the earliest American publications in the subject of economic history along these lines included Freeman Hunt, *Lives of American Merchants* (1858); James L. Bishop, *History of American Manufactures from 1608 to 1860* (1861); and Thomas P. Kettell, *One Hundred Years' Progress of the United States* (1870). And even earlier, there were accumulations of quantitative data in time series form, such as Timothy Pitkin, *Statistical View of the Commerce of the United States* (1816), and Adam Seybert, *Statistical Annals* (1818).

There was no specialized outlet for the publication of research in economic history prior to WWI, but more mainstream economics journals did occasionally publish research in the field. Among the earliest economic history articles were Charles F. Dunbar's "Economic Science in America" in the *North American Review* in January 1876 and Guy Callender's (1903) *Quarterly Journal of Economics* article on early US transportation and banking.

Harvard was the incubator of economic history in the USA. Dunbar, professor of political economy – the first in the USA to be so titled – and founder of the Harvard economics department, along with his colleague Frank W. Taussig, taught courses titled "Financial History of the United States" and "The Tariff History of the Country." In 1882, J. Laurence Laughlin, who later would found the University of Chicago economics department, and Taussig combined to offer a course on US banking and financial legislation, and Laughlin taught a history of political economy course. The following year, Dunbar offered "Economic history of Europe and America since the 7 Years' War," and Taussig taught "The history of tariff legislation." In 1888, he published the first edition of his *Tariff History of the United States*.[19]

[18]See Barker (1977), Berg (1992), and Harte (2001) for a history of the *Economic History Society*.
[19]Mason (1982)

In 1892, Dunbar and Taussig were responsible for the hiring of William J. Ashley to the first chair of economic history in the world. Ashley's reputation as an economic historian was made with the publication of his history of the English woolen industry (1888).

Ashley studied under Arnold Toynbee, the Oxford-based scholar who coined the term "industrial revolution," and Schmoller at Berlin. In 1885, he left Oxford to accept the position of Professor of Political Economy and Constitutional History at the University of Toronto and published his great work, *An Introduction to English Economic History and Theory.* Volume two came out in 1893, the year after he moved to Harvard, where he remained until 1901. Ashley (1927) argued for a course in economic history alongside the general economic theory (i.e., political economy) course. Later in his career, he promoted statistics, which he felt would become an integral part of every important economics department.

Ashley was strongly influenced by German scholarship, as was his Harvard successor, Edwin F. Gay. Gay went to Germany in 1890 to do graduate work in medieval and ecclesiastical history at Leipzig and then Berlin, where he attended the Schmoller economic history seminar in 1893 and became a convert. He wrote little but imparted the standards and techniques of the German academy – the methodological principle of sticking to the facts, of telling history as it really was – on his colleagues and students. Gay used a multidisciplinary approach when teaching. It was the same principle he learned from Schmoller, who was famous for his saying: "Aber, meine Herren, es ist alles so unendlich compliziert," to convey his insistence that the big picture always had to be kept in mind. To account for the complexity, Gay taught his students that hypotheses had to reflect several approaches, including social, political, international, and psychological, as well as economic. Dunbar, Ashley, and Gay had brought over the German concept of stages in development, the notion of a particularly important "take-off" era which Toynbee (1884) had labeled the "industrial revolution," and even many of the criticisms of the industrial system – a particular focus of Toynbee's writings, which were soon to energize the "muckrakers" and subsequent social reformers.

Gay produced a noteworthy assemblage of doctoral candidates including Chester Wright, Norman S. B. Gras, Abbott Usher, Julius Klein, and Earl J. Hamilton, all of whom manifested in one way or another their perception of economic (or later business) history as an adjunct of economic theories. Wright (1941) attempted to carry the relationship farthest as in his *Economic History of the United States*, as did Gras (1962) in his efforts to extend the Germanic scheme of economic stages to capitalism. But it is Frederick J. Turner (1893) of Wisconsin who may be credited with the first serious American contribution to economic history with his work on the American frontier.

In the first decades of the twentieth century, economic history spread across departments, if not in influence within the discipline. Chairs in economic history were created at many leading institutions, but the discipline had difficulty gaining traction due to the lack of a dedicated journal or society to promote its research. Contributing to the problem was the growing fascination with the scientific method and its potential applications to economics, exemplified by the theoretical approach

espoused by Marshall in the UK and soundly rejected by economic historians. In the USA, this manifested itself in the growth of economic forecasting. As Friedman (2014) details, this eventually led to the creation of the *National Bureau of Economic Research* (NBER).

The NBER

Wesley C. Mitchell believed that economic theories were not immutable laws, but rather that they depended on context and evolved over time. He was interested in developing the field of economics into one that took into account what human beings actually did. This was the influence of his mentor at Chicago, Thorstein Veblen. His *Business Cycles* (1913) was an assemblage of business data and his comments on the various series, which seemed to forecast a new theory of business cycle movements. Arthur Burns considered it the key economics text between Marshall's *Principles of Economics* (1890) and Keynes's *General Theory* (1936).[20]

New areas of exploration gave evidence of continued advance in research techniques after WWI. The business cycle provided a field exciting by reason of its relative novelty. Arthur Cole (1930) attempted to extend this analysis backward using antebellum time series data as early specimens of Warren Persons's (1919) A, B, and C curves. At Columbia's Council for Research in the Social Sciences, Gayer et al. (1953) collaborated on a somewhat comparable study using British data from 1790 to 1850. American economic historians had thus made considerable strides in the handling of statistical apparatus.

During his service to the US government during WWI, Edwin Gay became convinced of the need for better economic statistics. He and Mitchell headed the Central Bureau of Planning and Statistics, responsible for the gathering and reporting of statistical data. Together they helped found the NBER to stimulate the collection and interpretation of historical statistics.

Mitchell served as research director at the NBER from its founding in February of 1921 until 1945. He gathered tremendous amounts of empirical economic data in order to draw inductive generalizations from it. His vision was to improve society through the use of expert analysis and statistical investigation. He believed that disseminating scientifically objective data and improving knowledge about business cycles could aid government and business leaders in enacting countercyclical policy that would mute the business cycle.

He combined his historical approach to understanding cycles, which he saw as a global phenomenon, with an urgent call for more data collection from around the world. The NBER was central to this data collection effort and served as a sort of haven for statistical economists. The mission of the NBER was to gather empirical information of many kinds about the American economy in order to create a robust foundation for theoretical generalizations.

[20]Friedman (2014), p. 174

After WWI, this expansion and increased proficiency in the use of statistical materials took attention and resources away from economic history, and it began to lose resources and graduate students to "applied" fields such as international finance, statistics, and the business cycle. The Great Depression only made things worse. Enrollment in economic history courses held steady since major universities required a semester of it in their graduate programs, but writing it as a field declined. Norman Gras (1931) gloomily summarized the state of economic history as being neglected by universities, who regarded it as a very special subject, but one suffering a lack of intellectual resilience.

The NBER ultimately served as a catalyst for the change in emphasis from narrative to quantitative studies in economic history. Mitchell, Simon Kuznets, Arthur Burns, Solomon Fabricant, and Harold Barger produced a series of quantitative descriptions of American economic growth while at the NBER that measured growth as far back as the 1870s. The culmination of this quantitative approach to descriptive economic history was the *Historical Statistics of the United States* (1960) produced by a committee of scholars and sponsored by the Census Bureau.

Over time, economic history presented itself as empirical and multidisciplinary. Empirical in that it dealt with the facts of the past. The facts could be quantitative, as the NBER emphasized, or qualitative (as the German school believed was the responsibility of economic historians). It was also empirical in that economic historians saw history as a laboratory where they could test economic hypotheses.

Business History

The post WWI era also saw the blossoming of the field of business history. Wallace B. Donham initiated the study of business history at the Harvard Business School when he succeeded Edwin Gay as dean in 1919. Gay had encouraged some of his students to pursue the subject. Donham was more ambitious than merely directing graduate students toward the subject. He helped create the *Business History Society* in the early 1920s and raised funds for the endowment of a chair in business history that was filled by Norman S. B. Gras and oversaw the publication of the *Journal of Economic and Business History*, the first American journal dedicated to economic history and the first in the world to combine it with business history. A falling out between Gras and Gay eventually led to Gras's isolation from economic history, largely populated by Gay's progeny, and the gradual drift of business history into a separate field of study.

The differences between business history and economic history are many. Cole discusses them (1945) and Gras (1962) enumerates seven of them, most prominently that economic history comes from economic theory, whereas business history uses economic theory, but also psychology, politics, and sociology, among other disciplines, and none is more important than any other.

Gras, in true Schmoller fashion, conducted detailed research in the archives of corporations; and under his leadership, the genre of business history took shape as a narrative form. His colleague at Harvard, Arthur H. Cole, undertook to sponsor

research on the role of the entrepreneur in economic history, about which more will be said later.

Like his mentor Gay, who trained the first generation of American economic historians, Gras was responsible for training a generation of business historians. He and his followers believed that firms mattered because they were different (i.e., heterogenous) and not the homogenous profit maximizing entity modeled by economists. He defined the subject matter and approach that research in the discipline would take and wrote the first general treatise in the field (1939). He edited the *Harvard Studies in Business History* and served as editor of the *Bulletin of the Business Historical Society* from 1926 to 1953. In 1954, it was renamed the *Business History Review.*

The Depression was not kind to the *Business History Society.* Funding evaporated and the Harvard Business School had to curtail activities to the bare minimum. Business history switched from a general study to one of company histories and individual biographies of businessmen, for which funding was easier to obtain from private sources.

Another major contribution of business history was its focus on entrepreneurship. The *Research Center in Entrepreneurial History* (1948–1958) at Harvard was led by Arthur Cole and supported by a grant from the Rockefeller Foundation as part of the foundation's drive to support and encourage the study of economic history. The center was multidisciplinary in its approach and brought together sociologists, economists, and business historians including luminaries such as Joseph Schumpeter, Thomas Cochran, David Landes, and Alfred D. Chandler. The Center was distinguished by its willingness to address big issues related to explaining the role of entrepreneurship in economic development.

Cole focused on entrepreneurs as the unifying theme under which all issues (growth, change, development, etc.) could be understood. From the beginning, the focus on the center's study of entrepreneurship faced the problem of identifying just what entrepreneurship was. This problem would plague the center for its entire existence. The Center's grants dried up and it ceased to exist in 1958. Before it did, it launched *Explorations in Entrepreneurial History*, which was originally conceived in 1949 as an in-house organ and was published as such until 1958. It was reborn as a "second series" in 1963, renumbering with volume 1, number 1, and eventually renamed *Explorations in Economic History* in 1969.[21]

The *Committee on Research in Economic History* (CREH)[22] was launched in 1941 with the intellectual promotion and financial aid of the Rockefeller Foundation, which seeded it with a $300,000, 4-year grant. The CREH members were united in their worry that mathematical, technical economics would take over the discipline. Their primary concern was that perspective would be lost and current problems would lose their historical context. The CREH introduced and gained nearly

[21]Hugh Aitken (1965) edited a volume of some of the best work appearing in *Explorations in Entrepreneurial History.*

[22]See Cole (1953, 1970) for a history of the CREH.

universal support for an examination of the role of government, especially state government, in the period before 1860. It also underwrote the research for the Census Bureau's volume of *Historical Statistics of the United States*.

The *CREH* lacked a unifying theme until Cole introduced entrepreneurship as a field of potential interest. The eclectic and disjointed nature of the research carried out under the grant was frustrating to Rockefeller Foundation committee members Kuznets and Robert Warren. In 1942, Kuznets resigned over the issue, but was talked into staying until the end of the war. Warren was more proactive, lobbying against renewal of the grant in 1945, going so far as to question whether there was a justification for maintaining economic history as a field separate from economics. Despite such protestations, the grant was approved for a 5-year extension. Ultimately, the *CREH* was not refinanced, but instead the funds were given to the *Center for Entrepreneurial History*, headquartered at Harvard under the direction of Cole.

Founding of the EHA

The first hint of a move to create an American economic history society came in a letter from Earl J. Hamilton to Anne Bezanson urging her to raise the issue of the formation of an American *Economic History Association* with her mentor, Edwin Gay, widely regarded as the most influential American economic historian of the first half of the century.[23] Gay, who would ultimately become the first president of the *EHA*, was a well-respected figure in the field. He had retired from Harvard in 1936, where he had trained many of the current crop of economic historians, and was now located at the Huntington Library in Santa Monica, CA. Despite his distant location on the West Coast and his retirement from academia proper, he was still a dynamic force, widely considered to be the keystone to the creation of an American economic history organization. In his letter to Bezanson in May of 1937, Hamilton said, "... you and I know that he is the one man who would have a good chance to succeed in this difficult undertaking."[24]

The formation of a US economic history society was spurred in part by fear. Edwin Gay's students, who dominated the American field of economic history, had learned the empirical, inductive approach to economic history, and when it was threatened with extinction by the growing mathematical approach, they sought a refuge. Hamilton, a Gay protégé, was the first to attempt a rescue. In 1937, he tried to rally colleagues to create an economic history association in the USA. He feared that the *American Economic Association* (AEA) was planning to eliminate economic history sessions from their annual meetings. The endeavor failed, in part due to the concern that it would cannibalize the UK *Economic History Society* and result in two weak sisters. A renewed attempt succeeded a few years later.

[23]There are several histories of the EHA, among them Aitken (1963, 1975), Clough (1970), Cole (1968, 1974), de Rouvray (2004a), Haupert (2005), and Heaton (1941, 1965a, b).

[24]EHA archives

The outbreak of WWII, which was expected to lead to a decrease in scientific exchanges between the USA and Europe, was the factor that finally led to the formation of an American economic history society. At the *American Historical Association* (AHA) meetings in Washington in 1939, Herbert A. Kellar gathered a group of interested scholars and decided to form the *Industrial History Society*. That same month in Philadelphia at the AEA meetings, Hamilton convened a steering committee chaired by A. H. Cole, including Herbert Heaton (vice chair), Hamilton, and Anne Bezanson (secretary). They were charged with forming an organization independent of any existing society, cooperating with existing organizations, enrolling members, and planning for the publication of a journal.

The group was cautious in its approach, fearful of encroaching upon the membership and damaging the existing societies of mutual interest. Their first objectives included the arrangement of meetings in collaboration with the historians in New York and of the economists in New Orleans in December 1940 and a survey aimed at gauging interest in an *Economic History Association*. Bezanson wrote to more than 500 potential members, receiving positive responses from more than 400. In response, Hamilton arranged to organize sessions in New Orleans, while Heaton would do the same in New York.

On December 27, 1940, the *EHA* debuted at a joint session with the *American Historical Association* in New York City with a session on "The Business Cycle and the Historian," chaired by Herb Feis. The following day, they conducted their first solo session with Professor A.P. Usher chairing "The Next Decade in Economic Historiography." A business meeting followed during which the steps to form the Association were formalized. Edwin Gay was named its first president and Shepard B. Clough secretary-treasurer, and arrangements were made to secure publishers for the new journal. About a month later, E. A. J. Johnson was named editor, Clough associate editor, and Winifred Carroll assistant editor of the new *Journal of Economic History*. The *JEH* debuted in 1941, barely 4 months after the naming of the editor. Harold Innis was the chair of the first stand-alone *EHA* meeting, held in the fall of 1941 at Princeton. On the 30th, in New Orleans, there was a joint session with the *American Economic Association* followed by a luncheon and business meeting, which endorsed the actions of the steering committee. Thus, the Association was born – in two places, over a period of 5 days.

Gay reluctantly agreed to serve as the first president of the *EHA*. He clearly had the skills to do the job, having previously served as president of the *National Bureau of Economic Research* from 1919 to 1933, the *American Economic Association* in 1929, and the *Agricultural History Society* (AHS) in 1934. His organizational skills, which were so eagerly sought, were well known. He was cofounder of both the NBER and the Harvard Graduate School of Business Administration, serving as its first dean. His academic career began with an appointment to the economics faculty at Harvard University in 1902. He went on to become chair of the department and, more importantly for the future of economic history, a mentor to an entire generation of economic historians. His reluctance came from the fact that he had retired to California for the express purpose of reducing his administrative responsibilities in order to concentrate on research.

At the first stand-alone *EHA* conference, John Nef (1941) took on the lofty duty of describing the responsibility of economic history, noting that the creation of the EHA at such a critical time came with obligations as well as privileges. First and foremost, the association had the duty of considering its objectives, if, as seemed probable at the time, Western Civilization had reached the end of an epoch.

The *Economic History Association* became a trendsetter in 1941 when it held its first annual conference in September. The usual pattern was for academic societies to hold their conferences the week after Christmas. The steering committee wanted to avoid the prospect of competition with its potential members, many of whom held memberships with either the AHS or the AHA, both of which met during the traditional Christmas break period. In addition, most members were also members of the AEA, which regularly met over the break. The perfect solution seemed to be to move the *EHA* meeting to another timeslot. Early September was chosen because it was before classes started at most universities, and it would avoid competing with related societies and the glut of holiday meetings already crowding the calendar.

By 1941, Gay felt that the work of the historical economists had not been able to displace the "theoretical school," but did modify it. By then, the use of the deductive method had become more guarded, and the practitioners of this "dark art" had increased the range and depth of their contemporary observations, and their viewpoint had expanded to become less individualistic and more social. In conclusion, he called for the reunification of economic history and theory, noting that the economic historians knew a great deal about the long trends of productive energies and social pressures leading to economic growth, which could be combined with the tools of the theorist to lend greater insight into the growth process. Far from incompatible, he felt that true philosophical objectives and the careful assembling of data were complementary.

E. A. J. Johnson (1941) fretted that while the number of tools available to economic historians was increasing, there was still too much work in economic history that was a haphazard gathering of facts with little appraisal of whether they actually shed any light on economic development. With so much to do to have a more complete understanding of the past, he felt economists should focus only on the most important tasks, such as productivity, capital formation, changes in income, regulation, consumption, and its effect on the entire structure of the economy. All of which, he argued, could be advanced by employing the most efficient theoretical tools at the disposal of the economic historian. He cited Leo Rogin's (1931) work on farm productivity as an example of measuring productivity (in a nonstatistical age), and how its changes, brought about by evolving technology, could be measured.

Cochran (1943) saw the use of limited and modest economic hypotheses, such as monopolistic competition or location theory, as more useful tools for practical research than the sweeping assumptions of the historian. He felt that specific limited propositions could be the first steps in applying to the data of social science a logical technique, similar to that developed in the natural sciences for stating and testing hypotheses.

So as the EHA was in its infancy, formed in part as a defense against the encroaching "mathematicization" of the discipline, the seeds of the cliometric

movement were being sown. It would be the next generation of economic historians who propelled that movement forward.

The New Economic History Movement

Arguing against those who cliometricians would later label "old" economic historians (the likes of Anne Bezanson, Arthur Cole, Edwin Gay, Harold Innis, and Earl Hamilton, the founders of the EHA), Kuznets claimed that little would be gained from a study of the past unless it was systematic and quantitative. According to him, this was the only way to weigh the relative effects of factors and events. The reason for the small quantity of quantitative work in economic history was due to the extraordinary effort necessary before the computer to sift and classify quantitative information and the relatively recent development of statistical theory and techniques capable of handling these problems.

After WWII, with the American economy booming, economists gained cachet. Economics with its rigorous models, tested from an abundance of numerical data by use of advanced, mathematically expressed formulae, came to be regarded as the paradigm of the social sciences. William Parker (1986) quipped that if economics was the queen of the social sciences, then economic theory was the queen of economics, and its handmaiden was econometrics.

At the same time as this increasingly technical focus, economists were increasingly interested in the determinants of economic growth and what they saw as the widening gap between the so-called developed and underdeveloped regions of the world. They saw the study of economic history as a source of insight into the issues of economic growth and economic development and the new quantitative methods as the ideal tools for analysis.

Norman Gras (1962) had a different take on economic history. He believed that it was showing its old age and declining in influence while business history was in its ascendency. His primary evidence for this claim was the decline in Germany, the incubator for the discipline, where he thought economic history had lost prestige at the expense of general history. Time would prove that Gras had a point. The rise of the cliometric movement in the USA was not mirrored in Europe, and when it did finally begin to make headway, it was in the UK before the continent.[25]

The generation of economists who were trained in the postwar decades found ways to mesh mathematics and economics, although the idea that economics should appropriate ideas from mathematics was itself contested, especially by economic historians. By the 1960s, the battle was over and the results were clear: economics was a "science," constructing, testing, and applying technically sophisticated models. Econometrics was on the rise and economic historians were divided

[25]Cliometrics did not dominate the European scene as early or as completely as it did in North America. See Tilly (2001) for an overview of German clio, Grantham (1997) and Crouzet and Lescent-Gille (1998) for France, and Floud (2001) for the UK.

between those who abhorred it, and those who embraced it. The former faded in influence and their followers retreated to history departments.

The "new" economic history can be dated to the 1957 joint meeting of the *EHA* (founded in 1940 by "old" economic historians like Gay and Cole) and the Conference on Research in Income and Wealth (under the guidance of the NBER). In particular, two joint papers by Alfred Conrad and John Meyer (1958) constituted the manifesto for the new era. The first paper, on methodology, explained what scientific method was really all about and how it applied to economic historians. Parker (1980) cites the second paper as one of the most influential in the evolution of economic history. It added enormous force to the methodological prescription by claiming to follow it in an analysis of the profitability of slavery on the eve of the civil war. The analytical method, the data, the economic and accounting framework, and the choice of slavery as a subject were to have vast consequences for the next generation of economic historians. The meeting produced a volume edited by Parker (1960), which included such path-breaking work as Robert Gallman's estimates of commodity output, the farm gross product and investment series produced by Marvin Towne and Wayne Rasmussen, Douglass North's balance of payment estimates, and Stanley Lebergott's wage series.

Goldin (1995) noted that economic historians who came before the cliometric revolution distinguished themselves by mastering a wide array of facts and knowledge of institutions. But without the rigor of theory and econometrics, they could not avoid the occasional faulty reasoning. Important data were overlooked without the ability to properly test theories.

de Rouvray (2004b) argues that the timing of the cliometric movement corresponded to the success of Kuznets's quantitative growth studies; a reflection of the infatuation economists had developed for the national accounting approach. This predisposed them to view the past through this same lens and altered their definition of historical evidence. Fogel (1965) agreed, crediting his mentor Kuznets as the primary inspiration for the work of the new economic history.

But cliometrics is not identical to the Kuznets approach. Cliometricians are eager to apply neoclassical models to even marginal historical issues, which would not have been a priority of Kuznets, who was focused on the bigger issue of economic growth. But the emphasis on quantification and measurement and the decreased reliance on qualitative measures were certainly in synch with the Kuznets methodology.

Kuznets may have inspired the cliometric movement, but it was Robert Fogel who reunified economics and history. He used the latest techniques of modern economics and gathered reams of historical data to reinterpret American economic growth in sectors as diverse as railroads, slavery, and nutrition. Rather than conjecture about the causes of growth, he carefully measured them. He pioneered the use of large-scale cross-sectional and longitudinal data sets harvested from original sources to examine policy issues. McCloskey (1992) credited his contributions with opening new ways to the past.

Fogel's breakthrough work was *Railroads and American Economic Growth* (1964a). At the time of its publication, economists believed they had established

that modern economic growth was due to certain important industries having played a vital role in development. Fogel set out to measure this impact, which he did with extraordinary precision. He constructed a counterfactual to highlight the contributions of the railways to the growth of the American economy. The result was not what economists or historians expected. He famously found that the railroad was not absolutely necessary in explaining economic development and that its effect on the growth of GNP was minimal. Few books on the subject of economic history have made such an impression as Fogel's. His use of counterfactual arguments and cost-benefit analysis made him an innovator of economic historical methodology, but not universally loved. Fritz Redlich (1965), for example, accused him of "fictitious quasi-history" for his emphasis on the counterfactual. He acknowledged the value of counterfactual analysis, but thought it was social science research, not historical.[26]

This approach formed his major works on slavery and demography as well.[27] Fogel recognized early in his career that to answer such questions much greater use had to be made of quantitative evidence, so he mastered the most advanced analytical and statistical methods available and successfully employed them in his research. Herein was the difference between the "old" economic history and the "new:" the use of newly created data series and cutting edge techniques – made more useful, applicable, powerful, and easy to replicate and reconsider, with the growth of computing power, to bring a finely focused eye on a problem.

Fogel was not the first to use a form of identifying opportunity costs known as counterfactual analysis, but he was the most extensive user of it and became famous (infamous?) for his use of the technique in his landmark railroad study. Counterfactual analysis is the idea of determining the impact of an event or factor by considering what would have happened in its absence. Before Fogel, the concept was proposed by Fritz Machlup (1952), Meyer and Conrad (1957) and Conrad and Meyer (1958).

Like Fogel, Douglass North made his initial impact with research on the American economy. However, whereas Fogel disputed the importance of one sector of the economy in explaining economic growth, North focused on the impact that individual sectors could have in explaining economic outcomes. He sought to explain the causes of growth in the antebellum American economy. Starting with an export-based model he had previously formulated, he showed how one sector (the cotton industry) could stimulate development in other branches, ultimately leading to specialization and interregional trade.

North also focused on quantification early on, measuring the impact of decreased transoceanic shipping costs. His surprising finding was not that shipping costs decreased, which was widely recognized at the time, but that it was not technology, so much as institutional changes, such as a decrease in piracy and faster turnaround times in port, that was the source of the decreased costs. This focus on institutions would become North's mantra for the remainder of his career.

[26]For a different view of counterfactuals, see Engerman (1980).

[27]See Fogel and Engerman (1974) and Fogel (2000), for example.

Goldin (1995) notes that the cliometric revolution pitted young Turks, outsiders, "theorists" as they were called by the old timers, against those "old" economic historians who were more likely to be historians and less likely to rely on quantitative methods. They accused the newcomers of bringing economic theory to history without a proper understanding of the facts (a familiar battle cry). Cochran (1969) characterized the disagreement as one about the choice of models. The old guard claimed that realistic models had to be too highly generalized or too complex to allow the assumption of mathematical relationships. The "new" economic historians, however, were primarily interested in applying operative models to economic data. There was a difference in method between new and old economic historians that could not be ignored. The models preferred by the new economic historians were quantitative and mathematical, while those used by "sociological economic historians" tended to be narrative.

The schism was not just about methodology, but also orthodoxy. Cliometricians were using their new tools to overturn some long-held beliefs. Among the accepted wisdom they overturned was that railroads were indispensable to economic growth (Fogel 1964a), that they were built ahead of demand (Fishlow 1965), that President Jackson caused the financial panics of the 1830s (Temin 1969), and that slavery was unprofitable (Conrad and Meyer 1958).

Andreano (1970) collected a series of articles originally published in *Explorations in Entrepreneurial History, Second Series* that he felt reflected dialogue that had been taking place during the 1960s between economists and historians on the methodology of the "new" economic history. But the first attempts to bring together a body of work representative of the "new" economic history were the publication in 1971 of *The Reinterpretation of American Economic History*, a collection of essays edited by Fogel and Engerman, and Davis et al.'s *American Economic Growth: An Economist's History of the United States* (1972).

The reception of the "new" economic history was chilly by some due its perceived threat to traditional historical methods, but warmly welcomed by others for the possibilities it promised. Hughes and Reiter compared the computational effort it took them in their steamships paper (1958) to that of Newmarch (1857), who compiled more than 13,000 individual pieces of information and then performed a mere three arithmetical calculations, but all by hand. His efforts represented a lifetime of work, while the steamship paper was but one of many "big data" projects cliometricians would explore with the power of new techniques and technology.[28] The steamship study had a total of nearly twice as many observations (as the Newmarch data set) on 1945 punch cards, but the computer then did all of the computational work.

[28]For example, they cited four additional data-processing studies in economic history carried out at Purdue in the late 1950s that had developed entirely new statistical series and could not have been conducted without the latest technology or mathematical models: Lance Davis's textile studies (1957, 1958, 1960) and the Davis and Hughes exchange rate study (1960).

Cliometrics got the platform it needed to take off when North and Parker were named editors of the *Journal of Economic History* in 1960. Robert Whaples (1991) found that the journal led the *EHA* meetings (a selection of whose papers were represented in the *Tasks* issue) in the new cliometric methods. From 1956 to 1960, 10% of the papers were "clio," but only 6% of the *Tasks* articles featured cliometrics. From 1961 to 1965, the numbers were 16% and 15%; from 1966 to 1970, 43% and 18%, respectively; and from 1971 to 1975, they skyrocketed to 72% in the journal and 40% at the conference. This reflects the difference in editorship of the journal (North and Parker, proponents of the "new" economic history) and leadership of the *EHA*, whose presidents were of an older, less quantitatively, and decidedly not "new" economic history background.

The Shortcomings of Clio

Clio has not had an uncheckered history. Its rise has led to a rift between economists who practice cliometrics and historians who practice economic history without the use of the formal models, which they argue miss the context of the problem and have become too enamored of statistical significance at the cost of contextual relevance. Boldizzoni (2011) famously attacked cliometrics, focusing his sharpest criticism on the quantification of history at the perceived expense of its humanity. On the other side, cliometrics has lost some of its significance with economists, who do not see it as anything more than another application of economic theory. While applied economics is not seen as a bad thing, cliometrics is not seen as anything special, just applying theory and the latest quantitative techniques to old data instead of contemporary data.

As early as 1986, Parker observed that what was lost in the move to theory and econometric emphasis was the humane interest of the old British political economy and social welfare and the idealistic German historical economist's concern for the whole society – i.e., the Schmoller perspective. At the same time, Alex Field (1987) cited problems from another flank. Whereas the "new" economic historians had to fight to prove their technical skills belonged in the study of history, by the late 1980s, there were no more "old" economic historians left to challenge. Instead, the challenge came from the other side, where economic theorists questioned what value cliometricians added to departments strapped for resources. Most economists possess the same or even more sophisticated technical skills, which can be applied to any data set, contemporary or historical.

Even within the cliometric camp, there were those who cautioned against the over reliance on technique. In the early days of the cliometric movement, Jonathan Hughes (1966) warned that cliometrics is unkind to those who confuse ends and means in the pursuit of historical understanding. And Lance Davis (1968), though praising the new economic history for its contributions to both economics and history, criticized indiscriminate uses of theory applied to history. He argued that the greatest failure of the new economic history was the rush by some to apply any theory, even if irrelevant, to a historical issue or, even worse, a handy data set,

without understanding the context of the historical situation. And North (1965) warned that too much of the new economic history was dull and unimaginative because there was too much emphasis on econometric techniques as a substitute for theory and imagination.

Clio's Accomplishments

Clio's moment in the spotlight, or 15 min of fame, as Sam Williamson (1994) coined it, came at the 1964 AEA meetings. William Parker organized a session on "Economic History: It's Contribution to Economic Education, Research, and Policy," featuring papers by Douglass North (1965), Robert Fogel (1965), Barry Supple (1965), Richard Easterlin (1965), Robert Gallman (1965), and Rondo Cameron (1965), with comments by Evsey Domar and R. A. Gordon (1965). The session drew a crowd estimated at 200, generated lively discussion, and put cliometrics in a national spotlight that it had never previously experienced.

Fogel (1964b) highlighted the changes in economic history that justified its being "new." It was not a change in subject; they still remained interested in the description and explanation of economic growth. It was the approach to measurement and theory that was new. Economic history always had a quantitative dimension. But much of the past work had been limited to the simple organization of data contained in government and business records. While continuing this pursuit, the new economic history placed its primary emphasis on reconstructing measurements and organizing primary data in a manner allowing them to obtain measurements that were never before possible. It thus followed that the most critical issue in the work of the new economic historians was the logical and empirical validity of the theories on which their measurements were based.

The new economic historians made use of the whole gamut of economic theory and statistical models, and the measurements they obtain yielded considerably more precise information than previously available. The perfect example of this was Fogel's railroad study.

The publication of *Railroads* "represented a very major milestone – it was as if we now had proof that we had left the bumpy and unpaved dirt road of the first few years and could see ahead a straight and well-paved highway into the future," says Lance Davis (2014) in his review as part of Project 2000. The publication of Fogel's railroad study generated an entire subdiscipline of parallel studies and, more importantly, provided a methodological foundation for the systematic study of economic history and long-term economic growth.

Railroads showed how well economic history could benefit from the careful application of theory and econometrics. The work immediately generated substantial controversy, and even today some quibbling over minor details occurs. However, time has failed to overturn Fogel's major conclusions: that per capita income growth would have been set back only a few months had the railroads never been invented, and there was no other industry that was likely to have been more important than the

railroads. Since its publication, the great majority of economic history has been written by scholars employing those basic economic and econometric tools.

Basman (1970) and Field (1987) defended the growing reliance of economic historians on quantitative methods. Cliometrics advocates replacing imprecise qualitative judgments common in narrative history with more precise quantitative estimates. In this way, it does a great service to economic history and economics in general by taking a closer and better informed interest in the formulation and testing of explanatory economic models. They also tried, where possible, to move beyond simple description and informal explanation in historical scholarship to the investigation of causal relationships linking exogenous and endogenous variables. Cliometricians shifted attention from documentary to statistical primary source material and emphasized the use of statistical techniques to test posited relationships among variables.

It is the lack of relevant data more than the lack of relevant theory that is often the greater problem in research. In this way, economic historians have made some of the greatest contributions to the fields of economics and history by discovering and compiling new data sets that can then be used by future researchers to better understand the evolution and growth of economies over time.[29]

Perhaps, the most influential book to come from the new economic history is North's *Economic Growth of the United States, 1790–1860* (1961). What it lacked in thorough empirical research, it more than made up in the way it clearly demonstrated how an economic model, theoretically sophisticated yet nonmathematical, could be employed to explain the organization and evolution of the various regions of the American economy over several decades.

In North's early work (1961, 1966), he focused on the standard neoclassical explanations for economic growth (technology, human capital, technological change). But when he began to study European economic history, he concluded that the neoclassical model was not able to explain the kind of fundamental societal change that had characterized European economies for the past 500 years. This led him down the path of what would become the new institutional economics, making him an early proponent of two different revolutionary schools of economic practice: cliometrics and new institutional economics.[30]

In a number of books, beginning with *Institutional Change and American Economic Growth* (1971, with Lance Davis), North demonstrated the importance of the role played by institutions (including property rights) on economic development. In *Institutions, Institutional Change and Economic Performance* (1990), he posed the fundamental question of why some countries are rich and others poor. His conclusion was that institutions are a major determinant in the profitability and feasibility of economic activity. The greater the institutional uncertainty, the greater the transaction costs and the greater the drag on economic growth and development.

[29]The listing of all such databases publicly available is massive; for an example of the size and scope of such endeavors, see the list of databases on eh.net.

[30]See Basu et al (1987), Galiani and Sened (2014), and Menard and Shirley (2014) for discussions of North's role in the new institutional economics movement.

These views were a novel approach in both the history and development fields. Typical economic growth models focused on technological change and capital accumulation, assuming zero transactions costs and ignoring institutions altogether. He maintained that new institutions arise when groups in society see a possibility of profiting that is impossible under prevailing institutional conditions. If external factors make an increase in income possible, but institutional factors prevent it, then new institutional arrangements are likely to develop. Other pioneering work emphasizing the importance of institutions included R. C. O. Matthews (1986) and Oliver Williamson (1985).

The crown jewel of Clio's accomplishments came in 1993 when the Nobel Prize in economics was awarded to Robert Fogel and Douglass North. The Royal Swedish Academy of Sciences announced the award of the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel jointly to Professors Robert W. Fogel and Douglass C. North "for having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change."[31] As the committee pointed out, both men were leading figures within the field of "new economic history," which is now known as cliometrics.

Conclusion

Economic historians have contributed to the development of economics in many ways, combining theory with quantitative methods, constructing and revising databases, and discovering and creating entirely new ones. This has made it possible to question and reassess earlier findings, thus increasing our knowledge, refining earlier conclusions, and correcting mistakes. In addition, this field has added greatly to our understanding of economic growth and development, affording the economic historian the valuable element of time as a variable, which the traditional theorist does not enjoy. The use of history to examine economic theory has deepened our knowledge and understanding within fundamental areas of research as to how, why, and when economic change occurs. It is perhaps in this area where the greatest contributions of economic historians have appeared.

By merging economic history with modern techniques, cliometricians have not ended economic history but elevated it. The continuing evolution of technology has made a tremendous impact on the ability of cliometricians to handle ever larger data sets, share them with a wider audience, and access new data sets that previously took a lifetime to collate. In conjunction with the greater facility current economic historians have with econometrics, the future seems limitless. But as any good historian knows, predicting it is fraught with perils.[32]

[31] Engerman et al. (1994)

[32] For musings on the future of economic history see Jones et al. (2012), Baten (2004), Baten and Muschallik (2011), Dumke (1992), Field (1987), and Nicholas 1997.

References

Aitken HGJ (1963) The association's membership: growth and distribution. J Econ Hist 23(3):335–341

Aitken HGJ (ed) (1965) Explorations in enterprise. Harvard University Press, Cambridge

Aitken HGJ (1975) In the beginning. J Econ Hist 35(4):817–820

Andreano RL (ed) (1970) The new economic history: recent papers on methodology. Wiley, New York

Ashley WJ (1887) The early history of the English Woolen Industry. Am Econ Assoc II(4):297–380

Ashley WJ (1888) An introduction to english economic history and theory. Rivingtons, London

Ashley WJ (1893) The study of economic history. Q J Econ 7(2):115–136

Ashley WJ (1927) The place of economic history in university studies. Econ Hist Rev, 1st series 1(1):1–11

Ashton TS (1946) The relation of economic history to economic theory. Economica 13(50):81–96

Barker TC (1977) The beginnings of the Economic History Society. Econ Hist Rev 30(1):1–19

Basmann RL (1970) The role of the economic historian in predictive testing of proffered 'economic laws'. In: Andreano RL (ed) The new economic history: recent papers on methodology. Wiley, New York, pp 17–42

Basu K, Jones E, Schlicht E (1987) The growth and decay of custom: the role of the new institutional economics in economic history. Explor Econ Hist 24(1):1–21

Baten J (2004) Die Zukunft der kliometrischen Wirtschaftsgeschichte im deutschsprachigen Raum. In: Schulz G, Buchheim C, Fouquet G, Gömmel R, Henning FW, Kaufhold KH, Pohl H (eds) Sozial- und Wirtschaftsgeschichte. Arbeitsgebiete-Probleme-Perspektiven. Franz Steiner Verlag, Stuttgart, pp 639–653

Baten J, Muschallik J (2011) On the status and the future of economic history in the world. Munich personal RePEc archive

Berg M (1992) The first women economic historians. Econ Hist Rev 45(2):308–329

Bishop JL (1861) History of American manufactures from 1608–1860. Edward Young & Co, Philadelphia

Boldizzoni F (2011) The poverty of Clio: resurrecting economic history. Princeton University Press, Princeton

Braudel F (1949) La méditerranée et le monde méditerranéen à l'époque de Philippe II. A. Colin, Paris

Callender GS (1903) Early transportation and banking enterprises of the United States. Q J Econ XVII:111–162

Cameron R (1965) Has economic history a Role in an economist's education? Am Econ Rev Pap Proc 55(2):112–115

Cameron R (1976) Economic history, pure and applied. J Econ Hist 36(1):3–27

Carlos A (2010) Reflections on reflections: review essay on reflections on the cliometric revolution: conversations with economic historians. Cliometrica 4(1):97–111

Clapham JH (1929) The study of economic history. In: Harte NB (ed) The study of economic history: collected inaugural lectures, 1893–1970. Frank Cass, London, pp 55–70

Clapham JH (1931) Economic history as a discipline. In: Seligman ERA, Johnson A (eds) Encyclopedia of the social sciences. Macmillan, New York, pp 327–330

Clough SB (1970) A half-century in economic history: autobiographical reflections. J Econ Hist 30(1):4–17

Coats AW (1980) The historical context of the 'new' economic history. J Eur Econ Hist 9(1):185–207

Cochran TC (1943) Theory and history. J Econ Hist 3(December: supplement: The Tasks of Economic History):27–32

Cochran TC (1969) Economic history, old and new. Am Hist Rev 74(5):1561–1572

Cole AH (1930) Statistical background of the crisis of 1857. Rev Econ Stat XII(4):170–180

Cole AH (1945) Business history and economic history. J Econ Hist 5(Supplement: The Tasks of Economic History):45–53

Cole AH (1953) Committee on Research in Economic History: a description of its purposes, activities, and organization. J Econ Hist 13(1):79–87
Cole AH (1968) Economic history in the United States: formative years of a discipline. J Econ Hist 28(4):556–589
Cole AH (1970) The Committee on Research in Economic History: an historical sketch. J Econ Hist 30(4):723–741
Cole AH (1974) The birth of A new social science discipline: achievements of the first generation of American economic and business historians 1893–1974. Economic History Association, New York. Downloaded from http://eh.net/items/birth-of-a-new-social-science-discipline. Accessed Apr 2014
Conrad AH, Meyer JR (1958) The economics of slavery in the Antebellum south. J Polit Econ 66:75–92
Crafts NFR (1987) Cliometrics, 1971–1986: a survey. J Appl Econ 2(3):171–192
Crouzet F, Lescent-Gille I (1998) French economic history for the past 20 years. NEHA-Bull 12 (2):75–101, (Nederlandsch Economisch-Historisch Archief)
Cunningham W (1882) The growth of english industry and commerce. C.J. Clay, Cambridge, MA
Cunningham W (1892) The perversion of economic history. Econ J 2(7):491–506
D'Avenant C (1699) An essay upon the probable method of making a people gainers in the balance of trade. London
Databases, eh.net, http://eh.net/databases/
Davis LE (1957) Sources of industrial finance: the American Textile Industry, a case study. Explor Entrep Hist IX:189–203
Davis LE (1958) Stock ownership in the early New England Textile Industry. Bus Hist Rev XXXII:204–222
Davis LE (1960) The New England Textile Mills and the capital markets: a study of industrial borrowing, I840–1860. J Econ Hist XX:1–30
Davis LE (1968) And it will never be literature: the new economic history: a critique. Explora Entrep Hist, 2nd series 6(1):75–92
Davis L (2014) Review of railroads and American Economic Growth: essays in econometric history. Eh.net Project 2000/2001. http://eh.net/book-reviews/project-20002001/. Accessed 2014
Davis LE, Hughes JRT (1960) A dollar sterling exchange I803–i895. Econ Hist Rev 13(1):52–78
Davis LE, North DC (1971) Institutional change and American economic growth. Cambridge University Press, New York
Davis LE, Hughes JRT, Reiter S (1960) Aspects of quantitative research in economic history. J Econ Hist 20(4):539–547
Davis L et al (1972) American economic growth: an economist's history of the United States. Harper & Row, New York
de Rouvray C (2004a) 'Old' economic history in the United States, 1939–1954. J Hist Econ Thought 26(2):221–239
de Rouvray C (2004b) Seeing the world through a National Accounting Framework: economic history becomes quantitative. Presented at Economic History Society Annual Conference, University of London, Royal Holloway
de Rouvray C (2014) Joseph Willits, Anne Bezanson and economic history: 1939–1954. Rockefeller Archive Publications. http://www.rockarch.org/publications/resrep/derouvray.pdf. Accessed Apr 2014
Domar ED, Gordon RA (1965) Discussion. Am Econ Rev Pap Proc 55(2):116–118
Drukker JW (2006) The revolution that bit its own tail: how economic history has changed our ideas about economic growth. Aksant, Amsterdam
Dumke RH (1992) The future of cliometric history – a European view. Scand Econ Hist Rev 40(3):3–28
Dunbar CF (1876) Economic Science in America, 1776–1876. N Am Rev CXXII:124–153

Easterlin RA (1965) Is there need for historical research on underdevelopment? Am Econ Rev Pap Proc 55(2):104–108
Economic History Association archives, Hagley Museum and Library, Wilmington, DE, Accession # 1479, folders 1–11, 29–31
Edgeworth F (1877) New and old methods of ethics. James Parker, Oxford/London
Engerman SL (1980) Counterfactuals and the new economic history. Inquiry 23(2):157–172
Engerman SL (1996) Cliometrics. In: Kuper A, Kuper J (eds) The social science encyclopedia, 2nd edn. Routledge, London/New York, pp 96–98
Engerman SL, Hughes JRT, McCloskey DN, Sutch RC, Williamson SH (1994) Two pioneers of cliometrics: Robert W. Fogel and Douglass C. North, nobel laureates of 1993. The Cliometric Society, Miami
Evelyn J (1674) Navigation and commerce, their origins and progress. Printed by TR for Benjamin Tooke, London
Fenoaltea S (1973) The discipline and they: notes on counterfactual methodology and the 'new' economic history. J Eur Econ Hist 2(3):729–746
Field AJ (1987) The future of economic history. In: Field AJ (ed) The future of economic history. Kluwer-Nijhoff, Boston
Fishlow A (1965) American railroads and the transformation of the Ante-bellum economy. Harvard University Press, Cambridge, MA
Fleetwood W (1707) Chronicon Preciosum: or an account of English money, the price of corn and other commodities, for the last 600 years. Printed for Charles Harper, London
Floud R (1991) Cliometrics. In: Eatwell J, Milgate M, Newman P (eds) The new Palgrave: a dictionary of economics, vol 1, 2nd edn. Macmillan, London/New York/Tokyo, pp 452–454
Floud R (2001) In at the beginning of British cliometrics. In: Hudson P (ed) Living economic and social history. Economic History Society, Glasgow, pp 86–90
Fogel RW (1964a) Railroads and American economic growth: essays in econometric history. Johns Hopking University Press, Baltimore
Fogel RW (1964b) Discussion. Am Econ Rev 54(3):377–389
Fogel RW (1965) The reunification of economic history with economic theory. Am Econ Rev Pap Proc 55(2):92–98
Fogel RW (2000) The fourth great awakening and the future of egalitarianism. University of Chicago Press, Chicago
Fogel RW, Engerman SL (eds) (1971) The reinterpretation of American economic history. Harper & Row, New York
Fogel RW, Engerman SL (1974) Time on the cross: the economics of American Negro slavery, vols 1 and 2. Little, Brown, New York
Forster R (1978) Achievements of the Annales school. J Econ Hist 38(1):58–76
Friedman WA (2014) Fortune tellers: the story of America's first economic forecasters. Princeton University Press, Princeton
Galiani S, Sened I (eds) (2014) Institutions, property rights, and economic growth: the legacy of Douglass North. Cambridge University Press, New York
Gallman RE (1965) The role of economic history in the education of the economist. Am Econ Rev Pap Proc 55(2):109–111
Gay EF (1941) The tasks of economic history. J Econ Hist 1(Supplement: The Tasks of Economic History):9–16
Gayer AD, Rostow WW, Schwartz AJ (1953) The growth and fluctuation of the British economy 1790–1850, and historical, statistical, and theoretical study of Britain's economic development, vol 2. Clarendon, Oxford
Goldin C (1995) Cliometrics and the nobel. J Econ Perspect 9(2):191–208
Grantham G (1997) The French cliometric revolution: a survey of cliometric contributions to French economic history. Eur Rev Econ Hist 1(3):353–405
Gras NSB (1931) Economic history in the United States. In: Seligman ERA, Johnson A (eds) Encyclopedia of the social sciences, vol 5. Macmillan, New York

Gras NSB (1939) Business and capitalism: an introduction to business history. Crofts, New York
Gras NSB (1962) Development of business history up to 1950, selections from the unpublished work of Norman Scott Brien Gras, compiled and edited by Gras EC. Edwards Brothers, Ann Arbor
Graunt J (1662) Natural and political observations mentioned in a following index and made upon the bills of mortality. London
Greif A (1997) Cliometrics after 40 years. Am Econ Rev 87(2):400–403
Harte NB (1971) The making of economic history. In: Harte NB (ed) The study of economic history: collected inaugural lectures, 1893–1970. Frank Cass, London, pp xi–xxxix
Harte NB (2001) The economic history society, 1926–2001. In: Hudson P (ed) Living economic and social history. Economic History Society, Glasgow, pp 1–12
Haupert M (2005) The birth of the economic history association. Newslett Cliometric Soc 20(3):27–30
Heaton H (1941) The early history of the economic history association. J Econ Hist 1(Supplement: The Tasks of Economic History):107–109
Heaton H (1965a) A scholar in action, Edwin F. Gay. Harvard University Press, Cambridge
Heaton H (1965b) Twenty-five years of the economic history association: a reflective evaluation. J Econ Hist 25(4):465–479
Hughes JRT (1966) Fact and theory in economic history. Explor Entrep Hist, 2nd series 3(2):75–100
Hughes JRT, Reiter S (1958) The first 1,945 British steamships. J Am Stat Assoc LIII:360–381
Hunt F (1858) Lives of American merchants. Derby and Jackson, New York
Johnson EAJ (1941) New tools for the economic historian. J Econ Hist 1(Supplement: The Tasks of Economic History):30–38
Jones G, van Leeuwen MHF, Broadberry S (2012) The future of economic, business, and social history. Scand Econ Hist Rev 60(3):225–253
Kadish A (1989) Historians, economists, and economic history. Routledge, New York/London
Kettel TP (1870) One hundred years' progress of the United States. L. Stebbins, Hartford
Keynes JM (1936) The general theory of employment, interest and money. Macmillan, London
Knies K (1853) Die Politische Okonomie vom Standpunkt der Geschichtlichen Methode, Braunschweig, G. N. Schwetschte und Sohn
Lamoreaux NR (1998) Economic history and the cliometric revolution. In: Molho A, Wood GS (eds) Imagined histories: American historians interpret the past. Princeton University Press, Princeton, pp 59–84
Libecap GD (1997) The new institutional economics and economic history. J Econ Hist 57(3):718–721
List F (1877) Das Nationale System der Politischen Okonomie. Verlag der J.G. Cotta'sche Buchhandlung, Stuttgart
Lyons JS, Cain LP, Williamson SH (eds) (2008) Reflections on the cliometrics revolution: conversations with economic historians. Routledge, London
Machlup F (1952) The political economy of monopoly: business, labor and government policies. Johns Hopkins University Press, Baltimore
Maloney J (1976) Marshall, Cunningham, and the emerging economics profession. Econ Hist Rev 29(3):440–451
Marshall A (1890) Principles of economics. Macmillan, London/New York
Marshall A (1897) The old generation of economists and the new. Q J Econ XI, pp 115–135
Mason ES (1982) The Harvard department of economics from the beginning to world war II. Q J Econ XCVII:383–433
Matthews RCO (1986) The economics of institutions and the sources of economic growth. Econ J 96:903–918
McCloskey DN (1992) Robert William Fogel: an appreciation by an adopted student. In: Goldin C, Rockoff H (eds) Strategic factors in nineteenth century American economic history: a volume to honor Robert W. Fogel. University of Chicago Press, Chicago, pp 14–25

McCloskey DN (2006) The Bourgeois virtues: ethics for an age of commerce. University of Chicago Press, Chicago
McCloskey D [Donald] (1978) The achievements of the Cliometric School. J Econ Hist 38(1):13–28
McCloskey D [Donald] (1985) The rhetoric of economics. University of Wisconsin Press, Madison
McCloskey D [Donald] (1986) Economics as an historical science. In: Parker WN (ed) Economic history and the modern economist. Basil Blackwell, New York, pp 63–70
McCloskey D [Donald] (1987) Responses to my critics. East Econ J XIII(3):308–311
Menard C, Shirley MM (2014) The contribution of Douglass North to new institutional economics. In: Galiani S, Sened I (eds) Economic institutions, rights, growth, and sustainability: the legacy of Douglass North. Cambridge University Press, Cambridge
Menger C (1884) Die Irrthümer des Historismus in der deutschen Nationalökonomie. Alfred Hölder, Vienna
Meyer JR (1997) Notes on cliometrics' fortieth. Am Econ Rev Pap Proc 87(2):409–411
Meyer JR, Conrad AH (1957) Economic theory, statistical inference, and economic history. J Econ Hist 17(4):524–544
Mitch D (2010) Chicago and economic history. In: Emmett RB (ed) The Elgar companion to the Chicago School of Economics. Edward Elgar, Cheltenham/Northampton, MA, pp 114–127
Mitch D (2011) Economic history in Departments of Economics: the case of the University of Chicago, 1892 to the present. Soc Sci Hist 35(2):237–271
Mitchell WC (1913) Business cycles. University of California Press, Berkeley
Nef JU (1941) The responsibility of economic historians. J Econ Hist 1(Supplement: The Tasks of Economic History):1–8
Newmarch W, in collaboration with Tooke T (1857) A history of prices, and of the state of the circulation during the nine years, 1848–56, forming the fifth and sixth volumes of the history of prices from 1792 to the present time, vol 8, London
Nicholas S (1997) The future of economic history in Australia. Aust Econ Hist Rev 37(3):267–274
North DC (1961) The economic growth of the United States 1790–1860. Prentice-Hall, Englewood Cliffs
North DC (1965) The state of economic history. Am Econ Rev Pap Proc 55(2):86–91
North DC (1966) Growth and welfare in the American past: a new economic history. Prentice-Hall, Englewood Cliffs
North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, New York
North DC (1997) Cliometrics – 40 years later. Am Econ Rev 87(2):412–414
Parker WN (ed) (1960) Trends in the American economy in the nineteenth century. Studies in income and wealth, vol 24, conference on Research in Income and Wealth. Princeton University Press, Princeton
Parker WN (1980) The historiography of American economic history. In: Porter G (ed) Encyclopedia of American economic history: studies of the principal movements and ideas, vol 1. Charles Scribner's, New York, pp 3–16
Parker WN (ed) (1986) Economic history and the modern economist. Basil Blackwell, Oxford/New York
Persons WM (1919) An index of general business conditions. Rev Econ Stat 1(2):111–117
Pitkin T (1816) Statistical view of the commerce of the United States. James Eastburn, New York
Polanyi K (1944) The great transformation. Farrar & Rinehart, New York
Purdue University Department of Economics (1967) Purdue faculty papers in economic history, 1956–1966. Richard D. Irwin, Homewood
Redlich F (1965) 'New' and traditional approaches to economic history and their interdependence. J Econ Hist 25(4):480–495
Reinert ES, Carpenter K (2014) German language economic bestsellers before 1850. Working papers in technology governance and economic dynamics no 58

Rogin L (1931) The introduction of farm machinery in its relation to the productivity of labor in the agriculture of the United States during the nineteenth century. University of California Press, Berkeley
Roscher W (1843) Grundriss zu Volesungen uber die Saatswirtschaft nach geschichtlicher Methode. Dieterichsschen Buchhandlung, Göttingen
Seybert A (1818) Statistical annals. Thomas Dobson & Son, Philadelphia
Stoianovich T (1976) French historical method: the annales paradigm. Cornell University Press, Ithaca
Supple B (1965) Has the early history of developed countries any current relevance? Am Econ Rev Pap Proc 55(2):99–103
Taussig FW (1888) Tariff history of the United States. G.P. Putnam's, New York
Tawney RH (1933) The study of economic history. Economica 39:1–21
Temin P (1969) The Jacksonian economy. W. W. Norton, New York
Temin P (2014) Economic history and economic development: new economic history in retrospect and prospect, working paper 20107, NBER working paper series
Temple SW (1672) Observations upon the United Provinces of the Netherlands. Printed for Jacob Tonson, London
Tilly R (2001) German economic history and cliometrics: a selective survey of recent tendencies. Eur Rev Econ Hist 5(2):151–187
Toynbee A (1884) Lectures on the industrial revolution in England: public addresses, notes and other fragments, together with a short memoir. Rivington's, London
Tribe K (2000) The Cambridge Economics Tripos 1903–55 and the training of economists. Manch Sch 68(2):222–248
Turner FJ (1893) The significance of the frontier in American history. American Historical Association annual report. Government Printing Office, Washington, DC, pp 199–227
United States Census Bureau (1960) Historical statistics of the United States, Colonial Times to 1957. US Department of Commerce, Bureau of the Census, Washington, DC
Veblen T (1901) Gustav Schmoller's economics. Q J Econ 16(1):69–93
Weintraub ER (2002) How economics became mathematical science. Duke University Press, London/Durham
Whaples R (1991) A quantitative history of the Journal of Economic History and the Cliometric revolution. J Econ Hist 51(2):289–301
Williamson O (1985) The economic institutions of capitalism. Free Press, New York
Williamson SH (1991) The history of cliometrics. In: Mokyr J (ed) The vital one: essays in honor of Jonathan R. T. Hughes. JAI Press, Greenwich, pp 15–31. REH, supplement 6
Williamson SH (1994) The history of cliometrics. In: Engerman SL et al (eds) Two pioneers of cliometrics: Robert W. Fogel and Douglass C. North, nobel laureates of 1993. The Cliometric Society, Miami
Williamson SH, Whaples R (2003) Cliometrics. In: Mokyr J (ed) The Oxford encyclopedia of economic history, vol 1. Oxford University Press, Oxford/New York, pp 446–447
Wright C (1941) Economic history of the United States. McGraw-Hill, New York

The Contributions of Robert Fogel to Cliometrics

David Mitch

Contents

Abstract

Robert Fogel was one of the earliest and most forceful advocates for the use of quantitative methods and economic theory in the study of economic history and long-term economic change. He demonstrated through his work on the economic impact of the railroads and the economic history of US slavery that the cliometric approach had the potential to challenge and overturn long-standing views based on narrative approaches to economic history. The volume he edited with Stanley Engerman, *The Reinterpretation of American Economic History*, published in 1971 provided an early manifestation to economists and historians alike of the wide range of applications the cliometric approach could offer to various fields of economic history. Throughout his career, Fogel advocated for the cliometric approach to history more generally, not just to economic history.

D. Mitch (✉)
Department of Economics, University of Maryland, Baltimore County, Baltimore, MD, USA
e-mail: mitch@umbc.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_49

His contributions were recognized when he was jointly awarded the Nobel Prize in Economics with Douglass North in 1993. In the subsequent 20 years until his death in 2013, Fogel pursued an interdisciplinary research project focused on long-run changes in the interaction between technological advance, nutrition, human health, and mortality culminating in *The Changing Body* (co-authored with Roderick Floud, Bernard Harris, and Sok Chul Hong).

Keywords
Anthropometrics · Counterfactual · Economic growth · Railroads · Slavery · Social Savings · Technophysio evolution

Introduction: Robert Fogel as a Pioneer of Cliometrics

Robert Fogel was one of the earliest and most forceful advocates for the use of quantitative methods and economic theory in the study of economic history and long-term economic change. He demonstrated through his work on the economic impact of the railroads and the economic history of US slavery that the cliometric approach had the potential to challenge and overturn long-standing views based on narrative approaches to economic history. The volume he edited with Stanley Engerman, *The Reinterpretation of American Economic History* (1971), provided an early manifestation to economists and historians alike of the wide range of applications the cliometric approach could offer to various fields of economic history. He had a penchant for framing his conclusions in a provocative way which generated controversy but also sustained interest in his research areas. And throughout his career, Fogel advocated for the cliometric approach to history more generally, not just to economic history. Moreover, he had a major organizational impact on the practice of economic history through his establishment of a prominent university-based workshop in economic history, his training of a number of students who have gone on to prominent careers in economic history in their own right, and his establishment of the Development of the American Economy project through the National Bureau of Economic Research. His contributions were recognized when he was jointly awarded the Nobel Prize in Economics with Douglass North in 1993.

Fogel's own intellectual evolution presents a striking contrast with North. Both started out as Marxists. Both wrote dissertations focusing on business enterprise, in the case of Fogel, his M.A. thesis on the Union Pacific Railroad, in the case of North, large life insurance companies. However, throughout his scholarly career, Fogel's scholarship was strongly grounded in empirical data and evidence. He was the consummate miner of data and continued to seek new empirical sources throughout. His work maintained clear grounding in economic theory and he served as President of the American Economic Association. In contrast, North was much more conceptually oriented. Increasingly, he reached out to other social science disciplines and had major influences in political science and historical sociology. North never served as president of the AEA, and while some of his early work had empirical grounding,

he was not reliant on archival resources and large data bases as a major component of his scholarship.

Fogel shared with North the view that economic processes were fundamentally historical, but his views as to why this was the case put less emphasis on the role of institutions and political processes and more on the role of technology and biological processes. North, to a greater degree than Fogel, saw economic history as a requisite for improving the discipline of economics, while Fogel put more emphasis on how the tools of economics could remake the field of history. Fogel interestingly identified North as an economic theorist perhaps as much, if not more than, as an economic historian.

Throughout his scholarly career, Fogel saw economic history, and more specifically cliometrics, as inextricably tied to the study of economic growth. His first major impact on the study of economic growth was a response to the claims of W.W. Rostow about the role of leading sectors, such as the railroad, in processes of economic growth. It is sometimes overlooked that his work on slavery began with an attempt to explain the sources of relatively laggard economic growth in the US South. It was this pursuit that led Fogel to investigate the importance of the retarding effects of slavery on growth. His work on anthropometrics was part of longer-run views of the relationship between growth and human physiological development.

Fogel's work often responded to issues of the day. The Civil Rights movement and urban racial unrest in the 1960s, for example, seems to have been a major motivating factor behind his decision to pursue the cliometric study of antebellum Southern slavery.

Fogel's own intellectual interests changed considerably over his scholarly career. While he maintained an interest in what could be labeled the Kuznets tradition of examining long-term economic change, the focus of his research by the time he returned to the University of Chicago in 1981 increasingly focused on what he termed biodemography and health economics.

Robert Fogel's Biography and His Students

Robert William Fogel was born on July 1, 1926 in New York City to parents who were refugees from the Russian Revolution. He attended public school in New York City and then began undergraduate studies in electrical engineering at Cornell University. He changed his major to history with a minor in economics due to his growing awareness of the problems of unemployment in capitalist economies, graduating from Cornell in 1948. During the early 1950s, Fogel was a Communist Party organizer.

Fogel commenced graduate work in economic history at Columbia University in 1956 in order to pursue his interests in the great Marxian questions regarding the nature of long-term economic change. At Columbia, he completed a master's degree under the supervision of Carter Goodrich with a thesis on the Union Pacific as a case of premature enterprise. He also did course work in economics with George Stigler. At Goodrich's encouragement, Fogel enrolled in the doctoral program in economics

at Johns Hopkins University to study quantitative approaches to economic growth under the supervision of Simon Kuznets. He served as an instructor at Johns Hopkins in 1958 and then obtained an appointment as assistant professor of economics at the University of Rochester in 1960. In 1963, he obtained an appointment at the University of Chicago, first as Ford Foundation visiting research professor, completing his dissertation on the railroads and economic growth at Johns Hopkins in 1963.

The Chicago school of Economics associated with the Department of Economics at the University of Chicago is often depicted as ahistorical in its emphasis on rational choice and market processes. In fact by the time Fogel joined that department in 1963, it had a tradition of economic history going back to its founding in 1892. Chester Wright joined the faculty in 1907, John Nef in 1929, and then, to replace Wright, Chicago hired Earl Hamilton in 1947. Fogel was ultimately hired as the replacement for Hamilton, who had receded into the intellectual background of the department by 1963 and was no longer actively teaching economic history. Fogel was recruited both to teach economic history, a requirement for graduate students, and for his work on economic growth and development, an area of increasing interest at Chicago. But more importantly, a number of prominent members of the Department shared the view that economic history was a fundamental part of economics and were eager to recruit someone who would provide it with a more active presence. Fogel was invited to visit the department for the 1963–1964 academic year as the Ford Foundation visiting research professor. In the Fall of 1963, he was offered a tenured appointment as associate professor, and he joined the department on a permanent basis in 1964. According to Milton Friedman, the department was aware of the cliometric approach Fogel took to economic history and this was considered as a plus, less because of a definite commitment to hiring a cliometrician than because it demonstrated Fogel's originality and his promise for being a leader in the field of economic history.

The 1960s was a period of buoyant demand for new faculty with the marked expansion of higher education in the United States. In addition, there was particularly active demand in leading economics departments at this time for practitioners of the new economic history given its recent emergence. Thus, it is no surprise that Fogel was the frequent recipient of offers from other institutions. Coupled with Chicago's recognition that Fogel was one of the leading practitioners of an innovative approach to a field in economics that the department was particularly eager to promote at this time, he was promoted to full professor in 1965, just a year after joining the department as an associate professor. And he was able to command not only an attractive salary but also extensive research support. This included funding to establish the economic history workshop that began to meet in the fall of 1964. Further indication of the department's commitment to economic history at this time was its pursuit of Albert Fishlow, who twice spurned their offers to remain at Berkeley. Fogel's choice of Fishlow as a colleague, despite the potential rivalry given their common dissertation topics and specialties in US economic history, suggests both his high regard for Fishlow and his emphasis on intellectual merit over other considerations in choosing colleagues.

From 1975 to 1976, he was Pitt Professor of American Institutions at Cambridge University and then spent 1976–1981 on the faculty at Harvard. During this time, he played a central role in establishing the Development of the American Economy program affiliated with the National Bureau of Economic Research.

In 1981, he returned to the University of Chicago to succeed George Stigler as Walgreen Professor of American Institutions in the Graduate School of Business, where he established the Center for Population Economics. He also had an affiliation with the Committee on Social Thought at Chicago. In 1993, he was awarded the Nobel Memorial Prize in Economics (shared with Douglass North) for the development of quantitative and theoretical tools for the study of economic history. He never retired from his faculty position and continued to publish until his death in 2013.

At both Chicago and Harvard, Fogel had numerous students who went on to prominent careers in economic history in their own right. Examples include Alice Hanson Jones, Larry Wimmer, Peter Hill, Jacob Metzer, Clayne Pope, Claudia Goldin, Hugh Rockoff, Michael Bordo, Joseph Reid, Frank Lewis, Richard Steckel, David Galenson, Robert Margo, Kenneth Sokoloff, Jonathan Pritchett, Jenny Bourne Wahl, John Moen, John Komlos, Jonathan Pritchett, Dora Costa, and Joseph Ferrie.

The New Economic History: The Role of Theory and Quantification

Robert Fogel is commonly regarded as one of the pioneers of cliometrics (alternatively labeled the new economic history and sometimes referred to as econometric history or historical economics). Indeed, this is mentioned in the Royal Swedish Academy of Sciences press release (1993). In fact, the establishment of cliometrics is commonly dated to a conference in Williamstown, Massachusetts, in September of 1957 jointly sponsored by the Economic History Association and the National Bureau of Economic Research. At the time, Fogel was just starting his doctoral studies at Johns Hopkins. Fogel himself traces the origins of cliometrics even earlier, to advances in history, the social sciences, and mathematics underway before the Second World War (Fogel 1995). One general definition of cliometrics is the study of history using quantitative methods and social science perspectives. Definitions that have been offered of cliometrics have varied since the term was first coined in the early 1960s.[1] While Cliometrics is commonly distinguished from traditional approaches to economic history by the emphasis of the former on quantification and theoretical analysis, other basic distinctions seem to have been at stake in clashes between old and new economic historians. Indeed as has been noted above, pre-cliometric history frequently made extensive use of quantification. At least three

[1]See for example Sutch (1982), McCloskey (1978, 1987), and Williamson (1991). Fogel and Elton (1983), footnote 17 on p. 24 provides an extensive bibliography of pieces surveying the nature of cliometrics as do McCloskey and Hersh (1990). Histories of cliometrics include Williamson (1991) and Drukker (2006).

other basic contrasts, including one by Fogel, have been noted. The first is an emphasis on narration and description in traditional history versus an emphasis on causal explanation and application of formal models in the new economic history (Cochrane 1969; Hacker 1966). The second is a focus on institutions in traditional economic history versus one on processes in the new economic history (Redlich 1965). And the third, suggested by Fogel and Elton (1983, p. 29) is a focus on "specific individuals, on particular institutions, on particular ideas, and on non-repetitive occurrences" in traditional history versus a focus on "collections of individuals, on categories of institutions, and on repetitive occurrences" by cliometricians practicing scientific history.

Initially, in the late 1950s and early 1960s, the history to be studied by cliometrics was perhaps most obviously economic history and the social science perspective most clearly associated with the term was that of economics. However, during this same period, the use of quantitative methods and social science models was becoming more widespread in other fields of history as well. Thus, in 1962, the American Historical Association created an "Ad Hoc Committee to Collect the Basic Quantitative Data of American Political History" (Swierenga 1970). And in 1964, the Mathematical Social Science Board (MSSB) was formed with the joint sponsorship of the Social Science Research Council and the Center for Advanced Study in the Behavioral Sciences. Its stated purpose was "to foster advanced research and training in the application of mathematical methods in the social sciences" (Aydelotte et al. 1972, p. vii). The MSSB took an early interest in applications to history and in 1965 established a Committee on Mathematical and Statistical Methods in History. This Committee was chaired by Robert Fogel and also included Lionel McKenzie (a mathematical economist), Frederick Mosteller (a statistician), William O. Aydelotte (a political historian), Oscar Handlin (an American Historian), and Allan G. Bogue (a political and agricultural historian) (Bogue 1968). One of Fogel's aims in chairing this committee was to show the applicability of mathematical and statistical methods to a variety of historical issues, not just economic history. In a letter to Frederick Mosteller regarding the agenda of an early meeting of the Committee, Fogel stated:

> I view General History Project I as the most important of our potential undertakings . . .For I hope that out of this project will emerge a set of papers that could demonstrate to historians the full range of mathematical and statistical methods—from very simple statistical concepts to highly complex multi-equation systems—that are available for the analysis of historical issues as well as the variety of situations and ways in which such models are applied.[2]

The view that mathematical, statistical, and social science analysis could be applied to a broad range of historical issues was manifested in the establishment in 1975 of the Social Science History Association. That Fogel subscribed to such a broad, social science conception of cliometrics is suggested by his efforts in the early

[2]Letter from Robert W. Fogel to Frederick Mosteller dated April 9, 1965 located in Robert W. Fogel papers, Box 161, in Frederick Mosteller file.

1970s to establish a major graduate program in quantitative and social science history at the University of Chicago, to be called "Committee on Mathematical Methods in History."[3] It is also evident in his statements on cliometrics, which he more broadly termed scientific history (Fogel 1995; Fogel and Elton 1983).

Fogel conceived of the application of quantitative methods to history as entailing not only statistical analysis of empirical data but also the use of formal mathematics. Fogel (1995, p. 54) cites a 1982 American Heritage Dictionary definition of cliometrics as "The study of history using advanced mathematical methods of data processing and analysis." The committee on history that Fogel chaired for the Mathematical Social Science Board, as its title given above suggests, considered the application of mathematical as well as statistical methods to the study of history. In his 1973 letter to Frederick Mosteller regarding an early conference of the committee on history that Fogel organized, summarizes:

> The central objective of the conference then, as I see it, is the presentation of a set of papers which will dispel the notion that mathematical models are a straight-jacket; which will demonstrate that, when properly employed, mathematical models not only provide great flexibility, but greatly extend the range of opportunity for historical analysis.

And his proposed program at Chicago in applying social science and quantitative methods to history was entitled "Committee on Mathematical Methods in History," suggesting that he was prepared to take on board the use of mathematics generally as one of the hallmarks of cliometrics.

In considering what is distinctive about the use of quantification and perspectives from economic theory in the cliometric approach, Fogel emphasized the interrelationship between theory and measurement. He viewed theory as an aid to more effective measurement. Fogel (1966, pp. 7–8) observed:

> The methodological hallmarks of the new economic history are its emphasis on measurement and its recognition of the intimate relationship between measurement and theory. Economic history has always had a quantitative orientation. But much of the past numerical work was limited to the location and simple classification of data contained in business and government records. . .. The pioneers of the massive statistical reconstructions embodied in national income accounts were not economic historians but empirical economists, such as Simon Kuznets in the United States,. . .. While economic historians made considerable use of national income measures, they did not immediately attempt to extend the process of statistical reconstruction to the vast array of issues in their domain.

If in earlier statements Fogel (1964b) put more emphasis on theory rather than measurement as the distinctive trademark of cliometrics, he also conveyed in those statements that the contribution of theory was in enhancing the effectiveness of measurement.

[3]See memor from Fogel to Robert McAdams dated August 13, 1973, re: "A Proposal to establish a Committee on Mathematical Methods in History at the University of Chicago" located in Robert W. Fogel Papers, Box 145, Robert McAdams folder.

While Fogel has been one of the most forceful and persistent advocates for the use of quantification and social science perspectives in the study of history, as has already been noted, he was certainly not the first to do so. Indeed, he was awarded the Nobel Prize jointly with Douglass North, although the two of them were never direct collaborators. This suggests that cliometrics emerged from the efforts of various scholars. However, one tool of the cliometric approach that has been distinctively associated with Fogel is the employment of counter-factual analysis. The basic principle of counter-factual analysis is that in order to determine the impact of some factor, one must consider what would have occurred in the absence of that factor. Applying this principle to historical economic development, Fogel (1967, p. 285) states:

> one cannot determine the economic effects—negative or positive –of the tariff, slavery, the corporation, railroads, the Bessemer converter, the reaper, the telegraph, the Homestead Act, or interregional trade without considering how the economy would have developed in the absence of such institutions, processes, and artifacts. Obviously these counterfactual patterns of American development were not observed and are not recorded in historical documents. In order to determine what would have happened in the absence of a given institution, the economic historian needs a set of general statements that will allow him to deduce a counterfactual situation from institutions and relationships that actually existed.

Fogel did not originate the application of counterfactual analysis to economic history. This was proposed in Conrad and Meyer's methodological essay (1957). And Fogel (1967) himself acknowledged not only the influence of the Conrad and Meyer essay but also noted Fritz Machlup's (1952) discussion of counterfactual reasoning. Following Machlup, one can view counterfactual analysis as no more than systematic analytical reasoning. Or as John Meyer put it in a 40th anniversary retrospective on the 1957 Williamstown conference, "All policy proposals and advocacies are almost by their nature counterfactuals; what would have happened if a proposed policy had been adopted rather than rejected or overlooked" (Meyer 1997, p. 410). Fogel also points to the importance of Hempel's (1942) discussion of the role of general laws in history.

The employment of counterfactual analysis by Fogel in his study of the impact of the railroad on American economic growth provoked accusations by the traditional economic historian, Fritz Redlich, that Fogel's work was "fictitious" and "quasi history" (Redlich 1965, p. 486). And in a later version of his essay, Redlich states his "refusal to consider research based on counterfactual assumptions as genuinely historical research" (Redlich 1968 reprinted in Andreano 1970, p. 92). However, Redlich, in the same passage in this essay, also acknowledges the value of counter-factual analysis. However, he thinks that it should be classified as part of social science research rather than historical analysis: "I do not want to be misunderstood. I do not take a stand against this kind of research per se, nor do I consider it worthless; I only want to have it recognized as part and parcel of the social sciences and to stress its tool character as far as history is concerned" (Redlich 1968 [1970], p. 92).

Fogel has argued in response that not only is counterfactual analysis a useful tool in historical analysis, it is unavoidable if one wants to do any sort of causal analysis

or appeal to general principles. And he states that historians have actually quite frequently employed counterfactuals; they just have not acknowledged it or very fully examined the assumptions entailed in their use (Fogel and Engerman 1971, p. 10; also see Davis 1968, 1970).

Other new economic historians have frequently employed counterfactual analysis. Yet part of the reason why Fogel's work was singled out for attention and criticism was not just his explicit methodological employment of counterfactuals but the degree to which he pursued it in his work on the impact of the railroads. This is suggested by comparing his approach with that of Albert Fishlow's on the same topic. Fishlow also undertook the counterfactual comparison of what transportation costs would have been by water transport in the absence of railways; but in contrast to Fogel, Fishlow based his analysis on the existing canal system in 1859. Fogel instead argued in his analysis of the 1890 situation that in the absence of the railroad there would have been a considerable expansion of the nation's canal system. He then proceeded to propose a hypothetical canal system that might have been constructed by 1890 in the absence of the railway. He found that the presence of such an extended canal system substantially lowered his estimate of the social saving attributable to the railway. Fogel (1979) argues that Fishlow's employment of existing canal systems for his counterfactual analysis was just as hypothetical as Fogel's consideration of a hypothetical extended canal system. It is likely that developing such an elaborate counterfactual canal system earned Fogel's work the title of "figments" from his critic, Fritz Redlich (1965, 1968).

It has sometimes been suggested that the conflict between traditional and new economic history was at least partly generational, reflecting the brashness of "young turk" new economic historians (McCloskey 1985). Simon Kuznets, as Robert Fogel's dissertation supervisor, at times admonished his student for not showing sufficient respect for more traditional economic historians. This is of interest not only for exhibiting generational tensions, and the related ones between student and mentor, but also for highlighting Kuznets's distinctive role as a bridge between traditional and cliometric approaches. Thus, Kuznets wrote to Fogel in a letter dated August 17, 1962 of one passage in a dissertation draft:

> it conveys the unfortunate impression that so many of the traditional economic historians are men of limited understanding and imagination. In revising the text for the final version, I would urge you to go over the text with a fine tooth comb to try to eliminate this impression as much as possible.[4]

And in a letter to Fogel dated May 15, 1963, Kuznets says he wants to:

> repeat my urging to you to. . .edit out some of the statements which impress me as likely to irritate people and only make it more difficult for them to appreciate the value of your analysis. I am referring to statements that contain a generally high claim for the value of econometric analysis and by implication set a low value on more traditional economic

[4]Located in Robert W. Fogel papers, Box 157, Simon Kuznets folder.

> history. . .In general, it is best to let the analysis speak for itself, and err on the side of understating the possible general validity of the approach that you follow.[5]

In a reply to Kuznets dated June 5, 1963, Fogel writes:

> I do agree with you on the need to further edit my manuscript for statements which appear to implicitly repudiate the more traditional economic history. This was not my intention. I think that in history quantitative and qualitative methods of analysis supplement and reinforce each other. I do not view the increasing emphasis being put on the more rigorous use of theory and statistics by younger economic historians as a break with the past; I consider it a further development of a long existing trend in the discipline. At the same time I do not want to weaken my criticism of what I believe has been a serious underestimation of the opportunity that exists for extending quantitative methods worked out in other fields of economics to the domain of history. I also want to emphasize the same point with respect to theory. I realize that my methodological discussions lean heavily on standard principles of economics and statistics. I'm particularly conscious of the fact that many of my arguments are elaboration of points that you have made in various essays and lectures. However, I feel that when linked with specific applications to history, the methodological arguments become more meaningful to those who have previously misunderstood their import.
>
> I have tried to make these points without unduly offending those inclined to loose intuitive methods; but I realize that I have not been entirely successful in this effort.[6]

That Fogel saw continuity between traditional economic history and the new economic history becomes evident when considering his formulation of the contribution of economic history to the discipline of Economics.

The Reinterpretation of American Economic History

An important landmark in the development of cliometrics was the volume edited by Fogel and Engerman and published in 1971, *The Reinterpretation of American Economic History*. The volume consists of 36 essays, including a number by scholars who would not be regarded as economic historians. Among them are Zvi Griliches' article on the diffusion of hybrid corn, of T.W. Schultz on education as capital formation, and Simon Kuznets on the contribution of immigration to labor force growth. It also includes essays by scholars who might be regarded primarily as historians, including James Henretta and Allan and Margaret Bogue. The preface outlines three functions of the volume. Interestingly, the first is "to help teachers of undergraduate courses in American history introduce students to the quantitative revolution in historiography and the far-reaching substantive revisions produced by the new methodology" (Fogel and Engerman 1971, p. xv). The second is to give teachers of undergraduate economics courses materials to demonstrate the real world relevance of economic principles, while providing material for teachers of economic history is only third on the list. The volume opens with a piece by historian Daniel

[5]Located in Robert W. Fogel papers, Box 157, Simon Kuznets folder.

[6]Located in Robert W. Fogel papers, Box 157, Simon Kuznets folder.

Boorstin on "Expanding the historian's vocabulary." The volume perhaps had most impact on teachers of economic history. A projected second edition in 1976 was never realized.

Economic History as the Study of Economic Growth

As previously noted, one likely reason for the Chicago Department's keen interest in economic history during the late 1950s and early 1960s was a serious interest in the determinants of economic growth. Indeed, although Earl Hamilton's own research focused primarily on monetary issues, the research group he headed had the stated topic of the history of growth and development. While the workshop established by Robert Fogel had the more generally stated topic of economic history, Fogel's own research agenda in the mid-1960s saw economic history as primarily focusing on the determinants of economic growth. His work on the economic impact of the railroads was motivated by an interest in the contribution of a generally perceived key innovation to economic growth.

In the early 1960s, the study of economic growth was a common focus of cliometric activity (Drukker 2006). Both Douglass North's 1961 monograph, *The Economic Growth of the U.S.,* and his 1966 textbook survey, *Growth and Welfare in the American Past, A New Economic History,* featured growth in their titles. And North's students, Lance Davis and Jonathan Hughes, along with Duncan McDougall, focused on economic growth in their textbook first published in 1961. Indeed, Davis et al. (1961) discuss at some length in their introduction how, in contrast to the chronological focus typical of most textbooks in American Economic History, theirs is organized thematically, with sections and chapters focusing on key determinants of economic growth. A similar organizational format was featured in the textbook edited by Davis et al. in 1972, *American Economic Growth. An Economist's History of the United States*, which consisted of survey essays on key sectors and dimensions of the American economy.

Fogel began graduate work in economic history with an interest in the determinants of economic growth (Fogel 1994a). While in the midst of work on his dissertation on the railroads in the early 1960s, he had projected both a textbook on American economic history "within the framework of growth economics" and a research monograph entitled *Strategic Factors in American Economic Growth*.[7] Fogel appears to have had a definite conception behind the phrase "strategic factors" and for many years he offered courses alternatively entitled "Strategic Factors in American Economic Development" and "Strategic factors in American Economic Growth." In a memo to Lionel McKenzie, then chair of the Economics department at

[7]See Letter from Fogel to Kuznets dated August 28, 1961, pp. 3–4 located in Robert W. Fogel papers, Box 157, Simon Kuznets folder and Letter to Harold Barger dated July 15, 1963 located in Robert W. Fogel papers, Box 146, Harold Barger folder.

Rochester, dated Feb.1, 1961, Fogel described his view of the nature of strategic factors in economic growth:

> I would like to have the title of Economics 227 changed from Major Factors in American Economic Development to Strategic Factors in American Economic Development. The latter title is the one I originally submitted.
>
> My course attempts to single out and analyse those factors upon which the course of American economic growth depended; i.e., those factors whose absence would have fundamentally altered the record of development. The word "major" means "superior in quality or position"; it connotes only an ordering of importance. A major event need not be one which is capable of altering the design of a given pattern; it can have limited consequences. The development of cheap inland water transportation and the decision not to renew the charter of the bank of the Second Bank of the United States are both generally considered major events in American Economic history. The first was a necessary condition for economic growth during the first half of the nineteenth century; the second was not. The first had strategic consequences on the location of economic activity and the rate of growth; the second did not.[8]

Fogel saw economic history as contributing to an understanding of economic growth and he saw this as the traditional aim of economic history. He did emphasize the importance of using cliometric tools to come up with new answers to traditional questions; he thought that the fruitfulness of the new methods would depend ultimately on whether they led to and supported new interpretations (Fogel and Engerman 1971, p. 2).

Economic history as the study of continuity and change in economic activity can well be seen as entailing far more than just the study of economic growth. Fogel (1965, p. 94) explicitly conflates economic change with economic growth. In citing the historical school economic historian William Ashley, he suggests that Ashley thought that conflict between economics and economic history was avoidable because in Ashley's view "economics proper and economic history focused on different problems: the former on the static properties of modern economies; the latter on the evolution of economic societies – or as we now call it – economic growth."

However, Fogel's teacher at Columbia, Carter Goodrich (1960b, p. 536) noted that "there remain certain points to be considered before economic historians can agree to abdicate in favor of the new discipline of Economic Growth." He argued that "economic historians cannot accept its limitations as to time or to subject matter." He goes on to state that economic life in primitive societies and "issues involving human values and other effects of economic changes" have also been "central themes for economic historians in the past." And Paul David, a leading practitioner of cliometrics, appears to have shared Goodrich's concern when he noted in a letter on Fogel's 1965 reunification article that "it is important, I think, to preserve the distinction which recognizes 'growth problems' as a subset of the problems of secular change."[9]

[8]Located in Robert W. Fogel papers, Box 159, Lionel McKenzie folder.

[9]Letter from Paul David to Robert Fogel dated December 4, 1964 located in Robert W. Fogel papers, Box 149, Paul David folder.

In a 1971 evaluation of quantitative economic history co-authored with Albert Fishlow, Fogel himself acknowledges that the new economic history had emphasized issues of growth and development at the expense of distributional and welfare issues. Fogel lists, among other deficiencies of the new economic history, a focus on growth at the expense of equity and welfare issues (Fishlow and Fogel 1971). However, Fogel also argues for continuing work by economic historians on issues related to economic growth, while acknowledging the importance of other issues in economic history.

Robert Fogel's Substantive Contributions

Robert Fogel made major contributions in three basic areas: (a) estimates of the contribution of the railroad to American economic growth, (b) an examination of various dimensions of the antebellum US southern slave economy, and (c) anthromopetrics and technophysio evolution. Each of these contributions and their legacy on subsequent cliometric research illustrate different aspects of the power of cliometric research. Fogel's project first involved developing an innovative conceptual tool: social savings, and applying it with considerable skill and diligence. The social savings approach, as well as offshoots and reactions to it, have been applied to a wide range of other countries and historical settings. And recent research on the economic impact of the railroads and transportation more generally have employed new conceptual and computational tools to revisit the findings of the social savings approach. While Fogel's second project on slavery began with a focus on estimating the relative efficiency of slave versus free agriculture in the antebellum USA, it expanded into a quite wide-ranging examination of not only the economics of slavery but also of slave society and demography. And while this research made extensive use of the basic theoretical and quantitative tools of economics, narrative accounts, especially in Fogel's second book on slavery, *Without Consent or Contract* (1989), featured prominently as well. Fogel's final project was marked by its interdisciplinary character using methods and concepts from physical anthropology, nutrition, and related health science disciplines.

The Economic History of the US Railroad

Union Pacific as Premature Enterprise

Fogel's first published book on the railroads was *The Union Pacific Railroad: A Case in Premature Enterprise*. It appeared in 1960 and was based on his master's thesis at Columbia, written under the supervision of Carter Goodrich. The book is particularly relevant for understanding cliometrics because it represents a bridge between the "old" and "new" (i.e., cliometric) approaches to economic history. Fogel indicates in the preface that the topic for the book was suggested to him by Carter Goodrich. Goodrich was a more institutionally oriented economist who had serious interests in history and was president of the Economic History Association, but also

was active in various international labor organizations. He also wrote extensively on the role of government in the economy and one of his major books was *Government Promotion of American Canals and Railroads*. Thus, Fogel's book, as a study of a major American business enterprise reflecting a combined role for the government and private enterprise, clearly reflected long-standing interests of Goodrich (See Goodrich 1960a). The term in the book's subtitle, "premature enterprise," was the central one in the book. It has resonance, but contrasts with a parallel term, "building ahead of demand," articulated in Albert Fishlow's study of the antebellum railroad, published around the same time (Fishlow 1965).

The preface of the book indicates both continuities and differences with previous approaches to the history of the Union Pacific by previous historians. He outlines four differences, the first three of which pertain to issues raised by previous historians including (a) indicating that Congressional legislation on the Union Pacific reflected not a flight from a role for government in such enterprises, but instead a persistent commitment to them; (b) new material bearing on the difficulty of raising funds for the enterprise as reflected in new estimates of the market evaluation of the likelihood of failure; and (c) evaluating the wisdom of government involvement by providing new estimates of the social rate of return of the enterprise.

He then describes the fourth difference:

> ". . .it draws on formal economic theory in the determination and analysis of historical facts. Interest theory is combined with the theory of a 'fair game' to deduce, from the market price of the railroad's first mortgage bonds, the market's evaluation of the probability that the Union Pacific would fail. The theory of rent forms the basis for the estimation of the social rate of return on the capital invested in the railroad. The concept of present value is used in the determination of the relative efficiency of the various proposals that were put forth for the financing and construction of a Pacific road." He then goes on to defend the use of counterfactual analysis as articulated by Fritz Machlup (in advance of his 1964 book usually associated with this). He goes on to defend the role of theory "as helpful in the determination of facts as it can be in the explanation of them." (Fogel 1960, pp. 10–11)

Previous historians of the Union Pacific had claimed that the government subsidies of the construction of the railroad in the 1860s resulted in the corruption and excessive profits that characterized the "Gilded Age" more generally. In this study, Fogel undertook to construct detailed accounting measures of the profits earned by the railroad based on primary source material. He employed information related to bond prices to construct estimates of the expected "risk of failure" during the period of construction of the railroad. He then made adjustments in his measured accounting profits for the large measured risk premium his failure risk estimates implied. He thus concluded that claims of exorbitant profits were overstated, while also arguing that the actual mix of public and private sources of finance for the Union Pacific may not have been the most desirable choice. An important contribution of this work in applying economic theory to historical analysis was its utilization of financial models to estimate market risk premiums.

Another important contribution that he was to use more fully in his subsequent book on the railways and economic growth was his use of land rents. He used these

to measure the social benefit of the railroad not captured in the private rate of return to investors. This then allowed him to calculate the excess of the social rate of return to the railroad over and above the private rate of return.

Railroads and American Economic Growth

His doctoral dissertation, and the subsequent book *Railroads and American Economic Growth: Essays in Econometric History* (1964a), challenged the prevailing view that the railroad had a decisive influence on the growth of the American economy. He calculated how much higher the costs would have been to the US economy in 1890 of providing the same level of transportation services with alternative modes of water and land transportation. Fogel's counterfactual methodology proposed a hypothetical canal system that would have been built in the absence of railroads. His estimated "social saving" of the railroad was less than 5% of 1890 US gross national product. His findings spawned numerous challenges by other scholars. Fogel responded by arguing that, for the case of the United States, any plausible allowance for factors raised by critics (such as scale effects and problems of measuring freight rates on rail vs. water traffic) would still imply a modest rather than indispensable contribution to economic growth. However, Fogel also acknowledged that for other economies – such as Mexico, with its more limited access to water transport – the impact of the railroad on growth may well have been substantially larger. His counterfactual methodology generated considerable controversy among historians, with some stating that it was fundamentally ahistorical and fictive. Fogel replied that an analytical and causal approach to economic history inevitably requires the posing of counterfactual questions.

Some of Fogel's findings regarding the impact of the railroad involve relatively straightforward applications of basic price theory. Thus, one of his striking findings concerned the quite small, and by initial crude measures negative, impact of inter-regional railroads on transportation social savings. This was due to direct transport costs by water that were actually lower than rail costs given access to water routes for inter-regional transportation. When allowance is made to seasonal obstacles to water transportation and to the advantages of speed of the railroad over water, the contribution of the railroad increases, but not markedly.

However, his further striking finding was the greater contribution of intra-regional savings from the railroad. He uses two approaches for this calculation. His alpha estimates entail direct cost-savings. However, to allow for spillover effects, as in his Union Pacific study, he goes on to construct beta estimates that use changes in land rents due to railroad access.

Fogel's work initially generated lots of controversy over various dimensions. These include McClelland's (1968) critique of Fogel's (and Fishlow's) empirical evidence and David's (1969) critique of the book's analytical framework. A reprise can be found in Atack and Passell (1994) and Fogel (1979). Fogel (1979, p. 51) notes "some observers of the debate [over the social savings of the railroad] . . .have interpreted the sharp disagreements among the cliometricians as evidence of the failure of social science methodology, and particularly of quantitative methods, in history. There is in this view a confusion between artistic and scientific

processes. . .Scientific creations. . .are usually protracted over long periods, approach perfection quite gradually, and involve the efforts of a large number of investigators. . .[In the case of the social savings controversy] The results of the interaction between the investigators and the critics have been a gradual deepening of the analysis, an improvement of estimating procedures, and the searching out of additional, or more reliable, bodies of evidence bearing on the points at issue. Rather than being a sign of the failure of the cliometric method, the controversy is a sign that the method is working."

Although Fogel's basic work on this topic had been completed by the mid-1960s, his reprise of the controversies and his replies to critics formed his Economic History Association Presidential Address, delivered in 1978 and published in 1979. Over the few decades since, there has been renewed interest in returning to the impact of the railroad. One issue concerns examining the impact of the railroad in a wider range of countries than those considered in the first wave of railroad impact studies. Herranz-Loncan (2006) provides a quite useful survey. Summerhill's (2003, 2005) estimates of the social savings of the railroad for Brazil and Argentina in 1913 are of the same order of magnitude as Coatsworth's (1979) findings of 25% or more for Mexico in 1910. However, a further issue concerns using both considerably enhanced data bases, geographic information system techniques, and more sophisticated modeling tools based on general equilibrium theory to examine the impact of the railroad (Atack 2013; Atack et al. 2010; Donaldson and Hornbeck 2016).

The Study of Industrial Expansion: Antebellum US Iron and Steel

In the spirit of applying neoclassical tools to the study of economic growth, Fogel and his frequent co-author Stan Engerman (Fogel and Engerman 1969) developed a model of the USA's nineteenth-century iron and steel industry to consider the factors behind the ups and downs in its antebellum expansion. In particular, they sought to distinguish between the role of technological advance and tariff protection in influencing the expansion of the industry. They estimated a basic supply and demand model and a Cobb-Douglas production function to assess the relative importance of tariff protection and technological advance in influencing the relative fates of the charcoal and anthracite sectors of the industry. They argue that tariff protection would have facilitated ongoing expansion of the anthracite sector, which was sustained by growth in demand for its relatively crude metal products, such as rails, while this sector received only a modest boost from technological advance. However, they also argue that growing domestic demand would have been sufficient to boost domestic expansion even in the face of foreign competition. In contrast, the more refined charcoal iron sector would have declined in relative terms given the relatively slow growth in demand for its products, even in the face of tariff protection. Fogel and Engerman had anticipated continued work on this project extending the analysis to later time periods and the rise of the US steel industry but abandoned these plans in order to focus on the cliometrics of slavery (Fogel 1996).

The Cliometrics of Slavery

Time on the Cross and Without Consent or Contract

By the late 1960s, Fogel's attention was increasingly dominated by the cliometrics of slavery. Active interest in applying cliometric methods to the economic history of slavery predates Fogel and can be dated to the 1957 conference jointly sponsored by the Economic History Association and the National Bureau of Economic Research. At that conference, Alfred Conrad and John Meyer presented a paper on "The Economics of Slavery in the Ante-Bellum South" in which they found that rates of return on slave plantations throughout the deep South equaled or exceeded that for northern manufacturing industries. Conrad and Meyer's research was published in the *Journal of Political Economy* in 1958 (Conrad and Meyer 1958), and as Fogel notes in biographical memoirs, it was the focus of sustained debate among faculty and students in the Economics Department at Johns Hopkins University when he was a doctoral student there.

Fogel (1975c, p. 667) noted that he and Engerman made the decision in 1968 to conclude the iron industry project described above and "to throw our full energies into the study of slavery" after being struck by the anomalous result that Conrad and Meyer found that slave agriculture was more efficient than free agriculture. This research effort can be situated in both a long-standing historiography on slavery as such, and a long-standing historiography on the extent to which slavery contributed to Southern economic stagnation. With the latter issue, there was continuity with Fogel's previous focus on economic growth.

Fogel and Engerman initially decided in 1968 that a cliometric approach could contribute the most to ongoing debates about slavery among mainstream historians by estimating the relative efficiency of slave versus free agriculture, bypassing previous work on the profitability of slavery, or questions of paternalistic masters, the psychological impact of slavery on blacks, or the role of slavery in dismantling slave family life. Their preliminary efforts using total factor productivity measures to compare the relative efficiency of Southern slave agriculture with Northern free family agriculture yielded the result that slave agriculture was 6% more efficient than Northern free agriculture. They found this result quite surprising, given the presumption of many commentators that slave labor was inherently inefficient. Further adjustments to refine their measure actually increased the relative advantage of slave agriculture to almost 40% (as summarized in Fogel 1996). Subsequent research by Fogel and Engerman (1977) and others (Schaefer and Schmitz 1979; Field-Hendry 1995) considered the role of economies of scale in producing this outcome. Fogel and Engerman (1974, 1977) also focused on the role of the gang system in achieving high levels of labor productivity. They suggest that the gang system achieved high levels of productivity by setting a pace of work that forced all members of the gang to keep up with the most active members.[10] Although their initial calculations were

[10]Also see Toman (2005). For critical discussion of Fogel and Engerman's work see Wright (1979). For more recent surveys of the literature on the efficiency of slavery see Wright (2006) and Sutch's chapter on "African-American Slavery and the Cliometric Revolution" in this volume.

based on the Parker-Gallman sample of slave farms, they subsequently extended the data base and recruited James Faust and Fred Bateman to collect and code a sample of 20,000 Northern farms. As their data collection effort expanded, so did the research questions they considered, including slave demography and the material treatment of slaves.

Their efforts unearthed considerable amounts of new data sources to address a wide range of issues concerning slavery. It also generated considerable amounts of controversy both among cliometricians (see David et al. 1976) and between mainstream historians and cliometricians. The sources of controversy included technical issues, such as in the interpretation of their relative productivity measures of slave versus free agriculture. The controversy extended to the larger narrative about the slave system and imputations that they were defending slavery as a moral system. Further elements of controversy concerned their decision to publish a volume for the general reader accompanied by a volume of supporting technical findings, the speed with which the general volume was produced, and the lack of time allowed for feedback from other scholarly specialists.

After the publication of *Time on the Cross* in 1974, Fogel continued follow-up work addressing critics, and Engerman has continued with wide ranging scholarship covering slavery in a variety of countries. This culminated in Fogel's *Without Consent or Contract: The Rise and Fall of American Slavery,* published in 1989, with several volumes of supporting technical material. Two features noteworthy of this second volume are the 29 page "Afterward" on "the Moral Problem of Slavery," and the extended, almost 200 page, narrative account in Part II on "The Ideological and Political Campaign Against Slavery."

Much of the impetus for the interest in slavery in the late 1960s in the USA stemmed from the civil rights movement of the 1960s. Indeed, Fogel and Engerman (1974, Vol. 2, p. 17), in one of the appendices to *Time and the Cross,* mention "the national tension over race relations" as one reason why discussion became so angry in a 1967 session on "Slavery as an obstacle to Economic Growth." "It must be remembered," they continued, "that 1967 marked the third successive summer in which race riots engulfed American cities with arson, violence, and death." The implicit shift in Fogel's research to a historical topic that resonated with current events may have been of some consequence for the subsequent direction of cliometric work. It implied a shift away from studying the determinants of long-run growth, a topic that would seem to provide strong grounds for incorporating historical analysis into economics. In a 2005 interview with the author, Fogel emphasized that long-run change has always been the focus of his work and that he has never lost that interest. He said that he turned to the issue of slavery in order to understand how institutions affect economic growth (Mitch 2005).

In any event, Fogel certainly saw slavery as an important area for applying the tools of cliometrics and for stating with greater precision the issues formulated by previous historians of slavery.

The Decision to Aim at a Broad Public Audience

However, in his work on slavery, Fogel also sought to reach a much broader public audience than professional economists. Other economists associated with Chicago, including Friedrich Hayek and Milton Friedman, published work aimed at broad audiences. Fogel may have been distinctive for the more focused nature of the work he was trying to bring to public attention. He discussed this decision at some length in three articles published in 1975 (Fogel 1975a, b, c). He notes that recent research on slavery had increasingly touched on larger issues than those of narrow profitability and efficiency on which Conrad and Meyer's initial cliometric research had focused. Fogel explains that ongoing collection of plantation and probate records expanded the scope of issues beyond "such purely economic problems as profitability and efficiency. . .It became clear that our monograph should be broadened to cover such topics as the skill-composition of the slave labour force, the slave family, slave mortality, and slave morbidity" (1975c, p. 670). And "the principle feature of the third phase of cliometric research on slavery is the shift of emphasis from how the slave system worked to the recovery of black history. The onset of the third phase has not brought the second to an end. Rather the two phases coexist, each giving vitality to the other" (Fogel 1975b, p. 42). Finally, he explains his decision to reach out to a larger audience (Fogel 1975c, p. 670):

> We also decided to bring the cumulative findings of more than a decade and a half of cliometric work on slavery to public attention and to do so without waiting for the completion of the research in progress. The decision to write a book for the general public was not an easy one nor a sudden one. . .But it was only when we began to sift through the new data from plantation and probate records that we became convinced that the cumulative weight of cliometric research had reached a critical level. . .Since we viewed the interpretive volume as *initiating* a new debate rather than closing an old one, such a publication schedule seemed reasonable [italics in original]. Naturally, our colleagues would reserve the right to disagree with our findings until they had an opportunity to scrutinize thoroughly the technical procedures. . .Although most comments on our publishing strategy were quite positive, we encountered some sharply negative reactions. . .Still another reader, more perturbed than any of the others with the decision to publish technical findings in such popular form, warned that we were flirting with professional suicide. Stick to the monograph and the technical papers, he advised, adding that if we could not resist the urge to bring our findings to the public, we should write an article for *Scientific American*.

The Fallout from the Slavery Controversy

Nevertheless, to some in the economics profession the intensity of the debate over slavery raised doubts about the credibility of cliometrics (see Heckman 1997). And when Fogel was under consideration for the Walgreen Professorship in 1979, his predecessor in the chair, George Stigler, made a point of sending letters to a number of prominent economic historians both in the USA and England inquiring whether the controversy over *Time on the Cross* had seriously tainted Fogel's scholarly reputation. Their response was an unequivocal no, and that they still regarded Fogel as one of the preeminent economic historians in the world. However, John

Hope Franklin, the prominent African American historian in Chicago's history department, did have some reservations based on the arguments of some of Fogel's cliometric critics. He also noted Fogel's lack of engagement with Chicago's History Department, though he mentioned that he got along well with Fogel personally.

In summing up some two decades of work on slavery, Fogel (1989, p. 13) states:

> I began it like many other cliometricians, not because I was especially interested in the history of American slavery, but because an accident of scholarship made the economics of slavery a major testing ground for the application of cliometric methods. Once drawn into the subject, however, it was the substance of the issues that maintained my interest. Although my principal professional expertise was, and is, in the areas of economics and demography, I found myself led down a road that forced me to grapple with the work of colleagues in cultural, political, and religious history. . .

He concludes that book with an afterward on "The Moral Problem of Slavery" a point to which he returned in a series of published lectures (Fogel 2003, see especially, pp. 45–48; also see Fogel 1994b). Thus, Fogel acknowledges moral considerations that transcend and shape the interpretation of his economic findings, or might be implied by a narrowly defined "Chicago" approach employing rational, maximizing behavior. This interest in moral and social issues that transcend economic considerations comes to the fore in Fogel's 2000 qualitative exposition, *The Fourth Great Awakening and the Future of Egalitarianism*.

Fogel, in an epiphany reminiscent of John Nef, thus came to appreciate that work on the relatively restricted topic of the economics of American slavery had led him to pursue wide ranging cultural, ethical, and political issues. And one of Fogel's faculty appointments at the University of Chicago at the end of his career was with the Committee on Social Thought, established by John Nef. But if Fogel's work on slavery had resulted in a detour from his work on economic growth, it was also to lead to his return to the study of long-run economic change.

Demography, Anthropometrics and Technophysio Evolution

One major aspect of Fogel's research into slavery focused on demographic issues. Fertility and mortality measures provided key indices of the material conditions of slave life and of slave family patterns. A related element focused on nutrition. After publication of *Time on the Cross,* Fogel became aware that international public health specialists were using anthropometric measures (including not only measures of height but also weight and body mass index or BMI – a measure of weight controlled for height) to measure nutritional status of populations in less developed nations. Fogel and his collaborators realized that a variety of the sources they had collected, including, for example, coastal shipping manifests of slaves, reported information on slave heights and could be used for comparing heights of southern slaves with other populations, which in turn could provide evidence bearing on relative nutritional status.

Fogel spent much of his time as Pitt Professor of American Institutions at Cambridge in 1975–1976, reading demography. At this time, he became interested in more fully documenting trends in mortality in North America, with a view to resolving current uncertainty about demographic historians over whether trends in mortality rates during the eighteenth and nineteenth centuries in North America were increasing, decreasing, or level. He thus began a project with the working title "the Economics of Mortality in North America, 1650–1919." His commitment to demographic issues was evident when, upon his return to the University of Chicago in 1981 as the Walgreen Professor of American Institutions, he also established and became Director of the Center for Population Economics. Through his work on mortality, he became increasingly aware of the potential insights that anthropometric measures could provide on nutritional status and the health of populations. This led Fogel to initiate a second large-scale project he called "Secular Trends in Nutrition, Labor Welfare, and Labor Productivity," which aimed to collect numerous data sets on stature, mortality, and related measures for hundreds of thousands of people in North America and Europe. This, in turn, has generated a burgeoning literature using anthropometric measures to examine spatiotemporal patterns in nutrition and its relation to biological well-being and health status. Among other interesting findings generated by this research were the observations of cycles in nutrition as evidenced by cycles in median heights in populations (see Craig 2016 for survey).

Fogel's initial major source for his work on the USA was the pension records of the Civil War Union Army that provides very detailed medical histories of these veterans from the Civil War into the Twentieth century.

During his American Economic Association presidential address, Fogel (1999, p. 2) put forward the concept of "technophysio evolution," which he defined as "the existence of a synergism between technological and physiological improvements that has produced a form of human evolution that is biological but not genetic, rapid, culturally transmitted, and not necessarily stable." Floud et al. (2011, p. 3) set forth the following five building block mechanisms contributing to technophysio evolution:

1. The nutritional status of a generation – shown by the size and shape of their bodies – determines how long that generation will live and how much work its members will be able to do.
2. The work of a generation, measured both in hours, days, and weeks of work and in work intensity, when combined with the available technology, determines the output of that generation in terms of goods and services.
3. The output of a generation is partly determined by its inheritance from past generations; it also determines its standard of living and its distribution and wealth, together with the investment it makes in technology.
4. The standard of living of a generation determines, through its fertility and the distribution of income and wealth, the nutritional status of the next generation.
5. And so on ad infinitum.

What this translates into in long-run trends is that in the last few centuries, the human body has become taller and heavier per unit inch in response to

improvements in the food supply. This in turn led to substantial increases in effective labor supply as improved nutrition provided additional energy for the population to work intensively for many hours per day on average throughout the year.

One important exposition of the implications of this was *The Escape from Hunger and Premature Death, 1700–2100*, *Europe, America, and the Third World* (2004), which argued for nutritional improvements as a key driver of economic growth and improvements in health and declining mortality. Fogel's last major book, co-authored with Roderick Floud, Bernard Harris, and Suk Chul Hong, *The Changing Body. Health, Nutrition, and Human Development in the Western World since 1700* (2011) provides a survey and synthesis of evidence on how the five building block mechanisms interacted and played out in the cases of Britain, Continental Europe, and North America. It has been observed (Margo 2012, p. 542) that while richly interdisciplinary, incorporating work by medical doctors, nutritionists, demographers, statisticians, and historians, economic analysis plays "a supporting role" with not much coverage of "behavioral incentives, or market equilibrium, or the economics of the underlying institutions of food processing and distribution."

Conclusion: Fogel, Kuznets, and the Empirical Tradition in Economics

During his stay at Harvard, Fogel also became involved with the reorganization of the National Bureau of Economic Research and was influential in its Development of the American Economy (DAE) project. Both the DAE project and his work on demography and nutrition implied a focus on long-run change, arguably the basis for integrating economics and economic history. In this way, he followed in the footsteps of his mentor, Simon Kuznets. Kuznets did not define himself as an economic historian (see de Rouvray 2004). Fogel's own subsequent work has focused more on long-run trends in demography than economic history as such. He described his research areas as the bio-demography of aging and health economics rather than economic history (Mitch 2005); all the same, his influence on economic history continues through his students. His pursuit of the inter-relationships between nutrition, health, and human heights was a major impetus behind the rise of the field of historical anthropometrics (Meisel and Vega 2006).

In a 1978 interview with Simon Kuznets, Fogel observed "even though it [NBER project] started out being called the program in Economic History, my own conception of it is as a program in American economic development, with emphasis on trying to get a better picture of the long-term trends that have influenced the development of the economy – to determine whether or not these trends are still at work and what they are."[11]

[11]Travscript of interview located in Robert W. Fogel papers Box 84, NBER program folder, quote is from p. 4, 3/13/78 interview of Robert Fogel with Simon Kuznets.

While teaching at Harvard in the late 1970s, Fogel had continued to label his economic history course "Strategic Factors in American Economic Growth." But on returning to Chicago in the Fall of 1981, he offered a course entitled "Long-term factors in American Economic Growth." He indicates in his lecture notes that in the Winter 1982 Quarter, he would offer a course entitled "Problems in the development of American Economy," aimed at those intending to do research in economic history. The "Long-term factors" course was intended "to provide students interested in issues of current economic policy – interested in the policy issues of the 1980s, with the empirical background they need to adequately assess those issues." He goes on to elaborate, "I have placed an emphasis on long-term factors that affect current policy in order to emphasize that many issues that are today treated as of recent origin are quite long-standing and may not yield to treatment aimed at quick cures."[12]

A defining feature of his modus operandi on his return to Chicago was the research team. Fogel (1975c) noted that cliometrics had "ushered in a new style" featuring the research team made necessary by the large-scale collection and analysis of quantitative data. And he notes at the end of this article (p. 670) that:

> Interaction and cooperation among scholars is not new but it is now being practised on a new scale. The use of numerical evidence was always a feature of historical analysis, but earlier investigators did not have the large grants needed to finance the massive collection of data, nor the hardware required to process them.

And in his 1983 essay on traditional versus scientific history, Fogel notes that "large-scale, collaborative research. . ..[is] a hallmark of cliometric work (Fogel and Elton 1983, p. 61)." His work since the mid-1970s on long-run trends and determinants of nutrition, health, and mortality has featured the utilization of very large data bases, such as Union Army pension records, which has necessarily required large research teams. The theme of big data and use of population level sources was taken up by Fogel's student, Richard Steckel, in his SSHA presidential address (Steckel 2007).

Fogel had already employed a large research team for his work on slavery. His collaborative work with frequent co-author, Stanley Engerman, date back to the 1960s.[13]

Although Fogel achieved renown for his application of economics and quantitative methods to the study of history, throughout his career his research was informed by the view that economic history is indispensable for understanding economic processes and current economic issues. He believed, to quote Schumpeter (1954, p. 12), "the subject matter of economics is essentially a unique process in historic time" (Mitch 2005). His work on the railroads was intended to address a key issue regarding economic growth. His work on slavery was motivated by a desire to use an understanding of the past to inform contemporary social issues. His work

[12]Syllabus and Lecture notes for this course located in Robert W. Fogel papers Box 59.

[13]For Engerman's perspective on this collaboration see Engerman (1992).

on nutrition and mortality focused on the centrality of long-run change (2004). And his presidential address to the American Economic Association argued that the discipline of economics requires historical perspective to come to grips with the problem of accelerating technological change (Fogel 1999).

Fogel's overall assessment was an optimistic one in the tradition of Kuznets. He viewed technological change as a long-run driver of human improvement, which, with interactions with the human body itself, has led to dramatic improvements in life expectancy and well-being. In this he contrasts with North's underlying pessimism, or at least cynicism, about prospects for improvement given inherent difficulties with human institutions. Nevertheless, Fogel did have concerns about spiritual aspects of human experience, considered at length in his 2000 book *The Fourth Great Reawakening,* and one that he returned to in his final volume on the legacy of Fogel et al. (2013). In sum, Robert Fogel's remarkable career and contributions point to both the power and the challenges of pursuing cliometrics.

References

Selected Works by Robert William Fogel (in Order of Publication)

Fishlow A, Fogel R (1971) Quantitative economic history: an interim evaluation: past trends and present tendencies. J Econ Hist 31(1):15–42

Floud R, Fogel R, Harris B, Hong SC (2011) The changing body: health, nutrition, and human development in the Western World since 1700. Cambridge University Press, Cambridge

Fogel R (1960) The Union Pacific Railroad: a case in premature enterprise. Johns Hopkins University Press, Baltimore

Fogel R (1964a) Railroads and American economic growth: essays in econometric history. Johns Hopkins University Press, Baltimore

Fogel R (1964b) Discussion. Am Econ Rev 54(3):377–389

Fogel R (1965) The reunification of economic history with economic theory. Am Econ Rev 55(1/2):92–98

Fogel R (1966) The new economic history: its findings and methods. Econ Hist Rev 19(December):642–656

Fogel R (1967) The specification problem in economic history. J Econ Hist 27(3):283–308

Fogel R (1975a) The limits of quantitative methods in history. Am Hist Rev 80(2):329–350

Fogel R (1975b) Three phases of cliometric research on slavery and its aftermath. Am Econ Rev 65(2):37–46

Fogel R (1975c) From the Marxists to the Mormons. Times Literary Supplement, no 3823, June 13, pp 667–670

Fogel R (1979) Notes on the social savings controversy. J Econ Hist 39(1):1–54

Fogel R (1989) Without consent or contract: the rise and fall of American slavery. Norton, New York

Fogel R (1994a) Autobiography. In: Fransmyr T (ed) Les Prix Nobel. The Nobel Prizes 1993. Nobel Prize Foundation, Stockholm

Fogel R (1994b) The quest for the moral problem of slavery: an historiographic odyssey. The 33rd annual Robert Fortenbaugh memorial lecture. Gettysburg College, Gettysburg

Fogel R (1995) History with numbers: the American experience. In: Etemaud B, Batou J, David T (eds) Pour Une Histoire Economique et Sociale Internationale: Melanges offerts a Paul Bairoch. Editions Passe Present, Geneva

Fogel R (1996) A life of learning. Robert William Fogel. Charles Homer Haskins lecture for 1996. American Council of Learned Societies, New York
Fogel R (1999) Catching up with the economy. Am Econ Rev 89(1):1–21
Fogel R (2000) The Fourth Great Awakening and the future of egalitarianism. University of Chicago Press, Chicago
Fogel R (2003) The slavery debates, 1952–1990. A retrospective. Louisiana State University Press, Baton Rouge
Fogel R (2004) The escape from hunger and premature death, 1700–2100: Europe, America, and the Third World. Cambridge University Press, Cambridge
Fogel R, Douglass C (1997) North and economic theory. In: Drobak J, Nye J (eds) The frontiers of the new institutional economics. Academic, San Diego
Fogel R, Elton G (1983) Which road to the past? Two views of history. Yale University Press, New Haven
Fogel R, Engerman S (1969) A model for the explanation of industrial expansion during the nineteenth century: with an application to the American iron industry. J Polit Econ 77(3): 306–328
Fogel R, Engerman S (1971) The reinterpretation of American economic history. Harper & Row, New York
Fogel R, Engerman S (1974) Time on the cross: the economics of American Negro Slavery. Little Brown, Boston
Fogel R, Engerman S (1977) Explaining the relative efficiency of slave agriculture in the antebellum south. Am Econ Rev 67(3):275–296
Fogel R, Fogel E, Guglielmo M, Grotte N (2013) Political arithmetic. Simon Kuznets and the empirical tradition in economics. University of Chicago Press, Chicago

Archival and Primary Sources

Robert W. Fogel papers. Special collections Research Center. The University of Chicago Library

Published Items

Andreano R (ed) (1970) The new economic history. Recent papers on methodology. Wiley, New York
Atack J (2013) On the use of geographic information systems in economic history: the American transportation revolution revisited. J Econ Hist 73(2):313–338
Atack J, Passell P (1994) A new economic view of American history form Colonial Times to 1940, 2nd edn. Norton, New York
Atack J, Batemen F, Haines M, Margo R (2010) Did railroads induce or follow economic growth?: Urbanization and population growth in the American Midwest, 1850–1860. Soc Sci Hist 34(2): 171–197
Aydelotte W, Fogel R, Bogue A (eds) (1972) The dimensions of quantitative research in history. Princeton University Press, Princeton
Bogue A (1968) United States: the new political history. J Contemp Hist 3:5–27
Bogue A (1990) Fogel's journey through the slave states. J Econ Hist 50(3):699–710
Coatsworth J (1979) Indispensable railroads in a backward economy: the case of Mexico. J Econ Hist 39(4):939–960
Cochrane T (1969) Economic history, old and new. Am Hist Rev 74(5):1561–1572
Cole A, Crandall R (1964) The International Scientific Committee on Price History. J Econ Hist 24(3):381–388

Conrad A, Meyer J (1957) Economic theory, statistical inference, and economic history. J Econ Hist 17(4):524–544
Conrad A, Meyer J (1958) The economics of slavery in the Antebellum South. J Polit Econ 66(2): 95–122
Craig L (2016) Nutrition, the biological standard of living, and cliometrics. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer Reference, Heidelberg
David P (1969) Transportation innovation and economic growth: Professor Fogel on and off the rails. Econ Hist Rev 22(3):506–525
David P et al (1976) Reckoning with slavery: a critical study in the quantitative history of American Negro Slavery. Oxford University Press, New York
Davis L (1968) "And it will never be literature": the new economic history: a critique. Explor Entrep Hist. 2nd series 6(1):75–92. reprinted in Andreano ed. 1970
Davis L, Hughes J, McDougall D (1961) American economic history. The development of a national economy. Richard D. Irwin, Homewood
Davis L et al (1972) American economic growth. An economist's history of the United States. Harper & Row, New York
de Rouvray C (2004) 'Old' economic history in the United States: 1939–1954. J Hist Econ Thought 26(2):221–239
Donaldson D, Hornbeck R (2016) Railroads and American economic growth: a 'market access' approach. Q J Econ 131:799–858
Drukker J (2006) The revolution that bit its own tail. How economic history changed our ideas on economic growth. Aksant Academic Publishers, Amsterdam
Engerman S (1992) Robert William Fogel: an appreciation by a coauthor and colleague. In: Goldin C, Rockoff H (eds) Strategic factors in nineteenth century American economic history. University of Chicago Press, Chicago
Field E (1988) The relative efficiency of slavery revisited: a translog production function approach. Am Econ Rev 78(3):543–549
Field-Hendry E (1995) Application of a stochastic production frontier to slave agriculture: an extension. Appl Econ 27(4):363–368
Fishlow A (1965) American railroads and the transformation of the American Ante-Bellum economy. Harvard University Press, Cambridge, MA
Goodrich C (1960a) Government promotion of American canals and railroads, 1800–1890. Columbia University Press, New York
Goodrich C (1960b) Economic history: one field or two. J Econ Hist 20(4):531–538
Hacker L (1966) The new revolution in economic history. Review essay of railroads and economic growth: essays in econometric history by Robert Fogel. Explor Entrep Hist 3(3):159–175
Heckman J (1997) The value of quantitative evidence on the effect of the past on the present. Am Econ Rev 87(2):404–408
Hempel C (1942) The function of general laws in history. J Philos 39:35–48
Herranz-Loncan A (2006) Railroad impact in backward economies: Spain: 1850–1913. J Econ Hist 66(4):853–881
Machlup F (1952) The political economy of monopoly. Business, labor and government policies. Johns Hopkins University Press, Baltimore
Margo R (2012) Review of the changing body: health, nutrition and human development in the Western World since 1700 by Roderick Floud, Robert W. Fogel, Bernard Harris and Sok Chul Hong. J Econ Lit 50(2):541–543
McClelland P (1968) Railroads, American growth, and the new economic history: a critique. J Econ Hist 28(1):102–123
McCloskey D (1978) The achievements of the cliometric school. J Econ Hist 38(1):13–28
McCloskey D (1985) The rhetoric of economics, 1st edn. University of Wisconsin Press, Madison
McCloskey D (1987) Econometric history. Macmillan, London
McCloskey D, Hersh G (1990) A bibliography of historical economics to 1980. Cambridge University Press, Cambridge

Meisel A, Vega M (2006) Los Origenes de la Antropometrica Historica y su Estado Actual. Cuadernos Hist Econ Empresarial 18(18):2–70
Meyer J (1997) Notes on cliometrics; fortieth. Am Econ Rev 87(2):409–411
Mitch D (2005) Interview with Robert Fogel conducted on August 3
Mitch D (2011) Economic history in Departments of Economics: the case of the University of Chicago, 1892 to the Present. Soc Sci Hist 35(2):237–271
Nef J (1932) The rise of the British coal industry. Routledge, London
North D (1961) The economic growth of the United States, 1790 to 1860. Prentice Hall, Englewood Cliffs
North D (1966) Growth and welfare in the American past. A new economic history. Prentice Hall, Englewood Cliffs
Redlich F (1965) 'New' and traditional approaches to economic history and their interdependence. J Econ Hist 25(4):480–495
Redlich F (1968) Potentialities and pitfalls in economic history. Explor Entrep Hist. 2nd series 6(1):93–108. Reprinted in Andreano ed. 1970
Royal Swedish Academy of Sciences. Press Release: the Sveriges Riksbank Prize in Economic Sciences in memory of Alfred Nobel for 1993. www.Nobelprize.org/
Schaefer D, Schmitz M (1979) The relative efficiency of slave agriculture: a comment. Am Econ Rev 69(1):208–212
Schumpeter J (1954) History of economic analysis. Oxford University Press, New York
Steckel R (1986) A peculiar population: the nutrition, health, and mortality of American slaves from childhood to maturity. J Econ Hist 46(3):721–741
Steckel R (2007) Big social science history. Soc Sci Hist 31(1):1–34
Summerhill W (2003) Order against progress: government, foreign investment, and railroads in Brazil, 1854–1913. Stanford University Press, Stanford
Summerhill W (2005) Big social savings in a small laggard economy: railroad-led growth in Brazil. J Econ Hist 65(1):72–102
Sutch R (1982) Douglass North and the new economic history. In: Ransom R, Sutch R, Walton G (eds) Explorations in the new economic history. Essays in honor of Douglass C. North. Academic, New York
Swierenga R (ed) (1970) Quantification in American history: theory and research. Atheneum, New York
Toman J (2005) The gang system and comparative advantage. Explor Econ Hist 42(2):310–323
Williamson S (1991) The history of cliometrics. In: Mokyr J (ed) The vital one: essays in honor of Jonathan R.T. Hughes. Research in economic history: a research annual supplement, vol 6. pp 15–31. JAI Press, Greenwood, CT
Wright G (1979) The efficiency of slavery: another interpretation. Am Econ Rev 69(1):219–226
Wright G (2006) Slavery and American economic development. Louisiana State University Press, Baton Rouge

Douglass North and Cliometrics

Sumner La Croix

Contents

Abstract

Douglass North (1920–2015) was one of the founders of the disciplines of cliometrics and the New Institutional Economics. He spent over six decades teaching economics and economic history at the University of Washington (1950–1981) and Washington University in St. Louis (1983–2015). In the 1950s and 1960s, North applied neoclassical economic models and quantitative techniques to major problems in US economic history and made significant advances on such topics as interregional trade, ocean shipping productivity, the US balance of payments, and sources of US growth. Switching his attention to European economic history from the late 1960s, North became convinced that economic historians needed to adopt a broader approach to analyzing long-run economic change that explicitly accounted for how economies were embedded in political, economic, and cultural institutions. After the publication of two major books using his new approach, *Structure and Change in Economic History* and *Institutions, Institutional Change and Economic Performance*, North and Robert Fogel were

S. La Croix (✉)
Department of Economics, University of Hawai'i-Mānoa, Honolulu, HI, USA
e-mail: lacroix@hawaii.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_44

awarded the 1993 Nobel Memorial Prize in Economics. Over the next 22 years, North's frame of analysis continued to expand. In his 2005 book, *Understanding the Process of Economic Change*, he argued that economists needed to draw from the sciences of human cognition and social psychology to understand why institutions form and how they change. In his last book, *Violence and Social Orders* (co-authored with John Wallis and Barry Weingast), North made the case that institutions arise in most societies to control the use of violence and are capable of supporting an open political order only in limited circumstances.

Keywords
North · Cliometrics · Institutions · Institutional change · Open access order · Limited access order

Introduction

Douglass Cecil North was born on November 5, 1920, in Cambridge, Massachusetts. His childhood was spent in Canada, England, France, Switzerland, and Connecticut. He graduated from the prestigious Choate private school in Wallingford, Connecticut. His father worked in the insurance industry and would become the president of MedLife, a West Coast insurance company. North graduated with a B.A. from the University of California, Berkeley, where he became a Marxist. A conscientious objector during World War II, he spent the war years as a navigator with the Merchant Marine. After the war, he returned to Berkeley to study economics and worked with Professor M. M. Knight, the brother of the well-known University of Chicago professor, Frank Knight. His dissertation, a study of the history of the life insurance industry in the United States, was completed in 1952. North joined the Department of Economics at the University of Washington in 1950 as an acting assistant professor and would continue as a faculty member in the department through 1983. In 1983, North moved to St. Louis, Missouri, becoming the Henry R. Luce Professor of Law and Liberty in the Department of Economics at Washington University in St. Louis. In 1993, North and Robert Fogel were awarded the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel. He was a founder of the International Society for the New Institutional Economics (now the Society for Institutional & Organizational Economics) and became its second president in 1998. North was a Senior Fellow at the Hoover Institution, which he visited regularly. Many summers were spent in Benzonia, Michigan, and North and his wife, Elizabeth Case, moved there year-round in 2014. He passed on November 23, 2015.

Douglass North's long and creative career was aptly summed up by the Royal Swedish Academy of Sciences when it awarded the 1993 Nobel Prize in Economics to Robert Fogel and North "for having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change." The Academy then caught the essence of both the man and the economic historian by praising him as "an inspirer, a producer of ideas, who

identifies new problems and shows how economists can solve the old ones more effectively" (Royal Swedish Academy of Sciences 1993).

North's influence on the field of economic history was primarily due to his path-breaking research on institutional change but also to the large number of economic historians who he mentored, taught, and inspired at the University of Washington (UW) and Washington University in St. Louis (WUSTL). UW students who were supervised or strongly influenced by North include three undergraduates who became prominent economic historians – Lance Davis, Jonathan Hughes, and Richard Sutch – and numerous Ph.D. students, including Terry Anderson, Lee Alston, Ben Baack, Richard Bean, Gordon Bjork, Philip Coelho, Rodgers Taylor Dennen, Price Fishback, Gerald Gunderson, Sumner La Croix, Robert McGuire, Lloyd Mercer, Ramon Myers, Roger Ransom, Clyde Reed, Gaston Rimlinger, James Shepherd, John Tomaske, Richard Tretheway, Irwin Unger, John Joseph Wallis, Gary Walton, and Robert Willis. A comparable list for WUSTL Ph.D. students includes Lorena Alcazar, Eliana Balla, Mary Ann Boose, Art Carden, Hugo Eyzaguirre, Tawni Hunt Ferrarini, Daniel C. Giedeman, Peter Z. Grossman, Bradley Hansen, Michael Haupert, Shawn Humphrey, Mansor Haji Ibrahim, Iacovos Ioannou, Shilpi Kapur, Philip Keefer, Janice Rye Kinghorn, Jeanine Koenig, Felix Kwan, Noel Johnson, Ruey-Hua Liu, Jeremy Meiners, Michael Munger, Randall Nielsen, Michael J. Orlando, Brian Roberts, Andrew Rutten, Werner Troesken, Mark David Vaughan, and Timothy Yeager.

In an ode to North's graduate teaching, Jonathan Hughes wryly observed that "the charisma emanating from North's lectern was slight indeed" yet then described his graduate seminar as a "place of excitement" (Hughes 1982: 11). Hughes attributed North's effectiveness as a teacher to his constant quest for new ideas and to his instillation of an attitude in his students to strive to be original. Hughes (1982: 10) noted that North always said "something like this" to each of his graduate students:

> If you are one of those who can be original, then it is your duty to the profession to write and publish. Do your work and *never look back*. Don't listen to critics until you've finished. Everyone can criticize; very few can produce new ideas. We must have new ideas; without them the field will die.

And while North constantly challenged students with his critical attitude, Hughes (1982: 11) noted that "[h]e also handed his students a sense of fun and life . . . You were his student, so you were worth something in the world by right."

North's career as a teacher and scholar saw prodigious accomplishments but was also marked by a relentless and restless search for a better analytical framework to analyze long-term change in economies and institutions. In the late1950s, he was one of a small group of economic historians who started the theoretical and quantitative revolution known as "cliometrics," a movement that brought the analytical tools of neoclassical economics and econometrics to economic history. From the mid-1950s to the early 1970s, North used the analytical engine of neoclassical economics to write two books on American economic history (*The Economic Growth of the United States, 1790–1860* and *Institutional Change and American Economic Growth*) and one book on European economic history (*The Rise of the Western*

World). His brilliance is perhaps best seen in his willingness to announce that neoclassical economics would not suffice to understand the issues posed by historical economies and that its analytical tools needed to be augmented with insights from the emerging field of transaction cost economics. In his seminal 1981 book, *Structure and Change in Economic History*, North again challenged economic historians to further broaden the scope of their analytical framework by incorporating explicit theories of the state and ideology. His 1991 book, *Institutions, Institutional Change and Economic Performance*, followed up by providing a more lucid development of specific theoretical concepts and new applications to American and European history. Despite the successful application of his expanded analytical framework, North came to the realization during the 1990s that it was impossible for economic historians and other social scientists to understand processes of economic, political, and social change until they explicitly considered how individuals form their beliefs about the physical and human world. In his 2005 book, *Understanding the Process of Economic Change*, North urged social scientists studying institutions to incorporate insights from evolutionary biology and cognitive science into their models. North's last book (co-authored with John Wallis and Barry Weingast), *Violence and Social Orders*, is unique in that it makes no new calls for an expanded analytical framework yet is strikingly original in its attempts to consider fundamental relationships between violence and the formation of states.

North's Early Career as a Neoclassical Economist and Cliometrician

North's early career at the University of Washington was marked by innovative research in which he staked out bold theoretical positions regarding US growth and assembled important data series critical to understanding US economic growth. Consider four articles published in leading economics and economic history journals between 1955 and 1960. His 1955 article, "Location Theory and Regional Economic Growth," set forth an original explanation of the differential pattern of growth, education, and inequality across the South, Northeast, and Midwest regions (discussed further below). The article was widely read and cited (45 times) by economists and economic historians in the 1950s and cited 1,550 times by 2017. His 1959 article, "Agriculture and Regional Economic Growth," used the framework developed in the "Location Theory" article to criticize arguments put forth by J. Kenneth Galbraith (1951), Theodore Schultz (1953), and Walter Rostow (1956). In North's words, all contended that "[g]rowth is associated with industrialization and stagnation with agriculture." North (1959: 950–951) countered that:

> this misses the whole problem of economic change and reflects a basic misreading of the economic history of the past two centuries. Involvement in the larger market economies, despite the evident hazards entailed, has been the classic way by which regional economies have expanded. It has resulted in specialization, external economies, the development of residentiary industry, and the growth of vertical "dis-integration" as a result of the widening of the market . . .

> . . . the relevant problems of regional economic development revolve around the issues raised in the main body of this paper. They are not issues of agriculture versus industrialization but rather revolve around a region's ability to become integrated into the larger markets of the world through exports, and of the resultant structure of the regional economy which will influence its ability to achieve sustained growth and a diversified pattern of economic activity.

North (1959: 951) then boldly stated that regional development theories predicated on the necessity of achieving industrialization fail to explain "the 19th century economic history of the Midwest from 1815–1860, the Pacific Northwest from 1880–1920, or even California from 1848–1900 . . ."

North's 1958 article on ocean freight rates over the 1790–1913 period provided a short preview of what was to come in his 1968 article on the same topic. He presented a quick look at new data on ocean freight rates for wheat and timber that revealed a long secular decline in rates over the period, interrupted sporadically by increases due to war and capacity bottlenecks. A major contribution of the article was its conjecture that the secular decline in freight costs was due not just to technological change in ship construction but also to external economies that accompanied expansion of particular export markets, to better capacity utilization from the rise of immigration from Europe to the Americas, and to the overall expansion of global trade.

North's 1960 article, "The United States Balance of Payments 1790–1860," was a fundamental contribution to the construction of historical national accounts for the United States. Earlier estimates had presented 12-to-30-year aggregates of the balance of payments, whereas North (1960: 573) was able to construct annual series from 1820 and 5-year moving averages from 1790 through 1819. North (1960: 573) coordinated his "methods and procedures" with Matthew Simon who was conducting a similar project for the US balance of payment series from 1860 to 1900. The publication of both studies in 1960 "provided a consistent and continuous series for the entire period of 110 years" (North 1960: 573; Simon 1960). The new series were anchored in reliable estimates of the merchandise trade balance. Invisible items in the payment calculations were dominated by shipping earnings which after 1815 failed to rise as fast as the expansion in US trade primarily because of falling ocean freight rates. Calculation of the balance of payment series also enabled calculation of an aggregate US foreign debt series. This provided a robustness check for the project as its estimate of US foreign debt compared well with careful estimates of the foreign debt made for particular years (North 1960: 183–184).

North's first book, *The Economic Growth of the United States, 1790–1860*, brought together several themes developed in his earlier articles. He made a point in the preface of emphasizing the role of markets in US growth and downplaying the role of institutions (North 1961: vii): "Institutional and political policies have certainly been influential . . . But they have modified rather than replaced the underlying forces of a market economy." To a modern reader, the book is an intense blizzard of tables and graphs, but they served the purpose of providing basic empirical foundations for North's new ideas regarding patterns of regional growth in the antebellum economy. Two important theses were developed at some length in the book. One thesis was that a necessary (but not sufficient) condition for

regional economic growth to emerge is for an export sector to experience strong growth. A single export sector can suffice to jump-start regional growth, but regional diversification into multiple export sectors provides the path to long-term regional growth. North used this idea to explain why the South's reliance on a single export crop, cotton, led to stagnation of growth vis-à-vis the Northeast and Midwest which had multiple export sectors experiencing strong growth. North's ideas on the contribution of exports to regional economic growth drew heavily from Richard Baldwin's (1956) seminal article on the same topic which, notably, also triggered a decades-long debate among development and trade economists over the same topic.

North's second and more important thesis was that the different types of organizations used in agriculture in the North, Midwest, and South provided the foundations underpinning the different patterns of trade flows, educational achievement, income inequality, and innovation across these regions. Claudia Goldin (1995: 199) aptly summarized this thesis:

> According to Douglass North, the roots of southern stagnation are to be found in the geographic patterns of trade in the antebellum period. The South, using slave labor, grew cotton and exported it to the American North and to Britain. With the receipts from its northern shipments it purchased foodstuffs from the Midwest and industrial goods from the North. With its receipts from European shipments, it purchased luxury items and other industrial wares. Little was ploughed back into the South as internal improvements. Schooling was denied slaves and was poorly provided to southerners in general. Cities, those generators of agglomeration economies, were rare in the South. Innovation was thereby stifled.
>
> The North ran a very different ship. With far more equality of income and wealth, northerners purchased goods produced by local tradesmen and local firms. Its funds were ploughed back into local industry and internal improvements. Its people were the best educated in the world. The North established institutions that served an egalitarian society and that furthered an industrial and growing region. The South had norms that reinforced a caste and race-based society and that inhibited growth at the service of a master class. Such institutions have long lives.

Some scholars have questioned particulars of North's regional trade narrative, noting, for example, that most Southern farms were self-sufficient in food and that Southern purchases of food from the Midwest were probably smaller than depicted by North (Goldin 1995: 199). The overall story told by North still resonates with economic historians, and its straightforward analysis of the different conditions and institutions that emerge from large plantations and small family farms provided the basis for its later elaboration and generalization to the global stage by Stanley Engerman and Kenneth Sokoloff (2000).

Doug North was among the economic historians using cliometric methods who met in 1960 at Purdue University, and annual cliometric conferences quickly became a forum for interchange and discussion of ideas and methods by rising younger scholars and established senior scholars. North and William Parker became co-editors of *The Journal of Economic History* in 1960, and subsequently each issue of the *Journal* began to include one paper that used cliometric methods (Diebolt

and Haupert 2017). The supply of cliometric papers grew in the early 1960s, and North trumpeted this trend in his 1963 *American Economic Review* article that announced the new movement to the broader economics profession. North would continue to be an advocate for cliometrics throughout his career and encouraged those who used cliometric methods to continue to expand the scope of the questions they addressed and to incorporate learning from other disciplines (Libecap et al. 2008).

In 1966, North published a collection of essays, *Growth and Welfare in the American Past: A New Economic History*, which he thought showcased the new economic history. Richard Sutch (1982: 24–25) observed that the essays were designed to introduce both economic history teachers and students to the new economic history. He found that "[i]n each essay, North used simple economic theory and the tabular presentation of quantitative data to challenge views that were widely held at the time . . . On topic after topic North demonstrated how the most elementary economic theory could be used to cast doubt upon venerable interpretations of the past." Perhaps the major contribution of the book was that it stimulated economic historians to think further about the hypotheses put forth on the burden of the Navigation Acts, income growth in the South between 1840 and 1860, public land policies toward railroads, and agrarian discontent, among others. As Sutch (1982: 27) succinctly put it, "*Growth and Welfar*e sent more than a few hares off and running," and "a host of newly converted New Economic Historians enthusiastically gave chase."

The expansion of North's frame of analysis from an economist focused on historical markets to one focused on historical institutions and markets first came into focus with his 1968 publication, "Sources of Productivity Change in Ocean Shipping 1600–1850," in the influential *Journal of Political Economy*. Deidre McCloskey (2010) praised this article as North's "best scientific work," noting its careful attention to how productivity in this industry was improved by reducing transaction costs. The article used state-of-the-art methods to estimate change in productivity for ocean shipping, and it revealed big increases over the period. What was notable about this finding was that most of the gains in productivity were not due to technical changes in shipping and ship construction but rather to "a decline in piracy and an improvement in economic organization." The use of bigger ships increased load factors, the time spent in port declined, and the amount of armament and manpower per average ton of ship declined. North attributed the adjustment in the last factor to the big decline in piracy that occurred over this period. Knick Harley's (1988: 868–869) study of ocean shipping rates took issue with North's findings. He found that declines in shipping rates pre-1850 were primarily due to the integration of a "relatively young American economy" into "the wider European economy" and productivity advances in shipping cotton, while declines in shipping rates after 1850 were primarily due to changes in "mechanical and metallurgical technology" that reduced the costs of constructing and operating ships.

North's focus in his ocean shipping research was ultimately on how improvements in the institutional environment reduced the risk of piracy which, in turn, allowed organizations in the industry to make changes in their structure and operate more efficiently. This research was a harbinger of a recurrent theme in his work, that

institutional change led to changes in the structure and variety of organizations operating in the market and that these organizations often had incentives to effect institutional change to further their own objectives. From the mid-1960s through the early 1970s, North's research agenda focused on how changes in market prices could induce both efficient changes within organizations operating within the market and efficient institutional change. For North, this phase of his research would be relatively short-lived, yet the books that emerged from this agenda would prove to be long-lived, highly influential, and very controversial.

From Cliometrics to Neoclassical Theories of Institutional Change

From the mid-1960s to the mid-1970s, North's research refocused on the role of institutions in retarding or stimulating economic growth. North collaborated with fellow cliometrician Lance Davis to investigate how tools of neoclassical economics could be used to understand long-term institutional change in the American economy. The result was a 1971 book, *Institutional Change and American Economic Growth*, which showed how changes in relative prices had induced efficient changes in the American institutional framework. The late 1960s also marked a switch in North's research interests toward historical European economies. In 1973, North and his University of Washington colleague, Robert Paul Thomas, published a 158-page book, *The Rise of the Western World: A New Economic History*, which provided a panoramic view of 800 years of European history.

The book by Davis and North built upon their earlier co-authored article (Davis and North 1970), "Institutional Change and American Economic Growth: A First Step Towards a Theory of Institutional Change." The centerpiece of this line of research is their theory of efficient institutional change. Efficient institutional change is triggered when an "action group," which could be an organization, a group of organizations, or even an individual, perceives that a change in a particular institution would not only generate gains for them and others but that the aggregate gains would exceed the costs borne by other organizations and individuals. Whether action groups take on the task of advocating and organizing institutional change depends upon how burdensome existing arrangements and laws are, the society's current technology, and the degree of novelty involved in the proposed changes. An effective action group would need to devise arrangements that provide losing groups with some type of compensation for their losses. Compensation reduces opposition to the proposed changes and increases the likelihood that efficient institutional changes are realized. As these changes culminate over time, the result is sustained economic growth.

In the 1971 book, Davis and North apply their theory to central topics in American economic history. Topics covered include land policy and agriculture, organization and reorganization in financial markets, transportation developments and economic growth, economies of scale and unsuccessful cartelization in manufacturing, institutional change in the service industries, organization and education of the labor force, and the changing public-private mix of economic activities. Davis and North were remarkably humble in evaluating their applications, noting

in the introduction that "the explanation at times [are] incredibly simplistic" while proclaiming that the book "represents a first step towards a useful theory of economic growth." Richard Sutch (1982: 34–35) pointed out that Davis and North were themselves aware that their theory could not explain the timing of many institutional changes analyzed in the book (Davis and North 1971: 263). In his review of their book, Alan Bogue (1972: 962) also drew attention to the authors' own perceptions of "the inadequacies" of their theory:

> In their eyes these include among others: the static nature of the model, the problem of tracing inter-sectoral relationships, the fact that profit maximization may be less adequate as an explanation of institutional change than in other aspects of economic process, the paucity of our knowledge about information flows, the fact that redistributive potential may stimulate arrangements of a different nature than the theory would at first suggest and that aspects of the political process are not completely compatible with the model.

Inadequacies aside, North had enough confidence in the theory's explanatory power to showcase it again in his 1973 book with Robert Thomas, *The Rise of the Western World*. In the book's preface, North and Thomas announced that "this is intended to be a revolutionary book" in that they "have developed a comprehensive analytical framework to examine and explain the rise of the Western World; a framework consistent with and complementary to standard neo-classical theory" (North and Thomas 1973: vii). Their analysis of changes in economies and institutions over eight centuries focuses on how population change led to changes in the relative price of land and labor, which, in turn, led to institutional change.

The book "steps into history" in the tenth century "when feudalism and manorialism shaped the society of much of Western Europe" (North and Thomas 1973: 9). Consider now a very brief synopsis of their brief synopsis of eight centuries of complex history. With the breakdown of the order provided by the Carolingian Empire, North and Thomas contended that the institution of the manor arose to cope with the most basic social problem, the control of violence. In their model, fleshed out more fully in an earlier article (North and Thomas 1971), the lord offered to the tenant the protection of the castle as well as land to farm in common fields in exchange for the tenant's fealty. Over the course of the next three centuries, security improved, population slowly increased, and settlement expanded in the frontier areas between villages. In Southern Europe and parts of Northern Europe, cities grew, and as trade increased, new trade networks emerged, along with institutional structures designed to support trade. As European populations expanded, the bargaining power of lords improved. Wages fell, rents increased, and the change in factor prices induced institutional changes that favored lords. A series of famines and epidemics in the fourteenth century culminated with the Black Death and substantially reduced population in Britain and across the continent. The drastic declines in populations led to increases in wages and declines in land rents. Competition among lords for tenants led to gradual changes in manorial institutions that would eventually end serfdom in Western Europe. Declining population also reduced trade, and cities adopted institutional arrangements that were "more 'defensive' in nature." "By the

time population began to grow in the last half of the fifteenth century, the basic structure of a feudal society had given way." Better ships and navigational methods led to the discovery of the new world at the end of the fifteenth century, and the flood of treasure from the new world lubricated the emerging trade between Western Europe with "its centers of skilled trades and manufactures" and Eastern Europe where lands "were still abundant relative to population." Over the next two centuries, changes in military technology occurred that increased the "optimal size of the most efficient military unit." For the larger political units, they needed to raise revenue, and this typically meant that they encouraged trade, as it could be taxed by the ruler. North and Thomas concluded that the search for revenue by emerging nation states led countries down different paths. In France and Spain, monarchs were able to develop a system of taxation that effectively raised revenues while depressing economic growth. By contrast, in Holland, an "oligarchy of merchants" rose to power, while in England, the Glorious Revolution established "the ascendancy of parliament over the crown." In both countries a body of property rights emerged "which promoted institutional arrangements, leading to fee-simple absolute ownership in land, free labor, the protection of privately owned goods, patent laws . . . and a host of institutional arrangements to reduce market imperfections in product and capital markets" (North and Thomas 1973: 17–18).

Criticism of the narratives relayed in *The Rise of the Western World* came from many quarters. Stefano Fenoaltea (1975) presented a trenchant critique of North and Thomas's model of the English medieval manor, a model that provides the foundation for much of North and Thomas's analysis of the tenth-to-thirteenth centuries. In particular, Fenoaltea focused on North and Thomas's argument that "[s]erfdom in Western Europe was essentially a contractual arrangement where labor services were exchanged for the public good of protection and justice" (North and Thomas 1971: 778). Fenoaltea (1975: 387) characterized their "ingenious interpretation" as "empirically unacceptable," noting that their argument "is contradicted by a variety of facts, from the pattern of manorial organization to the technology of medieval warfare." Moreover, "[t]he argument that the 'classic' manorial arrangement of labor services minimized transaction costs also seems erroneous, since North and Thomas appear to misspecify the feasible alternatives (which in fact included indirect barter and market exchange) and misrank the alternatives they consider (as even rents in kind appear superior to labor dues)" (Fenoaltea 1975: 408).

Herman van der Wee's (1975) review of *The Rise of the Western World* extolled the book as offering "precious reward" while criticizing their model of institutional change as far too reductionist: "The reduction of the institutional framework to the problem of property rights in land and man seems to me too narrowly juridical and should be tempered by social, economic, geographical, and cultural points of view" (van der Wee 1975: 238). Van der Wee praised North and Thomas for criticizing the singular emphasis in the Marxist model on the role of technological change in generating institutional change and then sent the same criticism to them for their singular emphasis on population change in generating institutional change. Without explicit discussions of the interdependence of these factors, "both explanations are . . . one-sided" (van der Wee 1975: 238).

Alexander Field (1981) criticized the North-Thomas research agenda for using neoclassical economic principles to explain endogenous institutional variation. Field argued that neoclassical theory is built upon a central assumption that certain elements in a general equilibrium model are exogenous, e.g., "endowments, technologies, preferences, and rules" (Field 1981: 184). He noted that in *The Rise of the Western World*, North and Thomas treat all of the usually exogenous elements as endogenous variables to be explained within their model. Field then argues that "some subset of institutional structures or rules needs to be treated as parametric" in a model of institutional change or the model will not be able to identify all of the phenomena that it attempts to explain.

Despite the extensive criticism, *The Rise of the Western World* was successful in one big way: it refocused the attention of many economic historians on the role of institutions in generating sustained economic growth over long historical horizons. The criticisms and the difficulty in putting together the two books on the European and US economies also forced North to recognize that neoclassical economics was inadequate to the task of explaining institutional change and to explicitly acknowledge that institutional change has not always proceeded efficiently. North's 1978 article, "Structure and Performance: The Task of Economic History," in the *Journal of Economic Literature* was influential because it rejected the idea, used in his 1971 and 1973 books, that the standard tools of neoclassical economics could provide sufficient foundations to explain why institutions arise and how they change over time. North (1978: 963) criticized economic historians for failing to support a historical framework that would enable economists to place contemporary problems in perspective. He admonished that the "[f]ailure of economists to appreciate the transitory character of the assumed constraints and to understand the source and direction of these changing constraints is a fundamental handicap to further development of economic theory." North (1978: 974) then appealed to economists to "go beyond the traditional boundaries of economics to explore political behavior, the growth and application of scientific knowledge, and demographic change." A hint at the direction that North's future research would take came with his call in the article for economic historians to integrate transaction costs into the economic framework used to analyze historical institutions (North 1978: 974–975).

North's subsequent incorporation of transaction cost considerations into his research was partly due to his search for an expanded set of economic tools that could be applied to historical materials and partly due to the intellectual atmosphere in the economics department at the University of Washington that supported vigorous discussion of the role of transaction costs in understanding how markets, firms, and governments were organized. Two colleagues, Yoram Barzel and Steven N.S. Cheung, had both spent time at the University of Chicago and had been influenced by Ronald Coase's use of transaction costs to explain how actual markets and firms were organized. As their stimulating work unfolded during the 1970s, North saw that incorporating transaction costs into his analysis of institutions provided a vehicle for understanding why economic institutions did not always change promptly or efficiently, and from the late 1970s, his research explicitly incorporated transaction cost economics.

In 1982 North and his former doctoral student, John Wallis, published an influential short article that criticized "crude predatory theor[ies] of the state in which government is simply a gigantic transfer mechanism for redistributing wealth and income" (North and Wallis 1982: 336). They emphasized instead that the "wedding of science and technology in the late nineteenth century made possible a technology of production whose potential was only realizable with an enormous increase in the resources devoted to political and economic organization – the transactions sector of the economy. A substantial part of this increase has occurred in the market and through voluntary organization, and a substantial share has also been undertaken by government" (North and Wallis 1982: 336). North and Wallis examined their thesis by dividing government expenditures for 15 OECD countries over the 1953–1974 period into two categories, transfers and transaction services. During this period of strong economic growth, they found that the share of transaction services grew faster than GDP and was relatively stable across countries (North and Wallis 1982: 338).

In a 1986 article, Wallis and North conduct a more in-depth breakdown of the transaction cost components of private and public sectors of the US economy from 1870 to 1970. Because many components of transaction costs are not easily measured, they focused their analysis on measuring transaction services provided by particular industries and by particular occupations (Wallis and North 1986: 103). Banking, insurance, finance, real estate, wholesale services, and retail services are identified as the primary industries providing transaction services, and all of their revenues are counted as transaction services. Occupations such as lawyers, accountants, judges, notaries, police, guards, managers, foremen, inspectors, sales workers, clerical workers, and personnel and labor relations workers are identified as providing transaction services, and wages of these employees in firms outside the transaction service industries are counted as transaction services (Wallis and North 1986: 105–106). Their central finding is that the percent of GNP accounted for by transaction services rose from about 25% in 1870 to over 50% in 1970 (Wallis and North 1986: 120). They identify a number of reasons potentially responsible for the increase: "[I]ncreasing specialization and division of labor; technological change in production and transportation accompanied by increasing firm size; and the augmented role of government in relationship to the private sector" (Wallis and North 1986: 123). They conclude that the main contribution of this line of research is its identification of the "sheer volume" of resources devoted to supporting transactions in US markets.

In a third short article, Wallis and North consider whether the treatment of transaction service sectors in US GNP accounts biases its calculation. They find that excluding final goods in the transaction service sector from GNP reduced the level of GNP in both 1870 and 1970 but had little effect on GNP growth rates (Wallis and North 1988: 653). Their transaction cost quartet concludes in 1994 with a paper examining a classic proposition drawn from Ronald Coase's research: given the existence of a specific technology, firms will choose institutions to minimize transaction costs. North and Wallis expand the inquiry to ask whether Coase's proposition still holds when firms simultaneously choose institutions and technologies. In this broader context, they show that it "is demonstrably false. Institutions will be chosen

that minimize total costs, the sum of transformation and transaction costs, given the level of output" (North and Wallis 1994: 610). The distinction is important because the broader proposition allows for institutional change within a firm to be a vehicle for implementing technical change and "also allows for institutional change to be an important and independent source of growth" (North and Wallis 1994: 610–611).

Expanding the Frame of Institutional Economics

North's 1981 book, *Structure and Change in Economic History*, was completed when he was 60 years old, and it represents a watershed for his research, as it sets forth an agenda for research in global economic history and institutional change that he would pursue for more than three decades. North identified four new elements as necessary building blocks for any viable theory of institutional change. First, he urged economic historians modeling institutional change to reconsider their reliance on models based on neoclassical economics and instead use models that explicitly blended considerations of transaction costs with neoclassical economics. Once this broader framework became North's engine of analysis, then a central implication of earlier models based on neoclassical economics – that institutional change leads to more efficient outcomes – no longer holds. This was a critical change in perspective as it allowed North's new theory of institutional change to encompass a broader range of economic and political outcomes seen in Africa, Asia, and South America as well as a much longer range of human history. Second, North contended that a theory of institutions must be built upon a theory of property rights, but, third, for a theory of property rights to make sense, it needs to be grounded in a theory of the state. A theory of ideology was the fourth new element set forth by North, and it is a necessary element because it is a critical building block in understanding how "different perceptions of reality" lead individuals to respond differently to common changes in the objective environment.

North's model of the state has been highly influential in the social sciences, and this is due in part to its simple definition of a state: "[A]n organization with a comparative advantage in violence, extending over a geographic area whose boundaries are determined by its power to tax constituents" (North 1981: 21). "The basic services that the state provides are the underlying rules of the game" (North 1981: 24). To raise tax revenue, the ruler specifies property rights in man and land that will be at least partly enforced by the state. Typically, there is a trade-off between adoption of more efficient institutions and the ruler's interests, such as raising additional revenue (North 1981: Chap. 3). One of the model's well-known implications is that "the property rights structure that will maximize rents to the ruler (or the ruling class) is in conflict with that that would produce economic growth," and this produces tensions between the interests of powerful groups and the ruler (North 1981: 28).

North was emphatic that institutional change can only be understood when models take into account how ideology frames people's perceptions of reality. Models of institutional change derive rules that govern actions taken by individuals, but which maximizing individuals have incentives to disobey in some situations, i.e.,

to free ride. North (1981: 45–46) pointed out that in other situations, individuals stick to the rules even in instances when they could violate them with little fear of punishment. North then argued that ideology, defined as "intellectual efforts to rationalize the behavioral patterns of individuals and groups," accounts for much of these deviations, as it provides an "economizing device" to simplify decision-making in a complex environment. Ideology "is intrinsically interwoven with moral and ethical judgments about the fairness of the world the individual perceives," and individual ideologies tend to change when "their experiences are inconsistent with their ideology" (North 1981: 48–49). Most importantly, theories of institutional change that fail to incorporate a theory of ideology will flounder in explaining participation of individuals and organizations in politics as well as decision-making by politicians.

North's application of *Structure and Change*'s theory to historical institutions runs just 130 pages and sweeps over 10,000 years of human history, beginning with a model of the transition from hunting and gathering to agriculture and ending with an analysis of the progressive movement in the United States during the early twentieth century. Chapter 7, "The First Economic Revolution," provides a simple theory of the transition to agriculture that uses population pressure to induce simultaneous declines in the populations of migratory animals hunted by competing bands and increases in the formation of private property rights (enforced against outsiders) in land farmed by members of a band. Chapter 8, "The Organizational Consequences of the First Economic Revolution," considers the 8,000 years of the ancient era and focuses on the institutional changes required for societies to make the transition from hunting and gathering to agriculture, the rise of small states, and their consolidation into regional empires. Chapter 9, "Economic Change and Decline in the Ancient World," covers the decline of regional empires and the chaos that enveloped Europe until the end of the first millennium AD, while Chaps. 10 ("The Rise and Decline of Feudalism") and 11 ("Structure and Change in Early Modern Europe") provide a very short (34 pages), somewhat revised rendition of institutional change over the same centuries (900–1700) covered in *The Rise of the Western World*. Chapter 12, "The Industrial Revolution Reconsidered," provides a reinterpretation of the nature and causes of the industrial revolution, while Chapter 13, "The Second Economic Revolution and its Consequences," argues that the scientific advances of the "Second Economic Revolution" enabled the creation of "an elastic supply curve of new knowledge which built economic growth into the system" (North 1981: 171). North concluded that "at the heart of the modern problems of political and economic performance" is the "on-going tension between the gains from specialization and the costs arising from specialization . . ." (North 1981: 209).

Structure and Change was received much more favorably by social scientists and historians than *The Rise of the Western World*, which covered many of the same topics. For example, David Galenson (1983: 189) concluded that "North makes a significant departure from orthodoxy in arguing that a theory of ideology is needed to explain both how collective action can occur in many instances, in spite of the obstacles to it posed by the free-rider problem, and why societies invest resources in

establishing their political legitimacy." Jack Goldstone (1982) and Walter Rostow (1982) applauded the expansion of North's theories and range of historical analysis to cover "a broader canvas," though Rostow concluded that North's "failures all stem from the same source: his lack of a reasonably coherent view of the individual human being." Gordon Tullock (1983) criticized specifics of North's applications but called the book a "rich collection of new and stimulating ideas and perspectives [that] deserve a good deal more space." An outlier in this wave of good feeling was the review by Frederic Pryor (1982: 986–989) in *The Journal of Economic History* who found that North's applications of his theories were often wrong and otherwise were "at such a high level of abstraction that the textures of the economic systems of the past are lost."

North joined with political scientist Barry Weingast to study England's "Glorious Revolution," and the resulting article, "Constitutions and Commitment: The Evolution of Institutions Governing Public Choice in Seventeenth-Century England," is the most highly cited article ever published in *The Journal of Economic History.* This pathbreaking case study changed how a generation of historians and social scientists approached this signal historical event and, more generally, how they analyzed government institutions designed to restrain the power of the executive. North and Weingast argued that the 1688 Glorious Revolution in England was a watershed in the country's institutional history, as it enabled a set of political institutions to be put in place that allowed the revolutionaries:

> to solve the problem of controlling the Crown's exercise of arbitrary and confiscatory power. Parliamentary supremacy, central (parliamentary) control in financial matters, curtailment of royal prerogative powers, independence of the judiciary (at least from the Crown), and the supremacy of the common law courts were established. A major consequence was an increased security of property rights. (as summarized in North 1991: 139)

North and Weingast then argued that the reforms allowed England's capital markets and institutions supporting capital markets, such as the Bank of England, to develop. The government's ability to borrow funds in capital markets was a critical factor in the country's success in two wars with France during the 25 years after the Glorious Revolution. North (1991: 139) later concluded that "[t]he security of property rights and the development of the public and private capital market were instrumental factors not only in England's subsequent rapid economic development, but in its political hegemony and ultimate dominance of the world."

Scholars have questioned the North-Weingast interpretation of the Glorious Revolution, with some asking whether the particular mechanisms that induced Britain's revolution in public finance were correctly identified and others challenging whether their overall interpretation captures the essence of the conflict underlying the Glorious Revolution. Steven Pincus (2009) in his magisterial history, *1688: The First Modern Revolution*, forcefully argued that North and Weingast totally misunderstood the context of the Glorious Revolution. Rather than a battle between forces supporting the monarch's prerogatives and forces supporting a Parliament intent on restraining the monarch and reinforcing individual rights to property and liberty, the

Glorious Revolution should instead be viewed as a battle between two modernizing forces, each with different visions of the polity, society, and economy that would allow Britain to compete effectively with Louis XIV's France (See also Pincus and Robinson 2014).

Others have focused on whether it was the Glorious Revolution that precipitated a financial revolution in Britain or the postrevolutionary threats to William and Mary from the deposed king. John Wells and Douglas Wills (2000: 419) argued that it was the threat from the followers of James II (and later his son) to overturn the Glorious Revolution that "better explains *why* the state introduced these institutional changes than do" North and Weingast. They concluded that much better institutions were required, because the Jacobite threat "exacerbated" the Crown's credibility problems and "both Parliament and the Crown had to develop new institutional arrangements to overcome these credibility problems" (Wells and Wills 2000: 419). Bruce Carruthers (1990: 697) argued that it was James II's "Catholicism more than his absolutism that turned his subjects against him. The wealth holders of England who engineered the Glorious Revolution were more concerned with popery than with property." Carruthers also maintained that it was not just the constitutional changes of the Revolution that enabled Parliament to assert itself more effectively but also the emergence of organized competition between the Tory and Whig parties in Parliament and the influence of a decision by the House of Lords in the Bankers Case, which contested a default of crown debt under Charles II.

North and Weingast then joined with Paul Milgrom on a project that provided a second more extensive case study of the theories outlined in *Structure and Change*. In their 1989 article, they present a formal game-theoretic model that provides theoretic foundations underpinning the "Law Merchant," an institution used by participants in the European Champagne fairs of the early medieval period to enforce contracts made at the fairs. Their analysis considers this essential question: How, "even if no pair of traders come together frequently" at the fairs, can their reputations provide a bond to ensure that the seller provides the promised quality and the buyer pays for the product? Their model finds that "if each individual trades frequently enough within the community of traders, then transferable reputations for honesty can serve as an adequate bond for honest behavior *if members of the trading community can be kept informed about each other's past behavior*" (Milgrom et al. 1990: 3). This article is important, as it served to provide formal game-theoretic foundations for North's early informal theorizing about the linkage between growth in European trade in the eleventh and twelfth centuries and the rise of specific institutions to support the trade.

North's 1990 book, *Institutions, Institutional Change and Economic Performance*, is a short, clearly written volume focused on fleshing out his theoretical framework in a way that could be understood by a wider audience of social scientists and historians. This book is by far the most widely cited of North's works, with Google Scholar listing 57,734 citations as of February 2018. The volume contains relatively little history, although it neatly illustrates various chapters with a fresh set of applications to European and American history. Besides outlining a theory of institutional change, the book also explains "how the past influences the present

and the future, the way incremental institutional change affects the choice set at a moment of time, and the nature of path dependence" (North 1990: 3). Achieving "an understanding of the differential performance of economies through time" is portrayed as the book's "primary objective."

One of the contributions of the book to North's overall analytical framework is that he draws a clear distinction between organizations and institutions. Organizations are defined as "groups of individuals bound by some common purpose to achieve objectives" (North 1990: 5). They will be "designed to further the objectives of their creators," and their structure and actions will be determined by technology, preferences, transaction costs, relative prices, and institutional constraints. North emphasizes that "in the course of attempts to accomplish their objectives," organizations can become "a major agent of institutional change." This can occur because of incremental alteration of the informal constraints as a by-product of maximizing activities of organizations, because organizations become directly involved in the process of institutional change, and because they acquire knowledge that changes costs and benefits associated with existing institutions.

North (1990: 7) draws the reader's attention to his 1984 book, *Structure and Change*, where his adoption of an economic model incorporating transaction costs had led to the abandonment of the proposition that institutions would necessarily evolve efficiently over time. In his 1990 book, he further expands on this idea by identifying more explicit pathways down which institutional change could go awry. His main focus is on organizations that are created to take advantage of opportunities within the institutional framework and then directly or indirectly alter that framework. North (1990: 7–8) hypothesizes that "the symbiotic relationship between institutions and the organizations that have evolved" can lead to lock-in, as these organizations depend on the institutional framework for their profitability. Suppose, however, that "entrepreneurs within economic and political organizations" perceive that they could do better by altering the institutional structure. If transaction costs were zero, only efficient choices would be taken. But transaction costs of changing institutions are positive, and, moreover, "actors frequently must act on incomplete information and process the information that they do receive through mental constructs that can result in persistently inefficient paths." Given this setup, it is easy to find examples of some countries – the United States in the late nineteenth century – where efficient institutional change was undertaken and other countries, England in the early seventeenth century, where institutional change that favored redistributive activities was undertaken.

North (1990: Chap. 11) further develops these ideas in a chapter that emphasizes Paul David's (1985) newly developed concept of path dependence. The basic idea is that "if the process by which we arrive at today's institutions is relevant and constrains future choices, then not only does history matter but persistent poor performance" can be locked in (North 1990: 93). This is most likely when the choice of an institution leads to increasing returns (from a variety of sources) as organizations take advantage of the institution's incentives. The situation that develops is complex, because "ideological beliefs influence the subjective construction of the models that determine choices" (North 1990: 103). Informal institutional constraints,

rooted in a society's culture, can also serve to lock-in institutional choices. North (1990: 140) concludes that "[w]e need to know much more about culturally derived norms of behavior and how they interact with formal rules to get better answers" to questions pertaining to institutional change and persistence.

David Galenson's review (1993: 419–422) in *Economic Development and Cultural Change* finds that the book "provides a synthesis of much of the best of [North's recent work] and serves to point up both its strengths and weaknesses. Much of the most successful work has focused on the political economy of processes by which institutions are formed, and here considerable progress is evident. Much less progress has been made in evaluating the consequences of institutional design." By contrast, Walter Neale (1993: 422), in a review in the same journal, calls it "a strange book, even a sad one." This is because North:

> views institutions as essentially limiting: they are 'the humanly devised *constraints* that shape human interaction' and 'define and *limit* the set of choices of individuals' [North 1991: 3–4; emphasis added by Neale] … One wishes that he had seen his way to adopting the idea that institutions are always and everywhere essential to almost all of human life. Language, rules for raising and sustaining children, drawing a contract, buying a tomato, … are all patterns that we call institutions. Institutions do indeed forbid many activities but … they are always and everywhere liberating as well as limiting. (Neale 1993: 423)

Neale also points out, as did Fredric Pryor in his earlier review of *Structure and Change*, that historical materials used to develop North's grand theories were typically limited to the Americas and Europe, with historical materials from Africa, Asia, and the Pacific often ignored.

Expanding the Horizons of Economists: From Cognitive Science to Political Orders

In his 2005 book, *Understanding the Process of Economic Change*, North begins by forcefully criticizing the rationality assumption of neoclassical economics, arguing that "the uncritical acceptance of the rationality assumption is devastating for most of the major issues confronting social scientists and is a major stumbling block in the path of future progress." To understand how "we perceive the world and construct our explanations about that world requires that we delve into how the mind and brain work – the subject matter of cognitive science" (North 2005: 5). North believes that cognitive science has lessons to teach economic historians about "how humans respond to uncertainty and particularly the uncertainty arising from the changing human landscape, the nature of human learning, the relationship between human learning and belief systems, and the implications of consciousness and human intentionality for the structure that humans impose on their environment" (North 2005: 5–6).

One place to begin incorporating such considerations is to take a close look at "the genetic architecture that evolved in the several million years that humans

evolved as hunters and gatherers. Innate cooperative behavior among small groups does appear to be a genetic trait," and there is some experimental evidence in support (North 2005: 45). We observe, however, that various groups in different societies display substantial variation in the extent and forms of cooperative behavior. North (2005: 47) speculates that these differences may stem from variations in the processes by which individuals learn. They could be due to "(1) the way in which a given belief system filters the information derived from experiences, and (2) the different experiences confronting individuals and societies at different times." North argues that one of the factors affecting individual learning is that many of the important decisions that individuals and groups need to make are in response to novel "non-ergodic" situations, in which important stochastic processes relevant to the decision are unstable. In these "non-ergodic" situations, there is little that decision-makers can learn from past decisions, and this opens the door for players to use nonrational beliefs to make their decisions.

North (2005: viii) emphasizes, however, that "[h]uman learning is more than the accumulation of the experiences of an individual over a lifetime. It is also the cumulative experiences of past generations" which are "embodied in language, human memory, and symbol storage systems [which include] beliefs, myths, [and] ways of doing things." Because every society's culture changes slowly, this affects how successful changes in the formal institutional rules of a society can be in fixing social problems and grasping new opportunities.

The rest of the book sketches out applications of these ideas and many others (e.g., the importance of human intentionality for understanding evolution of institutions) presented in the book to traditional Northian topics, e.g., *The Rise of the Western World*, and new topics, e.g., *The Rise and Fall of the Soviet Union*. North (2005: 169) concludes by observing that "[all] societies throughout history have eventually decayed and disappeared." Substantial evidence suggests that "flexible, adaptively efficient" institutions have been key to the persistence of successful societies, yet he worries that "[t]he ubiquity of economic decline of civilizations in the past suggests that adaptive efficiency may have its limits."

Masahiko Aoki (2010: 139) considers North's 2005 book to be a watershed for research in institutional economics and praises its "dynamic perspective" as "original and comprehensive." Aoki views North as following in a long line of other social scientists who conceived of institutions as "rules of the game" and sees North's originality as following from "the particular attention paid to the crucial importance of enforcement of the rules." Aoki (2010: 241) agrees with North that "the 'normative/ideological beliefs' of political entrepreneurs . . . can be a driving force of institutional change, as such beliefs make a certain direction of change in the human landscape a 'focal orientation.' But it is one thing to say this and another to understand how such beliefs in fact become shared behavioral beliefs . . ." Aoki then attempts to clarify these issues by translating North's informal game-theoretic ideas into the formal language of game theory. Perhaps his most important insight was that while the different economic, political, and social games in which organizations and individuals participate are intrinsically linked, the types of games played in each domain are different, and this limits our attempts to understand their linkages until

we understand each type of game. Aoki (2010: 145) echoes North in his observation that "it is still at an unsatisfactory stage for us to formulate social and political games, if not so much economic games . . . To understand the nature of interactions of play across different domains of games, we first need to be clear about what makes one type of domain essentially distinct from another." Once that is clarified, then in order to understand the process of institutional change, bi-directional relationships between the ways strategic play of the political and economic games are coordinated by people and organizations need to be explored. These relationships may be amenable to explicit game-theoretic analysis in terms of dynamic strategic complementarities, and this would allow a flourishing interdisciplinary field of analysis to emerge.

North's 2009 book, *Violence and Social Orders*, is a creative and ambitious attempt to develop an analytical framework for examining transitions between political orders over a broad range of societies and historical periods. Co-authored with long-time collaborators John Wallis and Barry Weingast, the North-Wallis-Weingast (NWW) team identifies the control of violence as the critical challenge faced by every society. The basic insight is that it is impossible for people in a society to achieve a high level of welfare if they live in an atmosphere charged with frequent outbreaks of violence. If violence reigns, then, in the immortal words of Thomas Hobbes, "life is nasty, brutish, and short." While no society can ever eliminate violence, a society can only prosper if violence is "controlled and managed" (North et al. 2009: 13). Societies living in the shadow of violence try to solve this fundamental problem by devising social arrangements to "deter the use of violence by creating incentives for powerful individuals [and their supporters] to coordinate rather than fight." In the NWW conceptual framework, there are two general ways in which societies can be *ordered* to control violence. In a *limited access order*, violence is primarily controlled by devising formal rules that limit people's rights to form new organizations which could compete with established organizations. In an *open access order*, violence has been largely restricted to the military and police, which are controlled by well-established lines of authority in the government. Political institutions support open entry of new organizations to compete in economic and political markets with organizations run for and by established interests (North et al. 2009: 4).

What is the mechanism used in a limited access order to control the use of violence (North et al. 2009: 4)? Leaders of powerful groups and powerful individuals first agree to form a "dominant coalition" whose members divide up resources and opportunities among themselves and agree to maintain each other's privileged access to those resources. Such privileges generate economic rents (defined as a premium above and beyond the normal returns to an activity) and will, if incentives are properly structured, ensure that each leader and their group keep the peace rather than fight. Because self-enforcing agreements between the members of the dominant coalition often break down, the dominant coalition acts to provide an "organization of organizations" – often called "the state" – which provides third-party enforcement of privileges. If individuals or groups inside or outside the dominant coalition fail to respect member privileges, then individual members of the coalition and the state

have incentives to take action to enforce member rights. In fact, to maintain their exclusive access to economic rents, the key action that members of the dominant coalition must take is "to *limit* the possibility for others to start rival organizations."

How are the economic rents created by the dominant coalition distributed to its members? Distribution is heavily influenced both by the violence potential of powerful individuals and organizations and by established networks of unique personal, family, and group relationships. Networks are important for rent distribution because most of the relationships between elites in the dominant coalition are personal, not impersonal. Because important relationships among elites are personal in nature, enforcement of agreements between members of the elite depends on the identity and status of the particular member of the elite rather than impersonal rules of law. Use of impersonal rules to enforce agreements between people of different status is impossible, and this means that agreements requiring high levels of trust can only be concluded by people from the same social group.

NWW spend the rest of the book elaborating on their theory of social orders and providing particular examples of different types of limited access and open access orders. Additional chapters cover the emergence of three different types of limited access orders, namely fragile, basic, and mature natural states; the development of English land law; institutions, beliefs, and incentives supporting open access orders; doorstep conditions underpinning the transition to an open access order; and examples of successful transitions in Britain, France, and the United States. In their reviews of *Violence and Social Orders*, Robert Margo (2009) and Robert Bates (2010) praise the book for its originality and criticize it as being too devoted to providing taxonomic classifications of social orders rather than providing deeper micro-foundations of their origins, operation, and evolution. Margo identifies the book's most important and profound idea as "the notion that one cannot simply 'get rid' of the superficial exterior of natural states and thereby uncover the beating heart of an open access order yearning to be free . . ."

In 2012, the NWW team (with Stephen Webb) edited a second volume, *In the Shadow of Violence*, that provides case studies of economic and institutional change in ten developing countries using the analytical framework developed in *Violence and Social Orders*. Critics of North have noted his reliance on evidence from Europe and the Americas to illustrate his theories, and the NWW team commissioned this second volume to show the broad range of historical experiences that could be analyzed with their analytical framework.

Do Institutions Matter? North and His Critics

Critics of North have come at him for a variety of sins. Economists with ties to the Wisconsin version of institutionalism, pioneered in the writings of John Commons (1934) and Thorstein Veblen (1899; 1904), have viewed North's work as fatally flawed because his theoretical framework, however suitably modified to incorporate transaction costs, continued to have one foot in the silo of dysfunctional neoclassical economics. Ben Fine and Dimitris Milonakis (2003: 568) take a slightly different

tack, with their criticism of North stemming from his continued reliance on methodological individualism. They conclude that "North's own intellectual voyage, despite setting out under the mast of individualism, eventually leads him to the troubled waters of collectivities, power, and conflict" which his theories accommodate poorly.

North's definition of an institution, "the rules of the game in a society or, more formally, . . . the human-devised constraints that shape human interaction," and his emphasis on the importance of institutional change in generating economic growth have been criticized by several prominent economic historians and theorists, including Avner Greif, Deidre McCloskey, Joel Mokyr, and Gregory Clark. Brief summaries of their critiques follow.

In a series of pathbreaking studies analyzing institutions with the tools of game theory, Avner Greif (2006) contends that institutions are best defined as the expectations of individuals regarding "the regularity of others' behavior." Greif and Christopher Kingston (2011: 25) summarize "the institutions-as-equilibria approach" as follows:

> The core idea in the institutions-as-equilibria approach is that it is ultimately the behavior and the expected behavior of others rather than prescriptive rules of behavior that induce people to behave (or not to behave) in a particular way. The aggregated expected behavior of all the individuals in society, which is beyond any one individual's control, constitutes and creates a structure that influences each individual's behavior. A social situation is 'institutionalized' when this structure motivates each individual to follow a regularity of behavior in that social situation and to act in a manner contributing to the perpetuation of that structure.

Greif's definition and analysis of institutions differ from North's approach, but both economic historians view institutions as fundamental to understanding how economies change and why some have been able to generate sustained growth. Several prominent economic historians have, however, tried to make the case that better institutions were *not* the main factor behind the surge of economic growth in the nineteenth and twentieth centuries.

In Volume II of her three-volume opus, *Bourgeois Dignity: Why Economics Can't Explain the Modern World*, Deidre McCloskey (2010) takes several chapters to criticize both the methodological underpinnings and the usefulness of North's scholarship. McCloskey contends that despite the expanding frame of analysis embraced by North over his career, his analysis remains burdened by continued reliance on outdated concepts from neoclassical economics – profit maximization by firms subject to constraints and utility maximization by consumers subject to constraints. North is accused of being fixated on constraints and failing to understand how noneconomic dimensions of culture, religion, and politics affect human behavior. Within the terms of North's own analytical framework, McCloskey criticizes North's narratives explaining how changes in institutions in Britain and other countries triggered modern economic growth. She argues that the actual timing of changes in institutions and changes in economic growth is poorly aligned with North's theoretical predictions. North's emphasis on the emergence of a more credible commitment to property rights in Great Britain after the Glorious Revolution is dismissed, with

McCloskey arguing that strong property rights had been well established in Britain for several centuries and that, in any case, the English Civil War was more important for the development of modern British institutions than the Glorious Revolution. McCloskey then contends that the last two centuries of economic growth were made possible by the development of bourgeois values and the rising respect gained for those values during the seventeenth and eighteenth centuries.

Joel Mokyr (2017: 5) observes that "much of the literature in economic history that is trying to explain differences in economic performance and living standards . . . has accepted in one way or another Douglass North's call for the integration of institutions into our narrative of economic growth . . .." Mokyr (2017: 5) agrees that better institutions "can be credited with many positive economic developments in the past" and yet "better markets, more cooperative behavior, and more efficient allocations do not in themselves account for modern economic growth." He concludes that the Industrial Revolution "[a]t first blush does not seem to have been a response to any obvious institutional stimulus." Rather it "was made possible by cultural changes" that affected man's "attitudes towards the natural world" and made possible institutions that "stimulated and supported the accumulation and diffusion of 'useful knowledge'" (Mokyr 2017: 6–7).

Gregory Clark (2007) argues that neither institutions nor geography nor exploitation of colonies can explain the last two centuries of economic growth. Rather, the long-established culture of a society combines with demographic forces and chance to determine long-term economic growth. From Clark's perspective, institutions are relatively unimportant because they are endogenously determined. "Thus institutions vary across time and places mainly because differences in technology, relative prices, and people's consumption desires make different social arrangement efficient" (Clark 2007: 212). Clark sees a close correspondence between his view of institutions and the neoclassical theory of institutions espoused by North and Thomas in *The Rise of the Western World*, where changes in relative prices and other economic factors drive institutional change and institutions tend, over a long horizon, to be efficient.

North's Legacy

For Doug North, longevity and creativity paid off. Some of his impact was due to an unusually long and extremely productive career that spanned 65 years. The rest was due to a creative mind that challenged economic historians to address questions that were typically out of their comfort zone. The entire range of human history was within the scope of his research, and by the end of his career, he had completely changed how economic historians perceived their mission and vastly expanded the analytical framework used to analyze long-term change in economies and institutions.

From the early 1990s, North had become increasingly interested in how different people could have different perceptions of the same reality. In this context it is notable that different scholars perceive North and his work very differently. Oliver

Williamson (2000: 600), one of the founders of the New Institutional Economics and the recipient of the 2009 Nobel Prize in Economics, identifies North as one of the key people whose work and organizational activities led to the founding of the New Institutional Economics, standing alongside other Nobel Laureates such as Ronald Coase, Kenneth Arrow, Friedrich Hayek, Gunnar Myrdal, and Herbert Simon. Many economic historians see North, along with Robert Fogel, John Meyer, Robert Gallman, William Parker, and Lance Davis, as one of the founders of cliometrics, the theoretical and quantitative movement that transformed research in economic history. Other economists perceive North as a scholar who constantly hectored them to escape from the narrow silo of neoclassical economics and to consider how their analysis could be enhanced by incorporating insights from other disciplines.

Perhaps the most perceptive views of North come from scholars who recognize how radically he revised his views on institutional change over the course of his more than six-decade career. A look at his first book provides an extremely different view of economic change than a look at his last book, which focuses on the foundations of social orders and how they change. Claude Menard and Mary Shirley (2014: 3) perceive that his big achievement was to change "many economists' view of development from a process of growth spurred by new technology and capital accumulation to a dynamic process of institutional change." And John Wallis (2014: 48) concludes that "North's genius is figuring out what to ask next, which often comes as an answer to the question of what cannot be explained with the current conceptual framework." Asking new questions and developing new ways to answer them lie at the heart of all scientific inquiry, and posing new, extremely challenging questions may well prove to be North's fundamental contribution to history and the social sciences.

Cross-References

► Institutions
► The Contributions of Robert Fogel to Cliometrics

References

Selected References by Douglass C. North (In Order of Publication)

Cox GW, North DC, Weingast BR (2015) The violence trap: a political-economic approach to the problems of development. 13 Feb 2015. [cited on 27 January 2018]. Available from: SSRN: https://ssrn.com/abstract=2370622

Davis LE, North DC (1970) Institutional change and American economic growth: a first step towards a theory of institutional innovation. J Econ Hist 30(1):131–149

Davis LE, North DC, with assistance from Smorodin C (1971) Institutional change and American economic growth. Cambridge University Press, Cambridge

Denzau A, North DC (1994) Shared mental models: ideologies and institutions. Kyklos 47(1):3–31

Denzau A, North DC, Roy RK (2005) Shared mental models: a postscript. In: Roy RK, Denzau AT, Willet TD (eds) Neoliberalism: national and regional experiments with global ideas. Routledge, London/New York
Drobak JN, North DC (2008) Understanding judicial decision-making: the importance of constraints on non-rational deliberations. Wash Univ J Law Policy 56:131–152
Milgrom PR, North DC, Weingast BR (1990) The role of institutions in the revival of trade: the law merchant, private judges, and the Champagne fairs. Econ Polit 2(1):1–23
North DC (1955) Location theory and regional economic growth. J Polit Econ 63(3):243–258
North DC (1956) International capital flows and the development of the American West. J Econ Hist 16(4):493–505
North DC (1958) Ocean freight rates and economic development 1730–1913. J Econ Hist 18(4):537–555
North DC (1959) Agriculture and regional economic growth. J Farm Econ 41(5):943–951
North DC (1960) The United States balance of payments 1790–1860. In: Parker WN (ed) Trends in the American economy in the nineteenth century, 24th conference on income and wealth, National Bureau of Economic Research. Princeton University Press, Princeton
North DC (1961) The economic growth of the United States, 1790–1860. Prentice Hall, Englewood Cliffs
North DC (1963) Quantitative research in American economic history. Am Econ Rev 53(1/Part 1):128–130
North DC (1965) The state of economic history. Am Econ Rev 55(1/2):85–91
North DC (1966) Growth and welfare in the American past: a new economic history. Prentice-Hall, Englewood Cliffs
North DC (1968) Sources of productivity change in ocean shipping 1600–1850. J Polit Econ 76(5):953–970
North DC (1978) Structure and performance: the task of economic history. J Econ Lit 16(3):963–978
North DC (1981) Structure and change in economic history. Norton, New York
North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
North DC (1991) Institutions. J Econ Perspect 5(1):97–112
North DC (1994) Economic performance through time. Am Econ Rev 84(3):359–368
North DC (2005) Understanding the process of economic change. Princeton University Press, Princeton
North DC. Douglass C. North -biographical. [cited 28 January 2018]. Available at https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1993/north-bio.html
North DC, Thomas RP (1971) The rise and fall of the manorial system: a theoretical model. J Econ Hist 31:977–803
North DC, Thomas RP (1973) The rise of the Western world: a new economic history. University Press, Cambridge
North DC, Wallis JJ (1982) American government expenditures: a historical perspective. Am Econ Rev Pap Proc 72(2):336–340
North DC, Wallis JJ (1994) Integrating institutional change and technical change in economic history: a transaction cost approach. J Inst Theor Econ 150(4):609–624
North DC, Weingast B (1989) Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. J Econ Hist 49(4):803–832
North DC, Alston L, Eggertsson T (eds) (1996) Empirical studies in institutional change. Cambridge University Press, New York
North DC, Summerhill W, Weingast BR (2000) Order, disorder, and economic change: Latin America vs. North America. In: Root H, de Mesquita BB (eds) Governing for prosperity. Yale University Press, New Haven
North DC, Wallis JJ, Weingast BR (2009) Violence and social orders: a conceptual framework for interpreting recorded human history. Cambridge University Press, New York
North DC, Wallis JJ, Webb SB, Weingast BR (eds) (2012) In the shadow of violence. Cambridge University Press, New York

Wallis JJ, North DC (1986) Measuring the transaction sector in the American economy. In: Engerman S, Gallman R (eds) Long term factors in American economic growth. Studies in income and growth, vol 51. University of Chicago Press, Chicago
Wallis JJ, North DC (1988) Should transaction costs be subtracted from gross national product? J Econ Hist 48(3):651–654

Other Selected References

Aoki M (2010) Understanding Doug North in game-theoretic language. Struct Chang Econ Dyn 21(2):139–146
Baldwin RE (1956) Patterns of development in newly settled regions. Manchester Sch Econ Soc Stud 24(May):161–179
Bates RH (2010) A review of Douglass C. North, John Joseph Wallis, and Barry R. Weingast's violence and social orders: a conceptual framework for interpreting recorded human history. J Econ Lit 48(3):752–756
Bogue AG (1972) Review: institutional change and American economic growth by Lance Davis, Douglass C. North with assistance from Calla Smorodin. J Econ Hist 32(4):961–962
Carruthers BG (1990) Politics, popery, and property: a comment on North and Weingast. J Econ Hist 50(3):693–698
Clark GA (2007) Farewell to alms: a brief economic history of the world. Princeton University Press, Princeton/Oxford
Commons JR (1934) Institutional economics. Macmillan, New York
David PA (1985) Clio and the economics of QWERTY. Am Econ Rev 75(2):332–337
Diebolt C, Haupert M (2017) A cliometric counterfactual: what if there had been neither Fogel nor North? Cliometrica; https://doi.org/10.1007/s11698-017-0167-8
Engerman SL, Sokoloff KL (2000) Institutions, factor endowments, and paths of development in the new world. J Econ Perspect 14(3):217–232
Fenoaltea S (1975) The rise and fall of a theoretical model: the manorial system. J Econ Hist 35(2):386–409
Field AJ (1981) The problem with neoclassical institutional economics: a critique with special reference to the North/Thomas model of pre-1500 Europe. Explor Econ Hist 18(2):174–198
Fine B, Milonakis D (2003) From principle of pricing to pricing of principle: rationality and irrationality in the economic history of Douglass North. Comp Stud Soc Hist 45(3):120–144
Galbraith JK (1951) Conditions for economic change in underdeveloped countries. J Farm Econ 33(4/Part 2):689–696
Galenson DW (1983) Review: structure and change in economic history by Douglass C. North. J Polit Econ 91(1):188–190
Galenson DW (1993) Review: institutions, institutional change and economic performance by Douglass C. North. Econ Dev Cult Chang 41(2):419–422
Goldin C (1995) Cliometrics and the nobel. J Econ Perspect 9(2):191–208
Goldstone J (1982) Review: structure and change in economic history by Douglass C. North. Contemp Sociol 11(6):687–688
Greif A (2006) Institutions and the path to the modern economy: lessons from medieval trade. Cambridge University Press, New York
Greif A, Kingston C (2011) Institutions: rules or equilibria? In: Caballero G, Schofield N (eds) Political economy of institutions, democracy and voting. Springer, Berlin
Harley CK (1988) Ocean freight rates and productivity, 1740–1913: the primacy of mechanical invention. J Econ Hist 48(4):851–876
Hughes JRT (1982) Douglass North as a teacher. In: Ransom R, Sutch R, Walton G (eds) Explorations in the new economic history: essays in honor of Douglass C. North. Academic, San Diego

Libecap GD, Lyons JS, Williamson SH, interviewers (2008) Douglass C. North, further reflections. In: Lyons JS, Cain LP, Williamson SH (eds) Reflections on the cliometrics revolution: conversations with economic historians. Routledge, New York

Margo R (2009) Review: Douglass C. North, John Joseph Wallis, and Barry R. Weingast's violence and social orders: a conceptual framework for interpreting recorded human history. EH.NET. [cited on 27 January 2018]. Available at http://eh.net/book_reviews/violence-and-social-orders-a-conceptual-framework-for-interpreting-recorded-human-history

McCloskey DN (2010) Bourgeois dignity: why economics can't explain the modern. University of Chicago Press, Chicago

Menard C, Shirley MM (2014) The contribution of Douglass North to new institutional economics. In: Galiani S, Sened I (eds) Institutions, property rights, and economic growth: the legacy of Douglass North. Cambridge University Press, New York

Mokyr J (2017) Culture of growth: the origins of the modern economy. Princeton University Press, Princeton/Oxford

Neale W (1993) Review: institutions, institutional change and economic performance by Douglass C. North. Econ Dev Cult Chang 41(2):422–425

Pincus SCA (2009) 1688: the first modern revolution. Yale University Press, New Haven

Pincus SCA, Robinson JA (2014) What really happened during the Glorious Revolution? In: Galiani S, Sened I (eds) Institutions, property rights, and economic growth: the legacy of Douglass North. Cambridge University Press, New York

Pryor FL (1982) Review: structure and change in economic history by Douglass C. North. J Econ Hist 42(4):986–989

Rostow WW (1956) The takeoff into self-sustained growth. Econ J 66(261):25–48

Rostow WW (1982) Review: structure and change in economic history by Douglass C. North. Bus Hist Rev 56(2):299–301

Royal Swedish Academy of Sciences (1993) Press Release, Oct. 12, 1993. [cited on 18 July 2018] Available at https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1993/press.html

Schultz T (1953) The economic organization of agriculture. McGraw-Hill, New York

Simon M (1960) The United States balance of payments, 1861–1900. In: Parker WN (ed) Trends in the American economy in the nineteenth century, 24th conference on income and wealth, National Bureau of Economic Research. Princeton University Press, Princeton

Sutch R (1982) Douglass North and the new economic history. In: Ransom R, Sutch R, Walton G (eds) Explorations in the new economic history: essays in honor of Douglass C. North. Academic, San Diego

Tullock G (1983) Review: Douglass C. North, structure and change in economic history. Public Choice 40(2):233–234

van der Wee H (1975) Review: the rise of the Western world: a new economic history by Douglass C. North and Robert Paul Thomas. Bus Hist Rev 49(2):237–239

Veblen T (1899) The theory of the leisure class. Macmillan, New York

Veblen T (1904) The theory of business enterprise. Charles Scribner's Sons, New York

Wallis JJ (2014) Persistence and change: the evolution of Douglass C. North. In: Galiani S, Sened I (eds) Institutions, property rights, and economic growth: the legacy of Douglass North. Cambridge University Press, New York

Wells J, Wills D (2000) Revolution, restoration, and debt repudiation: the Jacobite threat to England's institutions and economic growth. J Econ Hist 60(2):418–441

Williamson OE (2000) The new institutional economics: taking stock, looking ahead. J Econ Lit 38(3):595–613

Economic History and Economic Development: New Economic History in Retrospect and Prospect

Peter Temin

Abstract

I argue in this chapter for more interaction between economic history and economic development. Both subfields study economic development; the difference is that economic history focuses on high-wage countries while economic development focuses on low-wage economies. My argument is based on recent research by Robert Allen, Joachim Voth and their colleagues. Voth demonstrated that Western Europe became a high-wage economy in the fourteenth century, using the European Marriage Pattern stimulated by the effects of the Black Death. This created economic conditions that led eventually to the Industrial Revolution in the eighteenth century. Allen found that the Industrial Revolution resulted from high wages and low power costs. He showed that the technology of industrialization was adapted to these factor prices and is not profitable in low-wage economies. The cross-over to economic development suggests that demography affects destiny now as in the past, and that lessons from economic history can inform current policy decisions. This argument is framed by a description of the origins of the New Economic History, also known as Cliometrics, and a non-random survey of recent research emphasizing the emerging methodology of the New Economic History.

Prepared for the 2014 annual BETA-Workshop in Historical Economics hosted by the University of Strasbourg from 9 to 10 May, and organized in association with the Bureau d'Economie Théorique et Appliquée (BETA, http://www.beta-umr7522.fr), the University of Strasbourg Institute for Advanced Study (USIAS, http://www.usias.fr/en/), the Association Française de Cliométrie (AFC, http://www.cliometrie.org) and Cliometrica (Springer Verlag, http://www.springer.com/journal/11698).

P. Temin (✉)
Department of Economics, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: ptemin@mit.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_28

Keywords
New economic history · Economic development · Black death · Industrial revolution · European marriage pattern

The New Economic History was born about 50 years ago. As economics changed after the Second World War, economic history changed as well. The New Economic History started in the 1960s as a part of economic history and has grown to become the dominant strain in economic history today. I survey this progress and think about the future of economic history in three stages. The first stage recalls some of the early days of the New Economic History, its origins and early development. The second stage reflects on the achievements of the New Economic History as shown in recent publications by Robert Allen and Joachim Voth. Taken together, these contributions build on a half-century of research and suggest promising areas for the future. The third stage surveys some other contributions to the New Economic history in a partial and idiosyncratic way and distils implications for the future.

Paul Samuelson arrived at MIT in 1940. Receiving his PhD from Harvard in that year, he was snatched up by MIT when Harvard failed to make him a faculty appointment (Keller and Keller 2001, pp. 81–82). From this event came both the birth of the MIT economics department, and a revolution of economics itself. Samuelson's PhD thesis, published as *The Foundations of Economic Analysis* (1947), championed the use of mathematics in economics. He was not the first economist to use math, but he showed how math could be systematically employed to reformulate familiar and unfamiliar economic arguments. He was like Adam Smith, organizing various strands of existing economics into a new coherent synthesis.

The MIT economics department started its graduate program after the war. It was constructed like a three-legged stool, resting on required courses in economic theory, econometrics, and economic history. But while the legs of a stable stool are equal, these required courses were not. Economic theory and measurement were in their ascendancy, and economic history needed to find a way to coexist with the new theories and econometrics to survive. As in older economics departments, economic history had been taught before the Samuelsonian revolution, but it had been more like history than what we now think of as economics.

One effect of the change in the focus of economics was to change the main mode of reasoning from inductive to deductive. This meant that papers in economics changed from being primarily narrative to starting with a model. New economics papers progressed from a model to data and then hypothesis tests. Economic historians responded to this change in economics by embracing the new tools of economic theory and measurement in what became known as the New Economic History.

This movement was led by the two recipients of the 1993 Nobel Prize in Economics, Douglass North and Robert Fogel. North was editor of *The Journal of Economic History* with William Parker in the 1960s with the conscious aim of attracting papers using formal economics in their analysis. He gained most fame

by stimulating the growth of the New Institutional Economics through his many publications. Fogel burst into this scene with publications first on the social savings of American railroads and then, with Stanley Engerman, on American slavery. These contributions were showcased first at annual meetings of what would come to be called cliometricians held in the 1960s at Purdue University in the dead of winter.

The New Economic Historians threw their lot in with the econometricians. They turned to the collection of historical data and their use in testing hypotheses about economic activity. In this way, the New Economic History brought itself into the mainstream of economics as it was developing, but it caused a growing problem for economic history as economics departments turned their face toward the new theories championed by Samuelson and Solow.

The economic history paper was central to one of the legs of the three-legged stool supporting the MIT economics department. The paper requirement began soon after the war when most field courses had term papers. It was only a remnant of this pedagogical approach to graduate studies by the beginning of the twenty-first century. The two surviving papers, the remnants of the omnipresent term papers in most courses in the 1950s and 1960s, shared several characteristics. Students had to select a question to answer or a hypothesis to test, drawing on their course work or their general knowledge. They had to answer their question or test their hypothesis with using evidence from empirical data. And they had to write this up in the form of an article for an economics journal. They were, in short, two variants of an assignment in applied economics. In fact they were hard to distinguish at the margin and sometimes overlapped.

The two papers also differed in important respects. The history paper drew from economic history – defined loosely to follow the economics convention of focusing on events a quarter-century or more past – for its questions and hypotheses. The aim was for the students to analyze events in a different institutional setting or with unfamiliar relative prices. Given the scarcity of historical data for many interesting historical questions, particularly those about foreign countries, many different quantitative techniques were used. The econometrics paper by contrast was focused on the econometric methodology being used and less on the context in which it was used. And the history paper came in the first year of graduate work, while the econometrics paper was a feature of the second year.

I began to teach economic history at MIT in 1965, and I attended the cliometrics conferences at that time. The dominant memory I have of the conferences was the attention to data. An econometrics professor at MIT had remarked to me that when he could not find data for 1800 that he needed for a regression, he used data from 1900 instead. This was not the culture at the cliometrics conferences. Great attention was taken to the collection and interpretation of data, and disagreements were as often concerned with these issues as with the arguments and hypotheses built on the data.

I presented a paper at one of my first cliometrics conferences on the American iron industry, the topic of my thesis. As I recall, I found the ante-bellum iron data hard to reconcile with my hypotheses, and I proposed what I thought was a reasonable revision of the data for future use. The conferees thought this was a

terrible idea, and there was a lot of critical commentary demonstrating to me that the worlds of economic historians and econometricians had drifted apart. Bob Fogel came up to me after the session, asking how I could remain so calm under the fire I had just sustained. I remarked that the criticism was directed at my paper, not at me. Bob shook his head and rejected that distinction. We went on from there to become friends who often disagreed with each other.

There was palpable excitement among New Economic Historians during the next two decades. Two well-known and controversial books from that time can help us remember this excitement. *A Monetary History of the United States, 1867–1960* by Milton Friedman and Anna J. Schwartz appeared in 1963. They offered a new interpretation of fluctuations in the United States for the previous century and promoted the view that changes in the stock of money were the prime determinants of economic activity. Their claims and Friedman's awesome debating skills made this book a *cause célèbre* among economists and economic historians alike. Their data continue to be used, and their point of view is relevant to current debates. Ben Bernanke, Chairman of the Federal Reserve Board, once said to Friedman that he would not repeat the mistakes Friedman claimed the Fed made in the 1930s.

Time on the Cross by Robert W. Fogel and Stanley L. Engerman (1974) appeared a decade later. It too became famous and controversial, albeit more among historians and economic historians than among economists. They offered a new interpretation of American slavery as a more benign institution than previous authors and in which the rate of exploitation of slaves was markedly lower than previously thought. It is interesting that they derived this latter result by assuming that slaves had to pay for their own upbringing. This approach has returned today as college students increasingly have to pay for their own education as public support for state universities has declined. The growth of student debt is analogous to the debts Fogel and Engerman asserted slaves owed to their owners.

A measure of this intellectual enterprise was taken at the annual meeting of the American Economic Association in 1984. The papers presented in this session were published in the annual *Papers and Proceedings of the American Economic Association*, and the whole session was published in *Economic History and the Modern Economist* (1986), edited by William N. Parker. The session consisted of two papers by economic historians and two by Nobel-laureate economists. The economists took it upon themselves to discuss the place of the New Economics in economics as a whole.

Kenneth Arrow concluded his essay by saying, "In an ideal theory, perhaps, the whole influence of the past would be summed up in observations on the present. But such a theory cannot be stated in any complex uncontrolled system, not even for the Earth, as we have seen. It will always be true that practical understanding of the present will require knowledge of the past (Parker 1986, pp. 19–20).

Robert Solow made essentially the same argument in different words:

> The economist is concerned with making and testing models of the economic world as it now is, or as we think it is. The economic historian can ask whether this or that story rings true when applied in earlier times or other places, and, if not, why not. So the economic historian

> can use the tools provided by the economist but will need, in addition, the ability to imagine how things might have been before they became as they now are. . . . It was once suggested—by my kind of economist—that the division of labor is limited by the extent of the market. Perhaps what I have just been doing can be thought of as suggesting that economists extend their market and accept the specialized services that, in a more capacious market, the historians as well as other scholars, can provide. (Parker 1986, pp. 28–29)

These eminent economists gave good advice. The New Economic History has endeavored to follow it by examining questions drawn from a wide range of places and times, ranging from prehistory to recent events and all around the world. Anywhere there are data or information that can be construed to text hypotheses is fair game.

Three techniques have emerged as particularly useful in these wide-ranging explorations. The first is modern econometrics. New Economic Historians of the first generation used simple econometrics, which were a new way to learn from data in the historical literature. In its new approach to economic history as economics, however, simple econometrics looked like undergraduate econometrics. The use of econometrics was enough to get the first generation employed at good universities, but it was not sufficient for the next generation.

Fortunately, these students had been educated in modern econometrics, and they began to use it in their research. Younger scholars interested in economic history consequently have been able to get jobs at good universities and their articles published in top economics journals. For example, compare my experience at MIT with my younger colleague, Dora Costa. I published largely in economic-history journals and used simple regressions in my work. (I cannot resist noting that my use of even a simple regression about trade in ancient Rome sent ancient historians into a tizzy.) Costa by contrast used cutting-edge econometrics in her work, published regularly in major economic journals and taught econometrics at MIT.

The second technique utilizes the ideas behind event studies to examine the effects of turning points and decisions in economic history. Discontinuities provide information on the structure of economic systems that may not be apparent from their smooth operation in normal times. Legal boundaries provide discontinuities over space, and events ranging from crises to discoveries provide discontinuities over time. These important historical events clarify the structure of economic activity and provide evidence to test preconceived ideas about economic history.

The third useful technique is to examine events over several generations, an opportunity given to economic historians and students of economic development that distinguishes them from some other fields of economics. We can study the effects of demography and education that often are simply held constant in current economic analyses. These two approaches run into each other as we go further back in the past, as we sometimes find the effects of dramatic events in the fortunes of people over several generations. As usual among economists, we distinguish ideal types to think about processes that can be seen as a continuum from another point of view.

The big events of economic history are the Black Death of the fourteenth century, the European discovery of America in the sixteenth century and the Industrial

Revolution in the eighteenth century. We keep going back over these dramatic and far-reaching events to learn more about the path from the slow-moving economies before them to the fast-moving ones today. We know more about the most recent of these events, and it has overshadowed studies of the earlier ones. I want to return to the first of them to illustrate how the New Economic History is reshaping our conception of this transition and to illustrate how much we have gained from this collective activity we call the New Economic History.

When I started teaching these events, we saw the Black Death in very simple ways. It was a demographic shock that sharply reduced the supply of labor while leaving the supply of land intact. The result was a dramatic rise in the real wage, chronicled for England by Phelps Brown and Hopkins (1962) and revised and explored further by Clark (2005, 2007). The English data were extended to continental Europe by two less well-known contributions. The first one was the discovery of what Hajnal (1965) called the European Marriage Pattern. This pattern, as I recall teaching it long ago, had three components. The age of female marriage was high, in their twenties; many women did not marry at all, and married women did not automatically join the household of their husbands. According to Hajnal, this contrasted with an Asian marriage pattern where almost all women married at menarche and moved into extended households of their husbands' families. Hajnal observed this pattern in the early modern period, but he offered no clues where it came from.

The second contribution came from Brenner (1976), who argued that the effects of the demographic changes generated by the Black Death were modified by social and political structures. In the West, that is, England, the monarchy was strong and the aristocracy weak. This left room for workers to take advantage of their relative scarcity and bid up their wages. In the East, vaguely identified as continental Europe, the aristocracy was strong, and it prevented workers from moving to better jobs. This reduced the bargaining power of labor, and wages in the East did not rise after the Black Death. Serfdom decreased in Western Europe and increased in Eastern Europe. Brenner's argument was more controversial than Hajnal's views, and it gave rise to extensive debate – although not to explicit hypotheses testing.

The Brenner debate took place largely outside economics, but it can be seen as an application of North's emphasis on the role of institutions (North 1990). This view gave rise to the New Institutional Economics, a group of economists and economic historians who emphasize the role of institutions in shaping economic affairs. Brenner's ideas can be rephrased as a hypothesis about the role of institutions in shaping responses to the Black Death. The difference between strong monarchies in the west and strong aristocracies in the east was the key to the treatment of labor in this view.

The New Institutional Economics has spread beyond the bounds of standard economic history. It motivates a new view of the economic history of the Greco-Roman world (Scheidel et al. 2007). The editors of this volume tried to move away from the traditional opposition of primitivists and modernists in the study of ancient history into what they considered a more fruitful approach. They found inspiration in North's work and employed the New Institutional Economics to explain differences

among provinces of the Roman Empire, providing insights which other ancient and economic historians have expanded (Temin 2013).

This welter of seemingly unrelated contributions has now been clarified and reformulated by the New Economic History. Voitländer and Voth (2013) argue that the Black Death gave rise to the European Marriage Pattern and set in motion a process that led to the Industrial Revolution. This is a large claim, and it leads to a sharp revision of Western economic history. It needs some explanation to be understood.

Voitländer and Voth argue that the scarcity of labor after the Black Death led to a change in agricultural technology. Moving along the wage-rental isoproductivity line, farmers changed from growing crops to tending animals, from arable farming to husbandry. In other words, movement along a smooth production-possibility curve was a sharp change in the underlying technology. Sir Thomas More expressed it most colorfully over a century after the Black Death in his *Utopia* (2012 [1516]): "Your sheep that were wont to be so meek and tame, and so small eaters, now, as I hear say, have become so great devourers and so wild, that they eat up and swallow down the very men themselves. They consume, destroy, and devour whole fields, houses and cities."

The result of this adaptation of agricultural technology changed the role of women in Medieval society. Switching from crops to husbandry reduced the demand for strength to push plows and expanded the scope of work that women could do. The result was a change in the status of women in society that Alesina et al. (2013) observed at other times and places as well. The reduction in plowing reduced the demand for men's labor and increased it for women's labor. Women's wages rose and their opportunity for work expanded. They delayed marriage, entered service and became more independent. This in turn led to the European Marriage Pattern and the family pattern described by Laslett (1965). It was a massive change in the structure of society, but at the household level analyzed by Hajnal rather than the societal level described by Brenner.

The opportunities open to women delayed their marriage and reduced the rate of population growth. The result was the birth of the high-wage economies of England and a few neighboring countries. Voitländer and Voth test this theory in two ways. They use unpublished data from Broadberry et al. (2011) to estimate that the share of pastoral production in English agricultural output rose dramatically from 47% to 70% between 1270 and 1450. And they show by regressions that the age of first marriage after 1600 – when data become available – was dependent on both the share of pastoral production and its increase since the Black Death in English counties. They conclude that the extensive use of pastoral production increased the age of female marriage by more than 4 years.

The rise in wages as a result of the Black Death was sustained by a shift in marriage patterns that increased the age of women's marriage and reduced the rate of population increase. The adaptation to the initial shock led to a durable rise in people's income. This in turn led to a demand for more meat in their diet, which of course was accommodated by more husbandry. The whole pattern fit together with the Black Death as a shock that shifted households and the economy from one equilibrium to another.

This all fits in with Allen's view of the Industrial Revolution being the result of a high-wage economy. In fact, Voitländer and Voth probably were inspired at least in part by Allen's work. Allen (2009a) argued that the initial innovations of the Industrial Revolution emerged from tinkering by producers to reduce the costs of expensive labor and reap the benefits of cheap power. In response to the awareness from other work by Allen et al. (2005) that wages were high generally in Western Europe, Allen went to some lengths to show that the marginal gains from these initial innovations were not large enough to be profitable in either France or the Netherlands (Allen 2009a, b).

Allen (2013) argues in more recent work that wages and energy prices in North America were close enough to the British pattern for policy initiatives like tariffs, education and infrastructure investments to create conditions hospitable to industrialization. This clearly was true of countries in Western Europe that also followed the British pattern once industrial productivity advanced from its initial level. These countries did not have the factor prices to make the initial innovations of the Industrial Revolution profitable, but further development of these innovations rendered them profitable at factor prices close to those in Britain. And, as Allen noted, policy changes helped industrialization along as it spread.

But this was all within the high-wage area described by Voitländer and Voth. They noted that the European Marriage Pattern extended only from the Atlantic to a line from St. Petersburg to Trieste. Other countries in Asia or Africa were low-wage economies subject to Malthusian pressure on wages, and their factor prices were not close to English prices. Small changes in economic policies were not sufficient to make industrialization profitable in India or Egypt. The story that links the Black Death to the Industrial Revolution therefore is also a story why Europe has industrialized most easily in the past two centuries.

This synthesis reveals that these specific papers extend and unify a generation of contributions to the New Economic History. One strand has been to look at real wages in many times and places, finding evidence where none was suspected before. Another strand has extended financial history back to agrarian economies to reveal a very different index of how economies operated. And a third strand has been insights about odd and interesting facets of economic history that seem at first glance to be only isolated curiosities, but which later turn up as parts of arguments about how all of these strands can be woven together.

Three implications emerge from these recent contributions by the New Economic History. First, they rewrite Western history from soon after the end of the Roman Empire to today. Second, they provide a guide to the role of economic history in economics departments. And third, they call out for a change in publication strategy. I consider these implications in turn.

David Landes (1998) began his magisterial economic history of the West from the discovery of America. The expansion of Europe was an important event, but we now know it was hardly the beginning of the high-wage story. High wages in Western Europe could have resulted from the rise in the ratio of land to labor by the opening up of American land. But we now realize that the start of the high-wage economy

came from the rise in the ratio of land to labor that resulted centuries earlier from the Black Death. The growth of commerce to the New World was helped by British and Dutch shipping and services, and the resulting prosperity kept wages in London and Amsterdam particularly high. The expansion of Europe is an important part of the story, but not the beginning.

Another part of the development of Western Europe was the invention of the printing press in the interval between the Black Death and the expansion of Europe. Printing clearly was a labor-saving innovation, and it is tempting to see it as the result of high wages. Dittmar (2011) however, argued that the spread of printing was related more to the distance from Mainz, where it was introduced, than to factor prices. In terms of this discussion, Dittmar argued that printing was not a marginal innovation like the spinning jenny, but rather a discontinuous change in costs that spread with knowledge. This can only be true in part, as printing spread for the first century or so only within the areas of the European Marriage Pattern.

This single example reveals a more complex story beneath the outline given here. We have to fill in the blanks to provide a new history that reveals the combination of shocks that produced Western history. And while this story is based on simple economics, it requires some modification of the simple Malthusian story. For the high-wage economies of Western Europe were not simply fluctuations around a pre-existing norm; they were a new equilibrium around which population fluctuated. The Malthusian model needs to be expanded to encompass important changes in production and distribution like those that followed the Black Death. The Industrial revolution was not the first escape from the dismal conclusion that real wages could not long stay above subsistence.

This is an important story; how does it fit in modern economics departments? I propose that economic history and economic development should both be considered relevant to modern economic growth. The difference is that economic history traditionally directs its attention to the high-wage economies just discussed, while economic development focuses on the low-wage economies outside Europe. These two inquiries are closely related. They both analyze the growth of economies with new technologies, and they both are concerned with the incentives people have to adopt new innovations.

There is now a large gap between the technologies being used in high-wage and low-wage economies which mirror the large gap between real wages in these two types of economies. If we want to bring the low-wage economies to the level of high-wage economies, we have to modify either the technology being used in the high-wage economies or change the factor prices in the low-wage economies. These are two different directions of research and policy, and they are complementary to each other. If the education and employment of women lead to population control, this will lead to higher wages in poor countries that will make modern technology more appropriate. And if technological innovations like cell phones broaden the factor prices at which they are useful, this too will promote economic development.

Once economic history and economic development are seen as two sides of the same coin, there should be interesting cross fertilization between economic historians and development economists. One interesting factor is the time involved in

economic change. The world appears to be moving rapidly today, but the story of Europe now stretches from the fourteenth to the eighteenth. It is an interesting question how an interaction between these two fields might suggest ways to make a faster transition.

This brings us to the third implication of the New Economic History of Europe. We have to change our publication strategy. Voitländer and Voth published their contribution to European history in the *American Economic Review*, while Allen published his views on economic development in the *Journal of Economic History*. The papers are written for their respective journals, and there would be little point in simply reversing their position – should that even be possible. Instead, we need to think how to get the message across to the relevant audiences. How can we get historians to understand that they must start the story of modern Europe from the Black Death? And how can we get economists to understand that they must start analyzing policy interventions with a consideration of factor prices?

I hesitate to suggest how to do this to these established and prolific economic historians, but I do so to illustrate the paradoxical position of the New Economic History. And just as these contributions build on the work of many New Economic Historians, the job of communicating these results to the appropriate audiences probably would be most effective as a group effort.

Voitländer and Voth need to change from presenting a hypothesis test – the hallmark of the New Economic History – to presenting a narrative that historians will appreciate. They need to place their test in a narrative of Western European history that distinguishes the areas that adopted the European Marriage Pattern from those areas that did not. I have suggested some of the writings that should be included in the intellectual background, but the narrative should focus on telling a persasive story of a critical time in European history.

Allen needs to move in the opposite direction, to extract hypothesis tests from his impressive manuscript that can appear in a good economics journal. He might anchor his tests in a theory like that in Acemoglu and Zilibotti (2001) to provide a bridge between economic growth and economic history. He might incorporate his test of the suitability of the spinning jenny (2009b) or the graphs in his recent survey, but the paper must stand as a test of the overall proposition he made in his presidential address (2013). And it of course needs to have the bells and whistles that current economic articles now sport.

These suggestions of course can be safely ignored. They do however illustrate the paradox of the New Economic History. New economic historians have turned their back on traditional historians and sought their place among economists. This has provided good jobs for many scholars, but the acceptance by economists is still incomplete. We therefore have two challenges ahead of ourselves. The first is to argue that economic development can only be fully understood if we understand the divergent histories of high-wage and low-wage economies. And the other big challenge is to translate our economic findings into historical lessons that historians will want to read. These challenges come from our place between economics and history, and both are important for the future of the New Economic History.

These papers signal the achievements of the New Economic History, but not its breadth. I therefore conclude this paper with a very partial and highly idiosyncratic review of varied contributions to the New Economic History. It should become clear that the list deals mostly with people in and around Cambridge, MA, or that I know personally in other locations.

The first papers deal with the expansion of Europe, but from a different point of view. The Black Death changed Europe, but not at the expense of other people. The expansion of Europe a few centuries later was not as big an event in the economic history of Europe – if you believe the story I have just recounted – but it had repercussions outside Europe that have had lasting effects.

Melissa Dell (2010) investigated the effects of the Spanish silver mines in South America that led to the great European inflation of the sixteenth century. The Potosi and Huancavelica mines that yielded silver and the mercury to refine it were operated by indigenous labor under a *mita* system. Between 1573 and 1812, villages located near the mines in the Andes Mountains were required to provide one-seventh of their adult males as rotating laborers. Dell revealed the effects of this labor system by comparing current conditions in villages under the *mita* with adjacent villages.

Using all three of the techniques listed above, Dell found that the effects of the mita were apparent today, five centuries after the expansion of Europe. She used a "regression discontinuity approach," examining conditions at the edges of the *mita* area. Given the length of time involved and the complex geography, this was not an easy task. Dell exploited both the Spanish preference for workers close to the mines and from the Andean highlands and modern mapping techniques showing altitude for any location. She found that the long-run effect of the *mita* reduced household consumption by one-quarter, resulting at these low income levels in significant stunting of children.

This dramatic finding raised an obvious question: how could it be that the costs of the Spanish exploitation could last over several centuries? Dell organized her explanation around haciendas, rural estates with attached labor force reminiscent of medieval manors. The Spaniards discouraged the growth of haciendas in the *mita* area to preserve their unimpeded access to their labor force. Here we see an inversion of the Brenner thesis that local aristocrats oppressed workers after the Black Death by limiting the extent of the labor market; haciendas would have limited the exploitation of workers by the central, Spanish government by limiting its access to the labor market.

The haciendas were a mixed bag. On the one hand, they expanded after the end of the *mita* by coercive activity ranging from using legal rules to physical violence. On the other hand, they built roads connecting the highlands to lowland urban markets. Access to markets was a critical factor in the history of the high-wage economies of early-modern Europe and North America; it appears to have had similar effects in the low-wage economy of South America. It is worth noting that haciendas cannot be the source of future progress. They were abolished in 1969.

Nathan Nunn (2008) looked at more labor-market effects of the expansion of Europe, this time in Africa. The effects of American slavery in the New World have been the subject of myriad research projects. Nunn inverted this question to ask

about the effects of the slave trade on Africa. In other words, Nunn did not look at slaves and their descendants, but at the people who escaped this fate. Like Dell, he found persistent deleterious effects.

The Atlantic slave trade ended two centuries ago, but Nunn found that African countries that had more slaves per square mile taken from them have lower per capita GDP today. Like the *mita*, the slave trade is gone, but its effects linger on. Nunn made sure that the direction of causation was from trade to economic development, rather than vice versa, or that some other cause was to blame. One demonstration was that slaves were not taken from previously well-organized areas, but rather the reverse. It was the most organized areas that exported the most slaves.

The explanation for this reversal of fortune is that slaves were obtained for export by villages or states raiding each other. The lure of the profits to be gained from slave exports discouraged the expansion of village federations and the growth of ethnic identities. Suspicion and distrust impeded state formation. It is a truism of current development research that the multitudinous ethnic divisions in Africa impede economic growth. Nunn provides at least a partial explanation why there are so many ethnicities in Africa.

This idea can be generalized. Dasgupta (2007) argued that trust is the basis of economic prosperity. He devoted a short summary of economics to this single proposition. Revealing for this discussion, Dasgupta started his discussion by contrasting the conditions of a young girl in the United States with one in Ethiopia, one of Nunn's observations. This rather esoteric exploration into the durable effects of a defunct activity has led directly into the center of economics.

I stated earlier that the New Economic History focused on high-wage economies, yet this survey started with two important papers about low-wage economies. They illustrate how economic history and economic development work together to construct full pictures of poor economies in the world that can lead to productive economic policies. These papers are significant contributions to both economic history and economic development.

Turning now to contributions to American economic history, I start with *Time on the Cross*, mentioned earlier. This innovational study combined the use of massive new data and explicit economic reasoning to reach surprising conclusions. It was not only controversial; it became emblematic of both the advantages and some possible drawbacks of the New Economic History. Its conclusions were contested by both other economic historians and more widely (David et al. 1976).

One aspect of that discussion is unexpectedly relevant today. Fogel and Engerman measured what they called the exploitation of slaves by assuming that slaves were responsible for the costs of their own upbringing. In contrast to the more usual family pattern where parents support children in an intergenerational transfer, they assumed that slaves were isolated individuals who needed to "borrow" from slave owners to eat before they could work. The low earnings of adult slaves then was interpreted more as repayment of these loans than exploitation.

This argument appeared strange to their critics as a description of the nineteenth century, but it seems accurate for the twenty-first century. Slavery is long gone, of course, but its influence remains strong. Margo (1990) described the poor

educational opportunities open to free slaves in the late nineteenth century, and education in urban areas today exhibits a similar pattern of purposeful neglect. Childhood and education have become longer as time has gone on, and a decent education today includes college.

Poor students in the second half of the twentieth century could get low-cost education in state universities where the costs were subsidized by their parents' generation in taxes in an extension of public schools. But as states were strapped for funds at the end of the century and more recent years, it has been the path of least resistance for states to reduce their spending for state universities. State universities are largely private now with state funds accounting for only a minor part of their costs. The universities have raised tuition in an effort to offset this loss of revenue, returning young Americans to the position Fogel and Engerman assumed for slaves.

Wise men and politicians are telling us that the federal debt will burden our children and must be reduced. But the real burden on young people is educational debt caused by state educational policies. Our children have been made responsible for their own college education, which has become an important part of their preparation for work. They are graduating college with overwhelming debts of $100,000 or more, and even those who fail to graduate still leave college with ample college debts. College debt has surpassed credit-card debt, and the President and Congress have wrangled about the interest rate to charge.

This is a historical parallel of some interest and another reason to integrate the New Economic History with current economics. The discussion could even extend to macroeconomics, as the high debt of many young people will depress their consumption in coming years. The analog of slaves presumed repayments to their owners is the low consumption of debt-ridden young people today. The large amount of student debt outstanding suggests that this low consumption may be a drag on the American recovery from the global financial crisis.

Costa and Kahn (2008) examined social debts in a study of Civil War soldiers. They looked at the interactions of soldiers in war and captivity to see the effects of friends and comrades. They found that some soldiers were willing to risk their lives for others (heroes) while others were more like the *homo economicus* of elementary economics (cowards). They reach out to other social sciences for other concerns about the effects of community ties and suggest a variety of hypotheses to be considered. Their research also recalls Adam Smith, using tools derived from *The Wealth of Nations* to raise questions about the topics of *The Theory of Moral Sentiments*.

Hornbeck (2012) extended our understanding of long-run effects of economic changes to natural disasters. The Great Depression is thought of as a macroeconomic event, but the dust bowl of the 1930s was an important part of the national experience. Hornbeck used the same kind of regression discontinuity as Dell to separate the effects of soil depletion and other factors. Land values fell 30% in high-erosion counties.

Hornbeck looked for the kind of substitution in production that Voitländer and Voth found after the Black Death, but found little movement along relevant cost curves. Instead, he found that people migrated out of the dust-bowl area rather

than adjust their agricultural practice to the new conditions. The Okies, as the migrants to California were called, revealed another path of adjustment to change. As Hornbeck noted, this geographical adjustment is typical of recent American labor-force adjustments to other changes in employment opportunities (Blanchard and Katz 1992).

The fall in land prices in the dust bowl is similar to the fall in house prices at the end of the recent housing boom. Many mortgage holders have found themselves "under water" with the value of their loans exceeding the value of the houses. Various forms of relief have been tried, but the banks have resisted writing down their loans. The result is that many people are unable to spend as they would like or move because of their outstanding mortgages. This then has macroeconomic effects as noted already for educational loans. Consumption is down and geographical mobility as an adjustment to labor-market difficulties is not available. The New Economic History of the United States reveals that some of the factors that enabled us to recover from natural and man-made disasters are not available to us now.

Finally, the New Economic History has informed us of recent demographic events other than the Black Death. The "baby boom" in the United States was created by the return of soldiers from the Second World War after a long depression that depressed birth rates. Easterlin (1987) studied how the baby boomers fared in subsequent years. He found crowded schools and increased labor-force competition. The important new observation was the persistence of the effects of the demographic shock. As baby boomers aged, their problems aged with them in age-appropriate ways. For example, as the baby boomers have reached retirement age, politicians are worrying how the Social Security System will be able to handle them. A presidential commission increased the normal retirement from 65 to 67 over many years to prepare for this shock. More changes are under discussion. Urban economists are now even asking if the postwar American growth of suburbs that accommodated all those children is now outmoded. Cause and effect are unclear at this point, but a lower birth rate and shifting technology have begun to have their effects of living patterns. The New Economic History does not have much to say about historical processes just beginning, but the history Easterlin studied is relevant to the work of economists who analyze these movements.

Let us now turn our attention to good fortunes that have been illuminated by the New Economic History. Even if economics is the dismal science, economic history need not be. The largest favorable shock that has been illuminated by the New Economic History has already been mentioned. The Industrial Revolution was a major change whose effects are all around us still. Allen (2009a) used the tools of the New Economic History to show that the Industrial Revolution emerged from the combination of high wages and low energy prices. As already noted, this was such a large historical event that the literature about is immense and ongoing. I can only allude to it here.

Instead, I focus on the good analog of the persistent damage done to people damaged in economic transitions. Clark (2014) has used the extensive data characteristic of the New Economic History to show that half of the variation in overall status of individuals is determined by their lineage. Clark and his colleagues showed

that this is true from the United States to China and Japan, and from Sweden to India. Regression to the mean is apparent in their data, but the process takes hundreds of years.

The methodology was to use surnames to identify descendants. Instead of relying on scarce censuses and family records, Clark and his colleagues identified unusual names characteristic of prosperity at some historical time. They then looked at more recent data on prosperity and social standing to see if these names were over represented. Surprisingly, they were, in many countries and over long periods of time.

This view of durable status has been reinforced recently by Ferrie, long a student of population mobility in the United States (Ferrie 1999). Using the more familiar approach of identifying families in censuses, Long and Ferrie (2014) extended the normal two-generation study of social mobility to three generations in a recent paper. They found more persistence over three generations than over two generations. Clearly, there is a great deal of noise in the mobility of individuals and in any single generation. But extending the length of study provides evidence of greater stability.

Goldin and Katz (2008) used a different approach to analyze the relative fortunes of different groups in America during the twentieth century. Their focus was on education and the difference between educated – and therefore skilled – workers and uneducated and unskilled workers. The progress of technology sets the demand for labor, and the interaction of supply and demand was characterized as a race between education and technology. This colorful metaphor drastically simplifies the many determinants of both education and technology. Their book goes into these complications in great detail.

This contribution is particularly relevant today. Economists examining the distribution of jobs have found that the progress of computers has hollowed out the demand for labor. There are demands for low-wage jobs and quite high-paying jobs, but the demand for factory jobs that were the mainstay of growing employment after the Second World War is down. This has created a need to rethink the simple macroeconomics of labor, since different aspects of technology have effects on different segments of the labor supply. The New Economic History provides a historical background that suggests several important lessons. The nature of technology has been exerting pressure on the wages structure for many generations before this one. Both progress in education (supply) and technology (demand) must be considered when trying to discover effective policies in this area. And, as Goldin (1990) observed in the history of women's work, participants in these sorts of changes cannot predict where they will end up.

Let me abandon this romp through economic history now and try to think more broadly about the future of the New Economic History. I do not like cherry picking in the work of others; I cannot imagine it is informative in much beyond methodology here. The brief sampling of work here does not lead directly to substantive conclusions; it rather suggests the scope of the New Economic History. The subject matter ranges over time from early history to recent events, and over space across continents. If there is one safe prediction, it is that the discovery of new data and of new ways to use existing data will encourage this wide geographical and temporal spread.

As suggested earlier, two aspects of the New Economic History are keys to the growth of scholarship in this area. One is the focus on institutions as carriers of economic structures across generations and sometimes centuries. The other is the focus on causality through imaginative use of identification strategies.

The importance of institutions is undeniable, but its role in research is problematical. The tradition takes its cues from North (1990) and the support of the New Institutional Economics that carries on this tradition. As I have described, the New Economic History often appeals to the role of institutions in the long-term effects of various short-run changes. But while the econometrics are fine in these studies, the accounts of institutional change often are less fully analyzed. Greif (2006) tried to clarify the issues involved, but his concern with theory of institutions may have made the empirical task of finding changes in institutions harder. One issue is that the evidence on institutions frequently is qualitative instead of quantitative. Ways need to be found to quantify what before was not considered quantifiable. In addition, institutions often change only infrequently or very slowly. Finally, it is not always clear how to define the institutions in question. Have morals declined in the United States? Are morals even considered an institutional framework? These are the kind of questions that need more research.

The other characteristic of the New Economic History is the attention given to causality. This typically involves a strong identification strategy to disentangle the motives of different parties to a decision. As shown in the brief selection of work above, the New Economic Historians are aware of this issue and devote a lot of thought to the process of identifying supply or demand influences. Voitländer and Voth went to great lengths to show that the Black Death was in fact the cause of the demographic transition in Western Europe, and Allen has supported his explanation of the Industrial Revolution by comparing factor prices in many other countries. Dell and Hornbeck used geographic boundaries to identify causal elements in their stories.

Let me illustrate these claims with two final examples, one from a young New Economic Historian and one from an old member of our tribe, one from far away and one from long ago. They both involve the consequences of plagues.

The first example is by Dan Li, a Chinese economic historian (Li and Li 2014). She and her coauthor are part of a geographical expansion of the New Economic History to Asia. A recent paper summarized the literature on the history of Chinese economic institutions and macroeconomics for a millennium (Brandt et al. 2014). The paper argues that economic history illuminates choices today – as I have stressed for issues in more familiar venues. Li examined migration from China to Manchuria in the early twentieth century, shortly after a plague that hit the destination of the migrants. The plague reduced population more strongly in some areas than in others. Migrants to areas where the plague hit hard fared better in future years than those to other areas. The question is why did migrants settle there. In other words, was this good fortune determined by design or by luck?

There are no records of individual choices being made, no questionnaires about why a specific destination was chosen. Instead Li and Li (2014) use their data to

distinguish migrants to different areas in the data we have. They found that migrants with higher socioeconomic status avoided plague-hit villages. Migrants to these areas were the least likely to do well in Manchuria.

The second example originated in a conference on quantification in the ancient world. My first reaction was that ancient data was an oxymoron. But my second reaction was that qualitative data – even if only the opinions of modern ancient historians – could be quantified. The process was made manageable by choosing only to quantify the data only in the binary way so typical of our modern electronic devices. By this metric, inflation was either present or not, and political instability was either present or not. American economic historians will recognize this approach as the technique used by Romer (1986) to compare the severity of business cycles throughout the twentieth century. She had to degrade the recent data to make it comparable to the older data. I had to simplify the desired information to quantify at all.

The quantification allowed a decision on timing. The empirical result was that both switches turned on at the same time. This suggested joint causation, and a third possible cause was likely. I looked for a plausible exogenous variable that could have set an interactive process of inflation and instability off together and argued that the preceding Antonine Plague was the cause of both inflation and instability. I commend you to my book for details of the change from the Early Roman Empire to the Late Roman Empire, an important institutional change in world history (Temin 2013).

These two final examples are presented only to highlight the extension of familiar techniques to new fields of inquiry and the opportunities open to the New Economic History. If there is a theme that runs through this survey of where we were, where we are, and where we might go, it is that the fields of economic history and economic growth have much to learn from their interaction.

References

Acemoglu D, Zilibotti F (2001) Productivity differences. Q J Econ 116:536–606

Alesina AF, Giuliano P, Nunn N (2013) On the origin of gender roles: women and plough. Q J Econ 128:469–530

Allen RC (2009a) The British industrial revolution in global perspective. Cambridge University Press, Cambridge

Allen RC (2009b) The industrial revolution in miniature: the spinning jenny in Britain, France, and India. J Econ Hist 69:901–927

Allen RC (2013) American exceptionalism as a problem in global history. J Econ Hist 71:901–927

Allen RC, Tommy B, Martin D (eds) (2005) Living standards in the past: new perspectives on well-being in Asia and Europe. Oxford University Press, Oxford

Blanchard OJ, Katz LF (1992) Regional evolutions. Brook Pap Econ Act 1:1–75

Brandt L, Ma D, Rawski TG (2014) From divergence to convergence: reevaluating the history behind China's economic boom. J Econ Lit 52:45–123

Brenner R (1976) Class structure and economic development in pre-industrial Europe. Past Present 70:30–75

Broadberry S, Campbell BMS, van Leeuwen B (2011) Arable acreage in England, 1270–1871. Unpublished
Clark G (2005) The condition of the working class in England, 1209–2004. J Polit Econ 113:1307–1340
Clark G (2007) A farewell to alms: a brief economic history of the world. Princeton University Press, Princeton
Clark G (2014) The son also rises: surnames and the history of social mobility. Princeton University Press, Princeton
Costa D, Kahn M (2008) Heroes and cowards: the social face of war. Princeton University Press, Princeton
Dasgupta P (2007) Economics: a very short introduction. Oxford University Press, Oxford
David PA et al (1976) Reckoning with slavery: a critical study in the quantitative history of American Negro slavery. Oxford University Press, New York
Dell M (2010) The persistent effects of Peru's mining mita. Econometrica 78:1863–1903
Dittmar JE (2011) Information technology and economic change: the impact of the printing press. Q J Econ 126:1133–1172
Easterlin RA (1987) Birth and fortune: the impact of numbers on personal welfare. University of Chicago Press, Chicago
Ferrie J (1999) Yankeys now: immigrants in the Antebellum United States, 1840–1860. Oxofrd University Press, New York
Fogel RW, Engerman SL (1974) Time on the cross. Little Brown, Boston
Friedman M, Schwartz AJ (1963) A monetary history of the United States, 1867–1960. Princeton University Press, Princeton
Goldin CD (1990) Understanding the gender gap: an economic history of American women. Oxford University Press, New York
Goldin CD, Katz LF (2008) The race between education and technology. Harvard University Press, Cambridge, MA
Greif A (2006) Institutions and the path to the modern economy: lessons from medieval trade. Cambridge University Press, Cambridge
Hajnal J (1965) European marriage patterns in perspective. In: Glass DV, Eversley DEC (eds) Population in history. Edward Arnold, London
Hornbeck R (2012) The enduring impact of the American Dustbowl: short- and long-run adjustments to environmental catastrophe. Am Econ Rev 102:1477–1507
Keller M, Keller P (2001) Making Harvard modern. Oxford University Press, New York
Landes DS (1998) The wealth and poverty of nations: why some are so rich and some so poor. Norton, New York
Laslett P (1965) The world we have lost. Methuen, London
Li D, Li N (2014) Moving to the right place at the right time: the economic consequences of the Manchurian plague of 1910-11 on migrants. Paper presented at the 10th Beta workshop in historical economics, Université de Strasbourg, Strasbourg, May 2014
Long J, Ferrie J (2014) Grandfathers matter(ed): occupational mobility across three generations in the U.S. and Britain, 1850–1910. Paper presented at the modern and comparative seminar, LSE, london, Feb 2014 http://www.lse.ac.uk/economicHistory/pdf/Broadberry/acreage.pdf
Margo RA (1990) Race and schooling in the South, 1880–1950. University of Chicago Press, Chicago
North DC (1990) Institutions, institutional change, and economic performance. Cambridge University Press, Cambridge
Nunn N (2008) The long-term effects of Africa's slave trades. Q J Econ 123:139–176
Parker WN (ed) (1986) Economic history and the modern economist. Basil Blackwell, Oxford
Phelps Brown H, Hopkins SV (1962) Seven centuries of the prices of consumables, compared with builders' wage rates. In: Carus-Wilson EM (ed) Essays in economic history. St. Martin's Press, London, pp 179–196

Romer C (1986) Is the stabilization of the postwar economy a figment of the data? Am Econ Rev 76:314–334
Samuelson PA (1947) Foundations of economic analysis. Harvard University Press, Cambridge, MA
Scheidel W, Morris I, Saller R (2007) The Cambridge economic history of the Greco-Roman world. Cambridge University Press, Cambridge
Temin P (2013) The Roman market economy. Princeton University Press, Princeton
Thomas More S 1478–1535 (2012) Utopia. Penguin, London
Voitländer N, Voth H-J (2013) How the west 'invented' fertility restriction. Am Econ Rev 103:2227–2264

Economic History as Humanomics

The Scientific Branch of Economics

Deirdre Nansen McCloskey

Contents

Abstract

Essays in any field of the intellect about "whither the future of X" have a deep intellectual problem of an economic character. The future is coming, whether we like it or not, and our bets on its outcome will determine how we personally do. But if good predictions were achievable by studying econometrics or by following Warren Buffett, we would all be above average, as in Lake Wobegon. And we are not. So in sober truth, such sessions are actually about "What Do I Want Economic History to Become." Herewith, I am therefore to be allowed to make unrealistic "predictions."

Keywords

Cliometrics · Science · Economic History

Distinguished Professor of Economics and of History Emerita, and Professor of/English and of Communication Emerita, University of Illinois at Chicago. deirdre2@uic.edu. deirdremccloskey.org. The paper originated as a contribution to the session "Whither the Future of Economic History?" at the American Economic Association, Philadelphia, January 2018.

D. N. McCloskey (✉)
University of Illinois at Chicago, Chicago, IL, USA
e-mail: deirdre2@uic.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_47

Introduction

Essays in any field of the intellect about "whither the future of X" have a deep intellectual problem of an economic character. The problem is that if you or I were so smart as to know the answer, then you or I would be rich. If anyone could predict the future of, say, mathematics, she could arbitrage between the present and the future. As the satirical song writer Tom Lehrer put it long ago, she would "publish first." She would achieve riches in a coin relevant to her preferences, namely immortal fame. She would be the Euler of the twenty-first century.

The principle is identical to the more obviously economic one that predictions of the stock market or housing prices or hem lines of skirts are useless. As they say in Hollywood, nobody knows anything. That *Rocky* was a hit doesn't mean that *Rocky 2* or *3* or *N* will be. We have to make predictions, of course, and necessarily we place bets on them. The future is coming, whether we like it or not, and our bets as producers of movies or of mathematics will determine how we personally do. But if good predictions – better than what the average punter makes with his bookie or in the forward markets – were achievable by studying econometrics or by following Warren Buffett, we would all be above average, as in Lake Wobegon. It ain't happenin'.

So in sober truth such sessions are actually about "What Do I *Want* Economic History to Become." I am therefore to be allowed to make unrealistic "predictions."

What Economic History Has Become

I actually think, to be realistic for a moment, that it is probable that economic history will continue for the next decade or so to be dominated by Scientism, quite different from actual science. Scientism is the belief that you are only scientific if you follow a method of science laid down by an amateur philosopher fifty or a hundred or two hundred years ago. In cliometrics, everything is supposed to be quantitative, because then we are Scientists. (I once believed this, so I know.) In science, generally the method is supposed to be Baconian, expressed by Sherlock Holmes in "A Study in Scarlet" as "it is a capital mistake to theorize before you have all the evidence. It biases the judgment." In history, the method comes from Leopold von Ranke's first book, in 1824, in the form of *wie es eigentlich gewesen*, "as [the past] actually was," and in American history from the 1880s to the present in the form of "that noble dream" of an objective historical science.[1] In economics, the method comes from Lionel Robbins (1935) in the 1930s, influenced by Austrian logical positivism already by then under devastating attack by actual philosophers. Nonetheless, the illogical method of logical positivism was enthusiastically seconded by Samuelson (1947) in the 1940s and Friedman (1953) in the 1950s.

[1]Novick 1988 on American history. Novick argues that *eigentlich* should actually be translated "essentially," which gives the phrase a less naively Baconian essence.

The method eventuated in the official constitution of Samuelsonian economics, drafted by Tjalling Koopmans in 1957, *Three Essays on the State of Economic Science.* Koopmans (whose name, by the way, means "salesman") recommended a theoretical/empirical specialization. He recommended that theorists spend their time gathering a "card file" of *qualitative* theorems attaching a sequence of axioms *A', A", A"'*, etc. to a sequence of conclusions *C′, C″, C‴*, etc., *separated from* the empirical work, "for the protection [note the word, you students of free trade] of both," both the theorist and the empiricist. Then the empirical econometricians down in the basement will get to work to see if in the actual world A' leads to C′ or to C″'.

The official method would be fine if the theorems were not merely *qualitative*, the way Samuelson in *Foundations* had laid down they could be. If they took instead the *quantitative* form of the math used by physicists or geologists, in contrast to the on/off existence theorems that mathematicians and economists so love, good. Then the duller wits like McCloskey, the economic historian, could be assigned to mere observation, filling in the blanks in the theory. But *there are no blanks to fill in*, no How-Much questions asked, in the sort of theory that economists admire and that absorbs much of their waking hours (in recent years a little less, I am glad to acknowledge, in favor of quantitative simulation, praise the Lord).

I am here to tell you that the Samuelson-Friedman-Koopmans method will go on being used in economic history for a while, until economic historians realize that whatever its prestige in economics, and its power to overawe in history, it is bankrupt.

In its theoretical branch, the excess of liabilities over assets in the method is well illustrated by non-cooperative game theory. For one thing, experimental economics has shown over and over again that the premise of noncooperation is factually mistaken in humans. For another, finite games unravel, and infinite games have infinite numbers of solutions. In Yiddish syntax, some theory. No one of sense is against theory, if it means economic ideas. Informational asymmetry. Computational general equilibrium. Good. But if all we have is Koopmans' card file of qualitative theorems out to A^{100}', none tested, even in the empty set of ones that can be, what do we have, scientifically speaking?

Ah, but you will reply, we *do* test, with that econometrician down in the basement. No we don't. Name the important factual economic proposition since the Second World War that has been rejected or accepted by an econometric test. Robert Fogel subtitled his book of 1964 on railroads *Essays in Econometric History.* But Bob did not use econometrics, even by the definition of 1964. He used simulation. Rich Weisskoff and I were the RAs for John Meyer about that same time, incompetently helping to edit his essays with Alfred Conrad for a book entitled *The Economics of Slavery: And Other Studies in Econometric History* (also 1964). John and Alf in fact used simulation and accounting and economic ideas. One of Meyer's simulations, an input-output study of British growth in the late nineteenth century, led me out of using the idea of Keynesian aggregate demand for the long run, when I realized that John had done so and that it did not make a lot of sense, in view of the opportunity cost of resources (Kain and Meyer 1968).

Three terms of econometrics, such as I took (with Meyer in one course and Guy Orcutt, that pioneering simulator, for the rest), with no graduate training in other empirical methods – such as simulation, archival research, experiment, surveys, graphing, national income accounting – makes modern PhDs into savants of tests of statistical significance. Test, test, test says David Hendry. The trouble is that such tests are themselves bankrupt, as, for example, Kenneth Arrow (1960) noted a few years after Koopmans' constitution.

You will not believe, I realize, that null hypothesis testing without substantive judgment of magnitudes is bankrupt. Perhaps, though, you can come to believe by reading the report in 2016 of an official committee of the American Statistical Association (ASA 2016). I'll make a prediction. Someday, if you are young and scholarly and intellectually flexible enough, you will get it, and will give up mechanical tests of statistical significance at the 0.05 level and will start doing real science.

On the side of theory and Koopmans' card file, I worry in economic history about "analytical narratives," which seem to be popular with neo-institutionalists of the Northian tendency. Certainly, the contribution of economic history to economics consists in part in a "narrativizing" of economic behavior, a moving picture as against the more usual snapshot. The trouble again is the lack of meaningful quantitative testing. Quantification doesn't seem to happen in neo-institutionalism to a high standard. If the analysis is "consistent with" some little piece of economic history, all is said to be fine. What one would like to see is quantitative oomph, or else the humanist's substitute for quantity, serious comparative histories. Either or both would do.

I hesitate to cast the first stone, because I am not without sin. True, as the men caught in the #MeToo movement nowadays often say in extenuation, the sinning was a long time ago. Still, by confessing my own sins here and now I can avoid the impoliteness of naming particular works by my beloved colleagues in economic history that routinely misuse analytic narratives – that is, existence theorems, weakly "consistent" with the data, and not comparative, either. It would be easy to name them. It would be even easier to name colleagues who use tests of statistical significance – also weakly "consistent" with the data, at low power, and anyway usually irrelevant to the economic and historical question at issue, which is almost always not fit, but coefficient size, substantive oomph. But I won't.

So, bless me, father, for I have sinned. It has been half a century since my last confession.

In 1971, I gave a paper to the meetings of the Economic History Association, published the next March in the *Journal of Economic History* as "The Enclosure of Open Fields: Preface to a Study of Its Impact on the Efficiency of English Agriculture in the Eighteenth Century (McCloskey 1972)." Sounds swell, eh? But I remember uneasily that my commentator in the session at the meetings, the law professor, and student of environmental law and economics, Earl Finbar Murphy, complained that I had not shown in the paper that my very clever analytical narrative had actual quantitative oomph. I was distressed at the complaint, and, in the way of young scholars, even angry at Murphy. But I must have taken his complaint to heart,

because the next time I ventured into the open fields and their enclosures, I made sure to provide the quantitative goods, in a paper in 1976 called "English Open Fields as Behavior Towards Risk (McCloskey 1976)."

My friend Stefano Fenoaltea (1970) challenged the argument that scattering of plots was behavior towards risk with another analytic narrative, also as in my original 1971 paper not tested for oomph. Uncharacteristically for Stefano, his paper did *not* provide the quantitative goods. So in reply I wrote with a student at Chicago, John Nash, a paper actually measuring the oomph of Stefano's suggestion of storage of grain as insurance alternative to scattering of plots (McCloskey and Nash 1984). Oomph. Good on me, even at age 42.

And in that paper with John, to speak of econometrics, in order to isolate the cost of storage per month, including interest, we regressed changes in the prices of grain in any particular location against the number of months the change was measured over. The slope is what mattered. We had the good sense not to use tests of significance to "test" what any economist already knows, that prices rise after a harvest by the full cost per month of storage. They have to, for elementary reasons of arbitrage. Our R^2s were derisory. But so what? We were filling in the factual blanks in a quantitatively specified theory.

And again in the same era (the light was dawning in me slowly, slowly) I wrote a paper with J. Richard Zecher on "How the Gold Standard Worked" (1976), which used regression analysis to articulate a quantitative standard of what "one market" means. We did not do what studies of market integration routinely do down to the present, which is to "test" against a 0.05 standard of significance "whether or not" separate markets were integrated. Such a test is meaningless. There is no sense, Dick persuaded me, in which on/off, yes/no is a scientific standard. One has to have a comparative standard, such as within-USA integration of the market in bricks compared with international integration, USA vs. UK, for example. It's true in physics and it's true in economic history.

What Economic History Could Be

Enough of reality. What do I *want* economic history to become? What are my unrealistic predictions? In brief: I want it to continue to be the scientific part of economics and of history, but to get even more scientific than it is now.

Many economic historians trained as economists lack self-confidence in the face of their proud if ignorant colleagues in theory and econometrics, and therefore do not realize that economic history is the Darwinian-scientific part of economics. Peter Temin, my first teacher of cliometrics in the first year he gave the course, has recently lamented the decline of economic history in economics departments, such as his own at MIT, in favor of more appointments in theory and econometrics (Temin 2016). One wonders why in his long career at MIT he did not arrange for succession, though I have to admit that at Chicago I did not either. Two months after I left the University of Chicago in 1980, the department abolished the requirement in the graduate

program of a course in economic history. Now PhDs from Chicago, as from MIT, know nothing of the past of the economy.

Economic historians trained in other fields, such as sociology or history itself, or in departments of economic history in the Old World, are less inclined than their cliometric colleagues in the United States to whore after the latest "insight" alleged to matter factually from the theorists, or the latest "technique" of the econometricians identifying the number of angels on the head of a pin. But anyway, I say again, with pride at my colleagues' accomplishments worldwide, economic history is the almost completely *scientific* portion of economics and of history. It just needs more.

Realize, though, that the word "science" is a big problem in English, and is misleading economic historians to try to imitate what they imagine happens in physics. In all other languages, from French to Tamil and back, the local science word means merely "systematic inquiry," as distinct from, say, casual journalism or unsupported opinion. In German for example *Geisteswissenschaften,* which means literally in English a spooky sounding "spirit sciences," is the normal German word for what American academics call the "humanities," the British "arts." The Dutch to this day speak of *kunstwetenschap,* "art science," which English speakers now would call "art history" or "theory of art" and place firmly in the humanities arrayed against Science. In Italy, a proud mother of a 12-year old girl who is doing well at school speaks of *mia scienziata,* which sounds strange indeed in recent English, "my scientist."

In earlier English, *Wissenschaft* or *wetenschap* or *scienza* is what "science" also meant in English. Thus Alexander Pope in 1711: "While from the bounded level of our mind/Short views we take, nor see the lengths behind: /But more advanced, behold with strange surprise/New distant scenes of endless science rise!" (Pope 1711, *Essay on Criticism*, lines 221–224). Then in the mid-nineteenth century, as a result of disputes over chairs of chemistry in Oxford and Cambridge, the word was specialized to the systematic study of the physical world. In the *Oxford English Dictionary,* the new meaning, slowly adopted from the 1860s on (Alfred Marshall never did adopt it, but by the time of Keynes everyone had), became sense 5b, the dominant sense now, the lexicographers of Oxford inform us, in ordinary usage.

The usage of the last century and a half makes for endless yet silly disputes about whether economics is a Science, and gives natural scientists permission to issue lofty sneers about social *science*. Yet what would it matter to the practice if after learned dispute we decided that economics or economic history were *not* Sciences? I suppose we would be banished from the National Science Foundation or the National Academy of Science, which would be sad and unprofitable. But would the banishment change the actual practice of economic or historical science?

In actual practice, indeed, the sort of categorical issues that occupy *humanistic* sciences are an essential step in any systematic inquiry, whether of physical or social or conceptual matters. The humanities – such as literary criticism in the Department of Literature, and number theory in the Department of Mathematics, and transcendent meaning in the Department of Theology – study categories, such as good/bad, lyric/epic, 12-tone/melodic, red giant/white dwarf, hominid/*Homo sapiens*, prime/not, consciousness/not, God/gods, exist/not. The crucial and neglected point in the

battle of the Two Cultures is that such humanistic and human categorization is a *necessary initial step in any scientific argument*. You have to know what your categories *are* by well-considered definitions, such as *Homo sapiens sapiens*/*Homo sapiens neanderthalensis*, before you can *count* their members. This is obvious – though not to the anti-humanistic George Stiglers or Michael C. Jensens or Murray Rothbards of economics.

For example, economic theory is entirely humanistic, dealing in definitions and their relations, sometimes called "theorems" or, more usefully for an empirical science, "derivations." Theory makes remarks about categories – as Coase did: Transactions costs may be important here, and this is how they should be defined. Or, as Irving Fisher and Milton Friedman said, MV = PT. Or, as Edgeworth and Samuelson said, $(dU/dx)/(dU/dy) = P_x/P_y$. Or, as the Austrian economists say, markets may be more about events out of equilibrium than about equilibrium. Or, as Israel Kirzner and now D. N. McCloskey might put it, discovery may be more important for human progress than is routine accumulation or routine maximization of known functions.

At the level of economic theorizing, such folk are humanists, dealing in categories and derivations, in advance of and sometimes in lieu of examining the history of actual markets. I recently spent some time browsing through 2014 Nobel laureate Jean Tirole's textbook on the theory of finance (2006). The book gathers some hundreds of theories, with no shred of evidence supplied about which of the theories might apply to actual financial markets. For good or bad, it is as much an exercise in humanism as is Kant's *Critique of Pure Reason* (2008) or Ramunajan's notebooks on number theory (Ramanujan et al. 2000).

Some definitions and their corresponding theorems are wise and helpful, some stupid and misleading. The humanities, and the humanistic steps in any science, study such questions, offering more or less sensible arguments for a proposed category being wise or stupid, before counting or comparison or other factual inquiry into the world. The humanities study the human mind and its curious products, as for example, John Milton's *Paradise Lost* or Mozart's Flute and Harp Concerto in C (K. 299) or the set of all prime pairs or the definition of GDP. The studies depend on categories, such as enjambed/run-on lines or single/double concerti or prime/not-prime or marketed/unmarketed, such as we humans use.

In the early twentieth century, for example, many economists and other scientists such as the great statistician Karl Pearson believed that the category "Aryan race" was wise and helpful for thinking about the economy and the society.[2] Around then the American Progressives, and especially the leading economists among them, believed passionately in racism. They advocated policies such as immigration restrictions (later passed into law with the assistance of the KKK) and the minimum

[2]A late example of his views is Pearson and Moul 1925, "Taken on the average, and regarding both sexes, this alien Jewish population is somewhat inferior physically and mentally to the native population." And an early one is Pearson 1882 (1900), pp. 26–28, "From a bad stock can come only bad offspring. . . ."

wage (now adopted by modern Progressives) and compelled sterilizations ("Three generations of imbeciles are enough") to achieve eugenic results in favor of the perfection of the Aryan race (Leonard 2016). Later we decided, after some truly disturbing experiences and more reflection, that "race" aside from *Homo sapiens* was actually a stupid and misleading and even evil category. The decision itself depended on reflections on the humanistic categories of helpful/misleading, wise/stupid, good/evil.

The necessity of the humanistic first step, note well, applies to physical and biological sciences as much as to *les sciences humaines* or *die Geisteswissenschaften.* Meaning is scientific, because scientists are humans asking questions interesting to them about the meaning of β decay. Such is the main conclusion of science studies since Thomas Kuhn. The Danish physicist Niels Bohr wrote in 1927, that "It is wrong to think that the task of physics is to find out what the world is. Physics concerns what we can *say* about it."[3] We. Humans. Say. With words. About such *geisteswissenschaftliche* categories the German-American poet Rose Äuslander (2014) wrote, "In the beginning/was the word/and the word was with God/And God gave us the word/and we lived in the word. /And the word is our dream/and the dream is our life."[4]

We dream of categories, in our metaphors and stories, and with them make our models and our economic histories and our lives, especially our scientific lives, saying the world. The poet Wallace Stevens (2011) exclaims to his companion, walking on a beach in Key West, "Oh! Blessed rage for order, pale Ramon, /The maker's rage to order words of the sea," the human arrangement of words imposing order on the world. Of the woman they had heard singing, Stevens sang, "when she sang, the sea, /Whatever self it had, became the self/That was her song, for she was the maker."

There is nothing scary or crazy or French or postmodern or nihilistic about such an idea. The "hardest" sciences rely on human categories, and therefore on human rhetoric and hermeneutics, the speaking and the listening sides of human conversation in the sciences. The category of "capital accumulation," for example, can be defined in an aggregate, Smithian-Keynesian way. Or it can be defined in a disaggregated, action-specific Austrian way. It matters to the science, changing what we then proceed to measure and to recommend by way of policy. The humanistic job of economic theory is to ponder the categories, to see their internal logic, to criticize and refine them, just as in the departments of English and of physics.

But the humanistic step – though I am saying it is quite necessary for scientific thought – is of course not in a factual or policy science like economics the whole

[3]Quoted in *Niels Bohr: Reflections on Subject and Object* (2001) by Paul McEvoy, p. 291. The provenance of the remark is a little hazy, but it is well known. In Danish, the philosopher Hans Siggaard Jensen informs me, it was something like "*Fysik er ikke om hvordan verden er, men om hvad vi kan sige om den*."

[4]*Am Anfang/war das Wort/und das Wort/war bei Gott/Und Gott gab uns das Wort/und wir wohnten/im Wort/Und das Wort ist unser Traum/und der Traum ist unser Leben*.

scientific job. The point is one that economists regularly miss in their fascination with the blessed rage for order. Theory is not science *tout court.* One could have a theory of epics or concerti that never applied to any actual epic or concerto, and indeed foolishly misrepresented them as they are in the actual human world. Specializing in humanistic theorizing of the sort that Kenneth Arrow or Frank Hahn did is dandy, but it does not do the entire scientific job unless it is at some point firmly attached to experiment or observation or other serious tests against the world, as much of the work of these two brilliant men never was. The philosopher and economist Arthur Diamond looked into the empirical uses down in the empirical basement of abstract general equilibrium theory such as Arrow and Hahn practiced, and found that there was none.[5] If you are making a quantitative point, as must happen in a policy science like economics or in a world-speaking science like physics or in the glorious systematic inquiry into the past of the business of ordinary life called economic history, then after the humanistic step you must proceed to the actual count or the testing comparison. Count the deaths from plague in the 1340s or compare its impact in China, from whence it came.

Too often in economics the count or comparison does not happen, because economists think, as I have said, that theorems offer factual "insight," and believe that statistical significance "tests" the theory against facts. The two sides, theory, and econometrics, they say, can therefore specialize and specialize and specialize. Never trade. Such a procedure believes it imitates physics but does not inquire into how physics actually works. Physicists, even theorists, as one can see in the lives and writings of Enrico Fermi and Richard Feynman, spend much of their time studying the physical equivalent of the *Journal of Economic History.*

Conclusion

So what? Here's what. Economic history should become as humanistic as it is now foolishly anti-humanistic. It will become so if we overcome our anxiety that we might not be worthy of the white coats of the Scientists, sense 5b.

The word is "humanomics," coined by the indubitably scientific experimental economist Bart Wilson and embodied now in a book in progress by Bart and the Nobel laureate and founder and first president in 1986 of the Economic Science Association, Vernon Smith (Smith and Wilson 2017). I quarreled with Vernon at the time about the appropriation of the word "science" for what is mainly an association of laboratory experimenters. By now, he and I thoroughly agree that "science" covers more than misinformed imitations of physical sciences. Humanomics does not abandon what we can learn from such imitations, and is certainly not against mathematics or statistics. That is what is disastrously wrong with the complaints of, say, Francesco Boldizzone against cliometrics (Boldizzone 2011; McCloskey 2013).

[5]Diamond 1988, though Leland Yeager noted correctly that it does provide a useful "integrating factor of the whole body of economic theory" (Yeager 1999, p. 28).

There is much of value in the old German Historical School and the old American Institutionalists, but their ignorant hostility to what they called "English economics" is not it.

A properly scientific economic history invites the methods of the humanities into economic science. Bart and Vernon and colleagues, for example, are increasingly studying the meaning that their experimental subjects attach to their actions, as revealed by the humanistic techniques of textual analysis of what the subjects say to each other during the experiment. It has been known for decades in experimental economics that letting the subjects speak to each other can radically change the results. The study of human meaning casts light on the ordinary business of life.

We in economic history are well placed to take advantage of humanomics, as for example, merging business and economic history. But to do so, clearly, we need to set aside our anxieties about the National Academy of Science, and listen to *all* the evidence about the economy, whether it comes in the form of statistics of exports or the letters of one businessperson to another or the themes of contemporary drama. That last, for example, is one piece of evidence that attitudes towards business were radically changing in England in the first decades of the eighteenth century (McCloskey 2016).

Our colleagues in economics are trudging in the other direction, with a behavioral economics that ignores human meaning in favor of insisting, in the manner of 1930s psychology, that all that matters is external behavior; or, more extremely anti-humanistic, neuro-economics, which studies the brain but ignores the mind, as though we could understand Jascha Heifetz' fiddle playing by a closer and closer study of his muscles.

The solution is not either to run after the latest declared crisis in the environment or the latest "current policy issue" in the labor market, though admittedly the temptation among young scholars cowed by their present-minded colleagues to do so is great. The TV and newspaper and present politics are not good guides to what is permanently important in the study of the nature and causes of the wealth of nations. Bob Fogel once said that his principle in choosing topics for research was to do nothing that would not matter in 50 years. It is why early in the 1970s he abandoned some tentative research on the history of federal land policy in the United States. What will still matter in 50 years in economic history is poverty and its ending, and in political history tyranny and its ending. If poverty and tyranny are ended, the rest will follow. Better stick to the important issues.

I do not wish merely to preach (although come to think of it there's nothing wrong with preaching the gospel of scientific common sense). So let me give a concrete example of the scientific payoff of humanomics. Somewhat embarrassingly, the example is my own recent work in economic history, my trilogy on The Bourgeois Era (2006, 2010, 2016). Forgive me, mother, for I have sinned.

The Great Enrichment per capita in real terms by a factor of 20 or 30 or much, much more since 1800 is the most astounding economic change since the domestication of plants and animals. Historians, economists, and economic historians have been trying to explain it since Smith. Recently some have come to concentrate on the role of ideas, as in the work of the economic historians Joel Mokyr (2016) and Eric

Jones (2003), the historian Margaret Jacob, the historical sociologist Jack Goldstone, the anthropologist Alan MacFarlane or the economist Richard Langlois, or myself (2006, 2010, 2016), among a courageous handful of others.

The Great Enrichment has usually been explained by material causes, such as expanding trade or rising saving rates or the exploitation of the poor or changes in the rules of the legal game. The trouble is that such events happened earlier and in other places. For example, any organized society from hunter-gatherers to the United States of America has private property and the rule of law, or else it is not a society but a war of all against all, contra North and Weingast (1989). For another example, foreign trade is ubiquitous in human history, and is not about power, contra Findlay and O'Rourke (2007). Such points in criticism of the instinctive materialism of economic scholars after 1890 or so is itself a use of the humanist's substitute for quantification, comparison. Such material events as legal change and foreign trade cannot therefore explain the Industrial Revolution (which in fact has earlier parallels, usually a mere doubling of incomes). Especially they cannot explain its astounding continuation in the Great Enrichment of 3000 percent per capita. One can show in considerable detail (McCloskey 2010) that the material causes we study in economics do not work. One can also show (McCloskey 2016) how attitudes towards the bourgeoisie began to change in the seventeenth century, first in Holland and then in an England with a new Dutch king and new Dutch institutions. And one can show that the Bourgeois Revaluation was not ethically corrupting (McCloskey 2006).

One hypothesis is that if the social position of the bourgeoisie had *not* been raised in the way people spoke of it, then the aristocrats and their governments, or the bourgeoisie itself in guilds and mercantilism, would have crushed innovation, by regulation or by tax, as they had always done. And the *bourgeois gentilhomme* himself would not have turned inventor, but would have continued attempting to rise into the gentle classes. Yet if the material methods of production had not thereby been transformed, especially after 1800, the social position of the bourgeoisie would not have continued to rise. One could put it shortly: without spoken honor to the bourgeoisie, no modern economic growth. (This last was in essence of Milton Friedman's Thesis.) And without modern economic growth, no spoken honor to the bourgeoisie. (This last is in essence of Benjamin Friedman's Thesis.) The two (unrelated) Friedmans capture the essence of poor men, and women and slaves and colonial people and all the others freed by the development of admiration for the bourgeois virtues.

The causes were liberalism (McCloskey 2016), the scientific revolution (Mokyr 2002; not, however, in science's direct technological effects, which were postponed largely until the twentieth century), and above all a change in the rhetoric of social conversations in Holland and then in England and Scotland and British North America about having a go. The change in rhetoric was in turn a result of accidents of politics and society in northwestern Europe from the Reformation to the French Revolution that made people bold. By happenstance in the other advanced societies at the time such as China or the Mughal and Ottoman empires, boldness in the economy was not encouraged, as it had been in those places earlier, and in many

other ways, such as always in military technology. Europe "won" the war of civilian betterment by accident.

You can ask how an explicitly and persuasive bourgeois *ideology* emerged after 1700 from a highly aristocratic and Christian Europe, a Europe entirely hostile – as some of our clerisy still are – to the very idea of bourgeois virtues beneficial to the poor. In 1946 Schumpeter declared that "a society is called capitalist if it entrusts its economic process to the guidance of the private businessman" (Schumpeter 1946). It is the best short definition of that essentially contested concept, and highly misleading word, "capitalism." (Misleading because it invites us to focus on aggregate capital accumulation rather than on the particular and Austrian discovery of ideas for betterment that actually made the modern world.) "Entrusting" the economy to businesspeople, Schumpeter explained, entails private property, private profit, and private credit. (In such terms, you can see the rockiness of the transition to capitalism in present-day Russia, say, where agricultural land is still not private, and where private profit from actual enterprise, as against extortion and theft, is still subject to prosecution by the state, the jailing of billionaires, the cutting down of tall poppies.)

Yet what Schumpeter leaves aside in the definition, though his life's work embodied it, is that the society – or at any rate the people who run it – must *admire* businesspeople. Schumpeter, as Richard Langlois (2014) has noted, had no sociological theory. People must come sociologically to think the bourgeoisie capable of virtue. It's this admiring of the bourgeois virtues that Russia lacks, and has always lacked, whether ruled by boyars or tsars or commissars or Putin and his friends, ever since Muscovy long ago fended off the Mongols, at the sacrifice of the commercial model of Novgorod.

Attributing great historical events to ideas was not popular in professional history for a long time, 1890–1980. A hardnosed calculation of interest was supposed to explain all. Men and women of the left believed in historical materialism, and so strong was the belief that many liberals or conservatives were embarrassed to claim otherwise. But such a result of the "dream of objectivity" hasn't worked out all that well. Actual interest – as against imagined and often enough fantasized material interest – did not cause World War I. The Pals Brigades did not go over the top at the Somme because it was in their prudent interest to do so. Non-slave-holding whites did not constitute the bulk of the Confederate armies for economic reasons. Nor had abolition become a motivating cause because it was good for capitalism. And on and on, back to Achilles and Abraham pursuing their honor and their faith.

We do well therefore as economic historians to watch for cognitive-moral revolutions, and not simply to assume that Matter Rules, every time. A showing that ideas matter is not so unusual nowadays among historians. But it is another project to show that the material base itself is determined by habits of the lip and mind – *that* conclusion evokes angry words among most people on the economistic side of the social sciences, and often enough from historical materialists in the humanities.

In short, the Great Enrichment is a world-historical example of the force of language in the economy – its linguistic embeddedness as the sociologists would put it. In the economy the force of language is not to be ignored. (Or that it *is* to be

ignored: if the research is genuine the possibility must be lively that the hypothesis turns out to be wrong.)

Thus "humanomics." Ignoring the burden of art and literature and philosophy in thinking about the economy is bizarrely unscientific. It throws away, under orders from an unargued and demonstrably silly Method, a good deal of the evidence of our human lives. I do not mean that "findings" are to be handed over from novels and philosophies like canapés at a cocktail party. I mean that the exploration of human meaning from the Greeks and Confucians down to Wittgenstein and *Citizen Kane* casts light on profane affairs, too. A human with a set of virtues and vices, beyond the monster of interest focusing on Prudence Only, characterizes our economies, if not our economics.

And so (the hypothesis goes) an economic history without meaning is incapable of understanding economic growth, business cycles, or many other of our mysteries. A humanomic economic history would extend but also to some degree call into question the techniques of modern economics, and the numerous other social sciences from law to sociology now influenced by an exclusively Max *U* theory.

Economic history, that is, can embrace the humanities, without forsaking measurement, and become more, not less, scientific.

References

American Statistical Association (2016) Statement on statistical significance and P-values. Am Stat 70(2):129–133. At http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108

Arrow KJ (1960) "Decision theory and the choice of a level of significance for the *t*-test." Pp. 70–78 in Olkin, Ingram, et al., eds. 1960. Contributions to probability and statistics: essays in honor of Harold Hotelling. Stanford University Press, Stanford

Ausländer R (2014) While I am drawing breath. Arc Publications, Todmorden

Boldizzoni F (2011) The poverty of Clio, arc publications: resurrecting economic history. Princeton University Press, Princeton

Conrad AH, Meyer J (1964) The economics of slavery, and other studies in econometric history. Aldine Publishing Co., Chicago

Diamond AMJ (1988) The empirical progressiveness of the general equilibrium research program. Hist Polit Econ 20(1):119–135

Fenoaltea S (1970) Risk, transaction cost, and the organization of medieval agriculture. Explor Econ Hist 13:129–151

Findlay R, O'Rourke KH (2007) Power and plenty: trade, war, and the world economy in the second millennium. Princeton University Press, Princeton

Fogel R (1964) Railroads and American economic growth: essays in econometric history. Johns Hopkins University Press, Baltimore

Friedman M (1953 reprint ed 1966) Essays in positive economics. University of Chicago Press, Chicago

Jones E (2003) The European miracle: environments, economies and geopolitics in the history of Europe and Asia. Cambridge University Press, Cambridge

Kain JF, Meyer JR (1968) Computer simulations, physio-economic systems, and intraregional models. Am Econ Rev 58(2):171–181

Kant I (2008) The critique of pure reason. Penguin Classics, New York

Koopmans T (1957) Three essays on the state of economic science. McGraw Hill, Chicago

Langlois R (2014) The dynamics of industrial capitalism: Schumpeter, Chandler, and the new economy. Routledge, London
Leonard TC (2016) Illiberal reformers: race, eugenics and American economics in the Progressive era. Princeton University Press, Princeton
McCloskey DN (2013) The poverty of Boldizzoni: resurrecting the German Historical School. In: Investigaciones de Historia Economica Feb 9(1):2–6
McCloskey DN, Zecher JR (1976) How the gold standard worked, 1880–1913. In: Frenkel JA, Johnson HG (eds) The monetary approach to the balance of payments. Allen and Unwin, London, pp 357–385
McCloskey DN (1972) The enclosure of open fields: preface to a study of its impact on the efficiency of English agriculture in the eighteenth century. J Econ Hist 32(1):15–35
McCloskey DN (1976) English open fields as behavior towards risk. Res Econ Hist 1(Fall):124–170
McCloskey DN (2006) The Bourgeois virtues: ethics for an age of commerce. University of Chicago Press, Chicago
McCloskey DN (2010) Bourgeois dignity: why economics can't explain the modern world. University of Chicago Press, Chicago
McCloskey DN (2016) Bourgeois equality: how ideas, not capital or institutions, enriched the world. University of Chicago Press, Chicago
McCloskey DN, Nash J (1984) Corn at interest: the extent and cost of grain storage in medieval England. Am Econ Rev 74:174–187
McEvoy P (2001) Niels Bohr: reflections on subject and object. Microanalytix, San Francisco
Mokyr J (2002) The gifts of Athena. Princeton University Press, Princeton
Mokyr J (2016) A culture of growth: origins of the modern economy. Princeton University Press, Princeton
North DC, Weingast BR (1989) Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. J Econ Hist 49:803–832
Novick P (1988) That Noble Dream: the 'Objectivity Question' and the American Historical Profession. Cambridge University Press, Cambridge/New York
Pearson K (1882, reprint 2004) The grammar of science. Walter Scott. Dover Publications, London
Pearson K, Moul M (1925) The problem of Alien Immigration into Great Britain, Illustrated by an Examination of Russian and Polish Jewish Children. Ann Eugenics 1(1/2):5–127
Pope A (2008) An essay on criticism. Forgotten Books, London (2008 reprint from the 1711 original)
Ramanujan Aiyangar S, Hardy GH, Seshu Aiyar PV, Wilson BM (2000) Collected papers of Srinivasa Ramanujan. AMS/Chelsea Publication, London
Robbins L (1935) An essay on the nature and significance of economic science, 2nd edn. Macmillan, London
Samuelson PA (1947) Foundations of economic analysis. Harvard University Press, Cambridge
Schumpeter JA (1946) Article "Capitalism". In: Encyclopedia Britannica. Encyclopedia Britannica Inc, Chicago
Smith V, Wilson B (2017) Sentiments, conduct, and trust in the laboratory. Soc Philos Policy 34 (1):1–33
Stevens W (2015) The collected poems of Wallace Stevens. Vintage International, New York
Temin P (2016) Economic history and economic development: new economic history in retrospect and prospect. In: Diebolt C, Haupert M (eds) The handbook of cliometrics. Springer Verlag, Berlin
Tirole J (2006) The theory of corporate finance. Princeton University Press, Princeton
Von Ranke L (1824) Geschichten der romanischen und germanischen Völker von 1494 bis 1514. University of Michigan Library, Ann Arbor (edition 1885)
Yeager L (1999) Should Austrians scorn general-equilibrium theory? Rev Austrian Econ 11(1–2):19–30

Cliometrics and the Study of Canadian Economic History

Ian Keay and Frank D. Lewis

This chapter is dedicated to the memory of Frank Lewis, my co-author, colleague, and friend. I.K.

Contents

Abstract

The long-run development of the Canadian economy has been marked by discontinuities in policy, structure, technology, and performance. By applying cliometric tools and techniques, our understanding of these discontinuities has evolved well beyond the narratives that traditionally documented Canadian economic history. The theoretical foundations and quantitative rigor of cliometric analysis allow us to describe the Canadian experience relative to contemporary and international parallels. In this chapter we illustrate the transformative effect cliometrics has had on the study of Canadian economic history through a survey

I. Keay (✉) · F. D. Lewis
Department of Economics, Queen's University, Kingston, ON, Canada
e-mail: keay@queensu.ca; lewisf@econ.queensu.ca

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_43

of the key contributing factors and episodes that affected Canadian development. These factors include the role of resource-led growth, including the impact of the fur trade and western wheat production; the relationship between policy and performance, including the adoption of trade protection, the subsidization of infrastructure, and the impact of immigration; and the connections between economic policy and entrepreneurial decision-making.

Keywords
Canadian economic history · Resource-intensive development · Fur trade · Wheat boom · Trade policy and growth · Immigration policy · Railway building · Technological choice · Entrepreneurial failure

Introduction

Canada is a big place, but its aggregate economy has always looked rather small relative to its main trading partners. The Canadian land mass is over 23% larger than the lower 48 US states, but both total population and GDP in Canada have consistently been no more than about one tenth that of the aggregate American economy. The extraction and processing of natural resources have been, and continue to be, a considerably larger share of the Canadian economy than in the USA. Urquhart (1993), and domestic manufacturing establishments during the late nineteenth- and early twentieth-century industrialization phase differed substantially from US establishments in the same industries, located in northern border counties. Specifically, Canadian establishments were smaller, more rural, more seasonal, and more resource and labor-intensive, and there were fewer factories using less steam power (Inwood and Keay 2008). There is a natural tendency to compare Canadian economic performance to that of the USA, particularly in a historical, long-run development context. As these stylized facts suggest, this often leads to an unflattering portrayal of the Canadian experience, and traditional narratives typically reflect a decidedly pessimistic perspective (Bliss 1990).

Despite the negative depictions in much of the traditional historiography, there is some evidence supporting a more favorable view of Canadian development. For example, from as early as the 1870s, relative to global averages, Canadian growth was strong, nearly matching US and even British industrial production per person between 1870 and 1913 (Urquhart 1993). Compared to other "new world" economies, including the USA, Canada has also been relatively urban, with a disproportionately large share of its population living in dense, albeit geographically disparate, cities and towns. Canadian total factor productivity (TFP) levels and growth rates have been comparable to the global leaders since the 1870s, and real GDP per capita growth between 1880 and 1913 was faster even than growth in the USA (Harris, et al. 2015). Canada's interactions with the global economy during the first era of globalization support this more optimistic characterization of Canadian performance, with large and growing volumes of net immigration, particularly after the mid-1890s, significant net capital inflows amounting to nearly 10% of GDP, and rapidly expanding trade flows, notably into the US market.

Canadian development defies overly deterministic descriptions, and in its complexity, we can find close parallels over time and across a wide range of countries. The long-run industrial and economic development of the Canadian economy may not have been as obviously and unambiguously exceptional as the US experience, but this marks Canada as more representative of a wide range of countries in, for example, Scandinavia, Australasia, South America, and among currently developing, resource-rich, strongly globalized economies. Added to these features of Canadian growth are the fact that the country is unusually well endowed with high-quality historical evidence from business archives, tax, court, and probate records, trade reports, and census manuscripts and that there was exceptionally rich policy, geographic, institutional, and sectoral variation available for identification, and we can see why the union of complexity, common experience, and evidence allows us to draw broad lessons from the study of Canadian economic history. It also helps to explain why the Canadian experience is unusually well suited to the application of cliometric tools and techniques.

Traditional narrative approaches struggle to identify the causal factors at work in economic environments that are characterized by fluid, rapidly evolving, and endogenously determined forces. The careful use of economic theory and statistical rigor helps us to sort, categorize, and weigh these forces. This chapter provides a survey of some of the fundamental questions and debates within Canadian economic history that have been the subject of cliometric study. More specifically, we review the role of resource-led growth, focusing on the fur trade, the resource curse, and the wheat boom, and the relationship between policy and performance, focusing on the adoption of trade protection, the subsidization of infrastructure building, the impact of immigration and immigration policies, and the impact of policy choice on entrepreneurial decision-making. From our survey we learn that the cliometric study of Canadian economic history has had a transformative effect on our understanding of some of the key contributing factors and episodes affecting Canadian development. The traditional narratives are not necessarily or unequivocally overturned in every case, but the confidence that cliometric tools bring to our analyses clearly illustrates the value of a quantitative, transparent, and structured approach to the study of Canadian economic history.

Resource-Led Growth: Curse or Blessing

Canada has long been a resource-intensive economy, with nearly 20% of domestic GDP originating in resource extraction and processing industries between 1900 and 1999 (Keay 2007). The threat of a "resource curse," therefore, has been significant in a Canadian development context. Using a standard growth equation, with a panel comprised of 69 countries spanning the years 1970–1989, Jeffrey Sachs and Andrew Warner (2001) identify a strongly significant negative relationship between the natural resource intensity of countries' exports and their real per capita income growth. Sachs and Warner famously dubbed this relationship the "curse of natural resources," and although the finding has been the subject of study across a wide

range of countries and time periods, it has proven to be remarkably robust (Auty 2001, Lederman and Maloney 2007). The causal channels through which the curse is supposed to operate are broadly associated with the fear that the pursuit of resource rent draws productive factors away from more growth-enhancing economic activities. Resource rent is considered undesirable as a source of growth due to its association with corruption, weak institutions – particularly with respect to property rights – and poor policy and entrepreneurial decisions (Lederman and Maloney 2007). More generally, resource extraction and processing are not viewed as growth enhancing because these activities are perceived to be relatively low-skill labor-intensive, low-technology capital-intensive, and, therefore, low-productivity. In addition to these structural shortcomings, international prices for resource-intensive products are considered to be inherently more volatile than more elastically supplied, less resource-intensive traded goods, and the Prebisch-Singer hypothesis suggests that the terms of trade for resource exporting nations have been in long-run decline since early in the twentieth century (Cuddington et al. 2007). If we combine these factors with the threat of currency appreciation and input price inflation associated with "Dutch disease," there appears to be ample reason to be concerned about the long-run growth effects of resource specialization.

Sachs and Warner recognize that there is a problem reconciling their contemporary evidence of a curse with the experience of early industrializing nations, such as the USA, the Scandinavian countries, Australia, and, of course, Canada. These nations were strongly resource intensive in production and trade during the late nineteenth and early twentieth centuries (Bhattacharyya and Williamson 2011; Keay 2007). Sachs and Warner (2001) suggest that perhaps these nations were less reliant on resource extraction and processing than more contemporary developing economies or the combination of high transport costs and energy using technological biases around the turn of the twentieth century may have mitigated the worst effects of the curse. Cliometricians have challenged the universality of the curse by integrating resource and growth theories and by working to accumulate a body of quantitative historical evidence, particularly for Canada, that calls into question the causal determinants and predictive capacity of the resource curse narrative.

Cliometric tools have helped to document the Canadian historical experience as a clear case of successful and persistent resource-led industrial development and diversification. W.A. Mackintosh (1923) originally articulated the "staples thesis" as a model describing how globalization and inter-industry externalities foster industrial development and diversification in an initially resource-rich but labor- and capital-poor economy. Export demand and the pursuit of resource rent draw foreign capital into infrastructure projects and resource extraction and processing industries. These industries and projects generate rents, which raise domestic income levels and expand domestic demand for less resource-intensive manufacturing, construction, and service sector industries. Transport cost advantages shift local supply curves for raw material inputs, providing an additional boost to geographically proximate resource using manufacturing industries. The aggregate impact of the resource sectors' direct and indirect contributions to the intensive and extensive

growth and diversification of an economy can be both large and persistent in this model, even as stock depletion occurs.

Keay (2007) uses a four-equation vector autoregressive system, with industry-level price and output data from 1900 to 1999, to test for long-run chronological patterns in Canadian resource output, raw material prices, non-resource-intensive manufacturing output, and service sector output. He finds that increases in domestic resource extraction and processing "Granger caused" reductions in Canadian raw material prices and increases in non-resource-intensive manufacturing and service sector output. Although these results do not provide structural evidence that the staples thesis was operating in Canada through the twentieth century, they are fully consistent with the chronological patterns predicted by the model. Measures of the aggregate economic value of these indirect inter-industry externalities over the 1900–1999 period, and the calculation of annual industry and stage-of-production-specific resource rents, reveal that resource specialization just slightly reduced real GNP/capita growth rates in Canada over the twentieth century (from 2.1% per year to 2.0%) but average income levels were nearly 18% higher as a result of the extraction and processing of Canada's resource endowments (Keay 2007).

The weakly negative impact of resource specialization on Canadian twentieth-century growth is not inconsistent with the more contemporary, panel data evidence of a resource curse, but the strongly positive effect of resource specialization on income levels is not at all what is implied by Sachs and Warner's curse. Again, cliometric tools can help us resolve this apparent inconsistency. Boyce and Emery (2011) augment a basic exogenous macro-growth environment by incorporating key aspects of Hotelling's (1931) optimal extraction model. They show that in panel data, under some reasonable assumptions about relative productivity differences (namely, that non-resource-intensive manufacturing productivity is higher than resource industries' TFP), it is theoretically consistent to suggest that resource specialization will be associated with slightly slower income growth rates but considerably higher income levels. In their model, accumulated resource rents raise income levels, but optimal extraction paths in a hotelling environment reflect a trade-off between rising resource rents (increasing scarcity) and the rate of return on other assets. As a result, initially resource-intensive economies will shift factors of production out of resource industries and into more productive alternate industries more slowly than economies with initially low levels of resource intensity. Thus, in the presence of optimal decision-making, resource specialization may be associated with slower income growth rates, just as the curse literature finds in the contemporary data, but higher income levels, just as Keay finds for Canada over the twentieth century.

Imposing a little economic structure, and the collection of long-run evidence with careful empirical identification, does not necessarily undermine the universality of the resource curse in a modern context, but it does help to reconcile the historical experience of Canada and other early industrializers with the results derived from contemporary panel data. Resource-led growth may be slightly slower than growth in less resource-intensive economies, but over the long run, the average person in a

resource-rich economy will be significantly better off than the average person in a resource poor economy.

Indigenous Peoples and the Fur Trade: Market Signals, Demography, and Depletion

Canada was engaged in resource-intensive activities long before 1900. Our understanding of the early interactions that occurred between the indigenous population of North America and Europeans engaged in the fur trade affects our perceptions of First Nations people and the nature of their relationship with commercial markets. The application of economic theory and statistical analysis to the substantial body of evidence left by French trading companies and, more importantly, by the Hudson's Bay Company, are transforming long-held views and traditional narratives (Ray and Freeman 1978). During the eighteenth century, the Hudson's Bay Company traded for a variety of animal pelts throughout the Hudson's Bay drainage basin, but their primary trade target was beaver fur, which they acquired in exchange for more than 60 distinct European products. Using the annual trade accounts from individual trading posts, which report trade volumes and the rates of exchange between pelts and European goods, Ann Carlos and Frank Lewis (2010) develop a price index for furs. This information about the value of fur earned by trappers over time and across trading posts reveals that the indigenous participants in the recorded trades were more like the "industrious workers" of the emerging Industrial Revolution, rather than the "docile peasants" of pre-modern Europe.

As the price of furs traded at the Hudson's Bay posts steadily increased through the eighteenth century, indigenous trappers dramatically increased their purchases of luxury goods, including cloth, jewelry, and tobacco while maintaining their purchases of necessities. Purchases of alcohol also rose during this period, but they were insignificant relative to the value of the overall trade, and the total volume purchased undermines any suggestion that alcohol had more than a minimal impact on indigenous societies during this era. Carlos and Lewis (1999) combine their evidence on the value of the various furs brought to the posts with a simple logistic stock depletion model for beaver. They show that indigenous trappers were increasing and redistributing their effort in response to changing relative fur prices and rising depletion rates.

The records of the Hudson's Bay Company, and other quantitative evidence, have also been used to re-examine the impact of European traders' introduction of smallpox into the indigenous population in northwestern Canada. Carlos and Lewis (2012) focus on one example of a particular smallpox epidemic that swept through the Hudson Bay region in the early 1780s. Narrative accounts put the death rate among the indigenous population as high as 60–90%. Carlos and Lewis trace the volume of trade in furs during the pre- and post-epidemic years, finding only small reductions and limited price movements. They then estimate total mortality from eye-witness accounts of company men and indigenous traders who visited the affected trading posts, and they reconstruct pre-epidemic indigenous population

levels by calculating the human population that might have been supported by the moose, bison, and other resources in the region. As a final assessment, they appeal to the medical evidence on smallpox mortality rates from other more recent, better documented epidemics. According to Carlos and Lewis, all the quantitative evidence suggests that indigenous mortality, while devastating, was likely closer to 10–20%, rather than 60–90%. This surprising result, like much of the growing body of quantitative, cliometric evidence on the economic lives of First Nations populations in Canada, is shifting and refining our views on the role played by indigenous peoples in long-run growth and resource-led development.

The Wheat Boom: Time Series Analysis and the Identification of Structural Breaks

Shifting our focus from beavers in the eighteenth century to wheat at the very end of the nineteenth, we again see a prominent role attributed to Canada's resource endowment. Between 1896 and Canada's declaration of war against Germany on August 5, 1914, the population of the country grew from just over five million to nearly eight million, GDP in 1900 dollars increased from $655 million to nearly $1.8 billion, and annual gross immigration increased from just over 17 thousand to 400 thousand (Urquhart 1993). This period of historically unprecedented income and population growth has become known as the "wheat boom," in no small part because even the most casual of observers at the time recognized that although growth was clearly accelerating in the aggregate economy, the three western prairie provinces – Manitoba, Saskatchewan, and Alberta – were being fundamentally transformed. Wheat exports increased by more than an order of magnitude between 1896 and 1913; new railroad track, which was being added almost entirely west of Ontario, nearly doubled the national total to just over 29 thousand miles in 1913; and the population of the prairie provinces alone rose from approximately 330 thousand in the early 1890s to more than 1.3 million by 1913.

More recently, cliometric analyses have begun to reveal that Canadian growth during the wheat boom was not as unbalanced as the accounts from the period and the broad national aggregates and proxies seem to suggest. Green and Urquhart (1994) provide a detailed investigation of the macroeconomic evidence from this period. They suggest that growth in Canada did discontinuously increase after the mid-1890s, but this expansion was not necessarily narrowly based on the production and export of wheat alone. Wheat exports from the prairie provinces did begin to rise in the early 1890s following the completion of Canada's transcontinental rail line, the Canadian Pacific Railway, and an increase in the relative price of wheat and flour on international markets, but the largest increases in wheat production came later, in the 1910s and early 1920s. More closely coincident with the boom in income per capita was a dramatic increase in the accumulation of fixed capital, particularly infrastructure and social capital, and a relatively balanced expansion in both the urban-industrial and agricultural sectors. Between 1896 and 1913, the value of gross fixed capital increased from just over 10% of GDP to more than 30%, domestic

saving rates more than doubled, and foreign capital inflows more than tripled. Over the same period, manufacturing maintained its GDP share at approximately 25%, while agriculture's share actually dropped slightly, from approximately 25% to just over 20% (Green and Urquhart 1994).

Looking at migrants' intended destinations upon arrival, Green and Green (1993) find that balanced growth also characterized the labor market during this boom. Their figures show that the urban share of the population increased after 1890, in part because the large inflow of new migrants ended up fairly evenly distributed between central Canadian urban centers and the prairie provinces. The cliometric evidence, therefore, suggests that Canada's wheat boom may in fact be more accurately described as a complex expansion in investment, geographically balanced population growth, terms of trade, and exports. Cliometricians have sought to map out the connections among these macro-variables, documenting the implied structural transformation of the Canadian economy during the late nineteenth and early twentieth centuries. The most common tools applied in this effort have been a range of time series econometric techniques.

Inwood and Stengos (1991), for example, test for the presence of a unit root in Canadian aggregate GNP and gross investment between 1870 and 1985. The presence of unit roots in the macro time series indicates that the series are not trend stationary, which suggests that some shocks, or deviations from the long-run trend, were not transitory. These shocks, therefore, permanently affected the underlying determinants of trend growth in Canada over this period. Inwood and Stengos cannot reject the presence of a unit root over the full 1870–1985 period, but when they impose structural breaks in 1896, 1914, and 1939, essentially segmenting the series into sub-periods and allowing for period-specific trends, unit roots can be safely rejected for both GNP and investment. The implication of this finding is that the start of the wheat boom in 1896, World War I in 1914, and World War II in 1939, all marked permanent breaks in Canada's underlying macroeconomic growth trends. Inwood and Stengos argue that all other shocks to the economy, including the collapse of Canada's terms of trade in the early 1920s, the Great Depression, the post-World War II baby boom, and the oil shocks during the 1970s, were transitory – meaning that growth returned to its underlying, stable trend following the dissipation of these shocks.

This is a powerful result that clearly marks the wheat boom as a transformative episode in the long-run development of the Canadian economy. However, the nature of the underlying growth determinants is not explicitly modeled by Inwood and Stengos, and their discretionary imposition of structural breaks is, in effect, based only on their reading of the historiography on Canadian development. Evans and Quigley (1995) argue that the use of univariate time series models does not allow for tests of the exogeneity of any particular structural break or series of breaks. They also show that trend segments imposed in 1896, 1914, and 1939 are not necessarily the only breaks that lead to the rejection of a unit root in the Canadian GNP and gross investment series. Alternate breaks imposed in chronologically proximate years, and the imposition of additional breaks in, for example, 1920, 1929, 1949, and 1973, can also support the confident rejection of the presence of permanent shocks affecting Canadian growth. In a rejoinder, Inwood and Stengos (1995) respond to Evans and Quigley's critiques of their methodological approach and their choice of 1896, 1914,

and 1939 as the unique set of exogenous structural breaks affecting trend growth in Canada. They demonstrate the importance of segment-specific critical values for their unit root testing procedure, while probing the statistical strength of their unique 1896, 1914, and 1939 break points, and they argue for the a priori exogeneity of the technological, climatic, and international demand shocks that triggers the expansion of wheat production and exports in 1896 and the strictly international origins of the two world wars.

Green and Sparks' (1999) contribution to the wheat boom literature also employs time series techniques to study shifts in the underlying trend determinants of Canadian macroeconomic growth. However, rather than relying on univariate models of GNP and gross investment in isolation, they use a dynamic vector autoregressive (VAR) system, which models the time series connections linking population, real investment, terms of trade, real exports, and aggregate real income in Canada over the period 1870–1939. After establishing stationarity and estimating the cointegrating relationships among their five macro-variables, impulse response functions parameterized from the estimation of their VAR model allow them to map out the dissemination of shocks through their system of equations. They find that the cointegrating coefficients among their five variables, which reflect the long-run equilibrium relationships among them, were distinct during the 1896–1913 period, suggesting that the wheat boom was, in fact, a structural discontinuity in trend growth in Canada. The relative size of the impulse responses reveals the substantial impact that exogenous export market shocks had on real Canadian GNP between 1896 and 1913, the endogenous origins of the real investment shocks, and the key role played by deviations from trend growth in population. Green and Sparks (1999: 57) report that, "(t)he most striking result is the substantial contribution coming from innovations in population. . .(which). . . shifted up the growth path of per capita income by 5.7%."

The value of these time series analyses of the late nineteenth- and early twentieth-century Canadian growth experience, and, more specifically, the adoption of a dynamic, multivariate approach, comes from the confident distinction that can be made among possible sources for the structural discontinuity in the underlying trends. We can not only see that the wheat boom marked a significant structural break – fundamentally transforming the Canadian macroeconomy – but we can quantify the importance of the exogenous shocks to the immigrant inflow into Canada (population growth) and, to a lesser extent, the exogenous shocks to foreign demand for Canadian export products (wheat and flour). These findings not only reveal much about this episode in Canadian development, but we also learn more general lessons about resource-led growth that can be applied in other export-oriented, resource-rich countries.

The Adoption of Protectionism: General Equilibrium and "New Trade" Models

Resource-led growth in a labor and capital-scarce economy is only feasible if international market access can be assured, and Canada's interactions on international markets have been largely determined by the structure of domestic trade

policy. Policy choices in general, and trade policy in particular, have played a key role in Canadian development. For Canada, the impact of policy choice can be identified due to the presence of sharp discontinuities.

Through the 1870s, inter- and intracontinental transport costs were falling sharply, capital and labor flows were rising, and international trade, particularly in industrial products and raw material inputs used in industrial production, was expanding rapidly. In the USA, a deeply divided, Republican-controlled Congress raised average tariff rates from 15% in 1859 to 45% in 1870, and although there were some selective reductions, US rates were maintained at or slightly above 30% for the next two decades (Irwin 2010). In Canada, John A. Macdonald's Conservative Party campaigned in the 1878 national election on a platform that promoted a new "National Policy," which included support for European immigration into the Canadian west, subsidization of a transcontinental rail line entirely within Canadian territory, and the adoption of explicitly protectionist trade policy objectives. The Conservatives won the election, and in their first budget, brought before Parliament in 1879, virtually the entire Canadian tariff schedule was rewritten. Tariff rates averaged over all imports rose from just under 14% to 21%, and there were further tariff increases in 1884 and 1887. Through the 1890s and early 1900s, average rates slowly slid back, but on the eve of Word War I in 1913, the average weighted tariff (AWT) in Canada remained at 18%. Protectionism remained a primary objective for Canadian trade policy until the signing of the free trade agreement with the USA in 1988.

Traditionally, economic historians have characterized the National Policy as a "necessary evil" (Easterbrook and Aitken 1956). The standard narrative recognized that trade protection likely resulted in market power and economic profits for domestic manufacturers, earned at the expense of consumers who faced higher prices as a result of the increased tariff rates, but any negative consequences were thought to have been offset by the positive effects of infant industry protection. Higher tariffs were considered necessary to support import substitution, increase the scale of domestic production, and provide investment incentives for Canadian industries that were struggling against rising imports and closing international (particularly USA) markets.

A revisionist literature shifted the focus away from infant industry arguments, toward static welfare losses in a neoclassical Ricardian trade setting (Easton et al. 1988). The revisionist view concluded that the National Policy tariffs reduced competitive pressures in the Canadian economy, allowing manufacturers to charge prices in excess of their marginal costs, which resulted in reductions in consumer surplus and, therefore, lower social welfare. The size of the estimated static, partial equilibrium welfare losses resulting from the move to protectionism in 1879 varies from approximately 4% of Canadian GDP to 0.6% (Pomfret 1993, Beaulieu and Cherniwchan 2014, Alexander and Keay 2018a). The differences in the size of the measured effects depend critically on the trade elasticity estimates used in the deadweight loss calculations. In general, greater substitutability between domestic and foreign-produced goods leads to greater deadweight losses. Reliable estimation of the required elasticities is problematic in a historical setting due to the need for

highly disaggregated price information, controls for multilateral resistance, and a recognition of potentially endogenous prices (Alexander and Keay 2018a).

Recently, the digitization of the Canadian trade tables from as early as 1867 has encouraged a reinvestigation of the impact of the National Policy tariffs based on cliometric approaches that use the newly available, highly granular, product-specific import, export and tariff data, and new trade models that allow for deviations from the standard partial equilibrium, neoclassical environment. Both the traditional and revisionist literature on Canada's adoption of protectionist policy objectives in 1879 rely on measures of tariff rates averaged over broad categories of products or all import products. Beaulieu and Cherniwchan (2014) were the first to use newly digitized trade data to document the highly selective nature of the tariff changes that were applied by Macdonald's Conservative government. The initial round of tariff increases was characterized by a narrow targeting of manufactured import products destined for final consumption. Non-manufactured raw material products, and products used primarily as intermediate inputs, faced significantly smaller tariff increases. Alexander and Keay (2018b) use theory-consistent specifications derived from Grossman and Helpman's (1994) protection-for-sale model to show that products produced by establishments with the most potential political influence also enjoyed particularly large increases in their protection. Beaulieu and Cherniwchan (2014) derive trade restrictiveness indexes which reveal a tendency, particularly with the 1884 and 1887 revisions of the tariff schedule, to target import products with relatively close domestic substitutes, thereby promoting import substitution, strongly increasing the restrictiveness of Canadian tariffs, and increasing the welfare cost of protection.

Keay (2018) relies on estimating equations based on models of optimal tariff setting in the presence of infant industries with strong learning-by-doing effects, to show that the piecemeal reductions in Canadian tariff rates that occurred after 1890 were just as selective as the tariff increases during the 1880s. The post-1890 tariff cuts appear to have targeted products that showed signs of maturation during the first decade following the initial adoption of protectionism. Maturing products are defined as those with rising net exports, and therefore falling learning potential, from 1880 to 1889. Cutting tariffs on products with lower learning potential reduced deadweight losses due to Canadian trade policy by approximately 0.4% of GDP per year between 1890 and 1913.

As these examples from the literature illustrate, more finely disaggregated trade data not only allow us to get a better sense of the complexity of late nineteenth- and early twentieth-century Canadian trade policy, but they also allow us to assess the impact of protective tariffs with more sophisticated modeling approaches. To derive partial equilibrium, static measures of deadweight loss in a Ricardian trade model, the standard neoclassical assumptions must hold, including perfect competition, market clearing, no externalities, and constant returns to scale. However, the "new trade" models that were developed in the 1980s and 1990s allowed for the assessment of the impact of trade restrictions in economic environments that do not satisfy these standard assumptions (Melitz and Trefler 2012).

Harris et al. (2015) rely on predictions derived from two new trade models to interpret the results from a series of difference-in-differences and treatment intensity specifications that use the newly digitized Canadian trade data from 5-year intervals, spanning the period 1870–1910. The first of these models relaxes the assumption of perfect competition by treating the domestic manufacturing sector as a Cournot oligopoly. In this environment, a tariff can promote a dynamic growth response through the reduction in markups and the exploitation of internal scale economies. The second model focuses on the impact that tariff protection has on the accumulation of experience and the productivity advantages of learning by doing. Treating the imposition of the National Policy as a "natural experiment" in which some producers received larger tariff increases than others, they find that the products with the largest tariff increases in 1879 experienced disproportionately rapid output growth, productivity improvement, and output price reductions between 1890 and 1913. These dynamic growth effects are consistent with the theoretical predictions from the new trade models. Namely, tariff protection allowed domestic output to grow more rapidly, which facilitated a movement down producers' long-run average cost curves, and up their learning curves, such that productivity improvement accelerated and, eventually, output prices fell.

The impact of output expansion, and longer-run productivity and price effects, cannot be captured in static, partial equilibrium measures of the National Policy's impact on the Canadian economy. The employment of new trade models provide us with a new perspective on Canada's early adoption of protectionist objectives. Another new perspective opens when we move away from partial equilibrium models, toward a general equilibrium approach to welfare measurement. Alexander and Keay's (2018a) static, general equilibrium trade model with multiple industries reveals that, even for small open economies like Canada in the late nineteenth century, increases in tariff protection do not necessarily lead to lower aggregate welfare. In a general equilibrium setting, tariff increases do trigger partial equilibrium price distortions, and hence deadweight loss, but they also increase government revenues and positively impact domestic terms of trade. The net effect on welfare need not be negative. Alexander and Keay (2018a) use product-specific trade data from Canada, the USA, and Britain, to show that although the general equilibrium welfare effect of the tariff increases imposed under the National Policy is sensitive to the choice among alternate estimates of historical trade elasticities for Canada, under most reasonable specifications, Canadian welfare likely rose by as much as 0.16% of domestic GDP between 1877 and 1880 as a result of its unilateral adoption of protectionism. Of course, they also show that welfare would have increased substantially more had Canada and its trade partners pursued multilateral free trade during the late nineteenth century, rather than unilateral protectionism.

As the findings in these studies illustrate, our understanding of the sophistication of Canadian trade policy, and the impact of the adoption of protectionism in 1879, has recently undergone a transformation as a result of the accumulation of highly disaggregated import, export, and tariff data, which in turn has facilitated the application of more flexible modeling approaches. The reliance on quantitative evidence analyzed through a cliometric lens and interpreted with theory-consistent

methods has changed the way we view Canadian industrial development and the impact that domestic trade policy had on that development.

Transport Costs: Intracontinental Shipping and the Subsidization of Railways

Canada's adoption of protectionism in 1879 was, at least in part, a response to the pressures associated with falling trade costs and rapid globalization. Trade costs are a function of tariff rates, but transport costs are another key contributing factor. Depending on the route and product being shipped, between 1870 and the start of World War I, global trans-oceanic freight rates fell by between 35% and 85% (Jacks and Pendakar 2010; Mohammed and Williamson 2004). In addition to these reductions in freight rates, there were equally sharp reductions in insurance, wharfage, and brokerage fees (Inwood and Keay 2015). For a trade-dependent nation like Canada, which consistently had gross import and export values totalling approximately 40% of GDP between 1870 and 1913, the economic impact of falling transport costs and rapidly integrating intercontinental goods markets was potentially dramatic.

Inwood and Keay (2015) explore the connection between trade volumes and transport costs in Canada by focussing on a product that has become a key indicator of industrial development during the late nineteenth and early twentieth centuries – pig iron. Between 1870 and 1913, the total cost to move one net ton of pig iron from the docks in Liverpool to the port of Montreal fell from $6.71 to $4.10. What is surprising is that this nearly 40% reduction in trans-oceanic shipping costs did not coincide with an expansion in British exports as a share of the Canadian market. Britain's share of Canadian pig iron consumption fell from 30% in 1870 to less than 5% in 1913. Using an instrumental variable approach to control for the endogenous relationships between transport costs, domestic pig iron prices, and British pig iron shipments, Inwood and Keay estimate flexible, log-linear Armington import demand specifications. Their results reveal that falling trans-Atlantic transport costs did increase British trade in pig iron, but the effect of falling ocean freight rates was swamped by rising tariff rates in Canada and, even more importantly, declining inland water and rail transport costs.

The primary inputs used in the production of pig iron are iron ore and coal. Most Canadian blast furnaces imported iron ore from northern Michigan and coal from Pennsylvania. The transport cost share of Canadian producers' coal and iron ore input prices fell from over 45% in 1870 to just 26% in 1913. The collapsing transport costs that characterize the first era of globalization were a greater benefit to the Canadian pig iron producers trying to compete against British exporters in their home market, than they were to British exporters shipping pig iron across the Atlantic. Inwood and Keay's study of a specific trading relationship illustrates how a detailed, theoretically founded, but narrowly focused analysis can help resolve more general, broadly applicable historical questions. Their findings highlight the key role played by intracontinental railway and inland shipping costs during this period of rapid industrial development and global market integration.

Given the geographic dispersion of Canadian economic activity and the importance of intracontinental transport costs, it is not surprising that railways have long held a prominent position in the study of domestic development in Canada (Lewis and MacKinnon 1987, Carlos and Lewis 1995). The largest and longest continuously operating railway in Canada, the Canadian Pacific Railway (CPR), completed its transcontinental trunk line in 1885. The building of a railway across Canada, entirely on Canadian soil, was one of the three pillars of John A. Macdonald's National Policy. The transcontinental line was a formidable technological achievement that helped to integrate an east-west market to rival the pull of US markets south of the border. The rail link has come to be recognized as a cornerstone of early Canadian economic development. However, the role of Macdonald's Conservative government in the building of the railway has been controversial almost from the birth of the company.

The federal government gave the CPR a vast area of prime agricultural land in the prairie provinces, a large cash grant, unfettered access to construction materials along the line, and a long list of other benefits. Whether or not these subsidies were inefficient and excessive was first addressed by Peter George (1968) and Lloyd Mercer (1973), who used the company's financial reports to make "ex post" rate-of-return calculations. Both authors show that the CPR's private investors ultimately earned significantly more than the opportunity cost of their investment funds. Emery and McKenzie (1996) point out, however, that the decision facing the government and the investors at the time was in fact an "ex ante" one. There was considerable uncertainty surrounding the future of the CPR, and more generally, late-nineteenth-century railway investments were prone to highly volatile, unpredictable returns. Based on a recognition that construction of the new transcontinental line precluded the benefit of waiting for more information, and the resulting resolution of uncertainty, Emery and McKenzie employ a financial options model that captures the impact of both ex ante uncertainty and the irreversible nature of the decision to build. Emery and McKenzie use reasonable assumptions about the ex ante uncertainty faced by private investors, determined by measuring the volatility in the returns earned on comparable US railway investments. They find that the subsidies received by the CPR were not only necessary to get the project started, but they were in fact quite modest, at least in light of the risks associated with the commitment to privately fund such a massive infrastructure project. The National Policy's railway subsidization, therefore, was probably just adequate to spur the private construction of the railway, which contributed to the integration of the widely dispersed domestic market, lowered intracontinental shipping costs, and, as a result, promoted industrial development. As this finding suggests, with respect to the National Policy's tariff and railway building objectives, the exoneration of government decision-makers and the longer-run development consequences of their policy choices have become clear only with the application of cliometric tools in micro-level studies of narrowly identified industries (pig iron) and individual firms (the CPR).

Immigration: Self-Selection and Assimilation

The third and final pillar of John A. Macdonald's National Policy was the promotion of immigration into Canada. Nearly as many migrants moved south across the open border to the USA as arrived from Europe, so through most of the late-nineteenth-century, Canadian net migration was near zero. On average, between 1870 and 1895, for example, over 55,000 new arrivals were registered at Canadian ports and overland border crossing points in each year. However, 60,000 left the country annually over the same period. This net loss of population and human capital to the USA came to an abrupt end not with the passage of the National Policy in 1879 but with the start of the wheat boom in 1896. Gross immigration increased by more than 18% per year between 1896 and World War I, while net immigration became strongly positive, with new immigrants outnumbering emigrants in each year by more than 50,000 over this period (Green and Urquhart 1994). Although net immigrant flows fell slightly during the interwar period, between 1896 and 1930, immigration played a key role in boosting Canada's population by more than a factor of two, to just over ten million. This period of growth transformed Canadian labor markets and the demographic characteristics of the population as a whole.

Traditional narratives focus on the role played by the National Policy's direct subsidization of transport costs and the maintenance of open immigration rules in encouraging Ukrainian, Russian, Scandinavian, and more generally eastern European farmers and agricultural laborers destined for the wheat fields in the newly settled prairie provinces. Alan Green and David Green (1993) use both census data and micro-data drawn from ship manifests, which includes detailed information on the characteristics and intended destinations of individual migrants and their families, to explore how new arrivals in 1912 were distributing their human capital across Canadian labor markets. A multinomial logit model links the probability of migrants' intended destinations to their personal and family characteristics and the characteristics of their destination labor markets. Green and Green (1993) find that the new arrivals did not disproportionately favor western or agricultural labor markets, but instead they sought locations and occupations that not only matched their skills but also satisfied sectoral and regional labor market demands. One implication of this result is that even the large inflows of migrant labor during the post-1896 period did not substantively impact Canadian wage distributions.

Green and Green (2016) confirm and extend this finding by using random samples from the population census manuscripts from 1911 through 1941 to estimate age-occupation-birth place-specific elasticities of substitution, derived from a nested CES production model. Again, they find that through some combination of self-selection, post-arrival redistribution, and general equilibrium wage effects, the market-specific human capital brought by migrants roughly matched the existing distribution of human capital across Canadian markets.

Of course, a balanced distribution of the new arrivals' human capital, and evidence of responsiveness to market signals, does not necessarily imply either

smooth or rapid assimilation of immigrants into the Canadian labor market. Alan Green and Mary MacKinnon (2001) use samples from the early twentieth-century population census manuscripts for Toronto and Montreal to probe the labor market experiences of the foreign-born relative to the native born. Their findings suggest that unemployment for the foreign born was both higher and more persistent during the Great Depression, and estimates from fully flexible age-earning profiles reveal very slow wage assimilation at the start of the twentieth century, in part due to the disproportionately low number of clerical positions held by the foreign born. Looking more narrowly at British and Irish immigrants, Dean and Dilmaghani (2016) employ a pseudo-cohort methodology with census manuscript samples from 1901 and 1911, which allow them to draw a distinction between entry and assimilation effects. They also find slow assimilation, and they can show that initial earnings as immigrants first entered Canadian labor markets fell over successive post-1896 immigrant waves. Inwood et al. (2016) extend the time series evidence on assimilation, again estimating age-earning profiles using census manuscript samples from 1911, 1921, and 1931. They find relatively rapid immigrant assimilation in 1911 and 1921 but a sharp reversal in 1931. Even among the cohorts who had immigrated many years earlier, earning declines during the Great Depression were much steeper for the foreign born than the native born. Again, these findings indicate the extent to which labor market adjustment and self-selection forces were at work in domestic markets.

Alex Armstrong and Frank Lewis (2012) rely on evidence from ships' passenger manifests to look more closely at the selection of migrants arriving in Canada during the 1920s. They note that average earning differences between Canadian labor markets and European home markets could be as high as 400%. This could be taken as evidence of strong self-selection from the very top of the human capital distribution in migrants' home countries. However, drawing insights from a life-cycle model in which a taste preference for the home country is introduced and borrowing is assumed to be constrained, Armstrong and Lewis show that large European-Canadian average earning differentials do not necessarily reflect strong self-selection tendencies. In their model, potential migrants must save for several years before acting on their decision to migrate, in part because the adjustment period upon arrival can be long and costly. New arrivals could face a decline in occupational status, especially those from poorer countries, and there will almost certainly be a loss of amenities associated with immigrants' home markets. Together, these factors suggest that market-specific characteristics and human capital are important, but very high earning differentials were required simply to trigger migration, without necessarily reflecting severe self-selection. These insights into immigrants' decisions to move, their experiences with assimilation, the role of policy, and their impact on Canadian labor markets and earning distributions are only revealed as a result of the careful application of theoretical and statistical cliometric tools, which in turn can only be used where reliable micro-data has been carefully compiled and digitized.

Entrepreneurial Failure: Measuring Productivity and Technological Change

M.C. Urquhart (1993) has provided us with income per capita estimates for Canada that extend from 1870 to 1926, the first year that official *Dominion Bureau of Statistics* GDP figures are available. As a result of Urquhart's monumental effort to produce these estimates, we have a continuous and consistent basis for comparing Canadian macroeconomic performance to a wide range of international benchmarks over virtually the entire span of industrial development in Canada. Geographic, cultural, historical, and institutional proximity makes the USA by far the most obvious and most common comparator. Over the very long run, Canadian GDP per capita has been consistently lower than American GDP per capita. Depending on exactly how prices are measured, the average income gap between Canada and the USA has been stubbornly stable at something between 10% and 30%, in favor of the USA, with convergence rates statistically indistinguishable from zero (Urquhart 1993).

Standard macroeconomic growth models tell us that income growth and convergence depend critically on the accumulation of physical capital, human capital, and technological development. At least since the mid-eighteenth century, most capital accumulation and technological change has originated from within industrial production processes. As a result of these considerations, economic nationalists have confidently blamed domestic industrialists for the development and persistence of Canada's income per capital shortfall relative to the USA (Williams 1994). The "industrial failure" narrative focuses on the perceived consequences of government trade, industrial, and migration policy decisions stretching back to the National Policy in 1879 (Eastman and Stykolt 1967). These policy choices are said to have led to a wide range of economic, political, and social ills, including excessive US ownership of Canadian capital and poor entrepreneurial decision-making. The evidence offered in support of this traditional view is typically anecdotal descriptions of "backward" or "parochial" technologies, drawn from business histories and biographies (Naylor 2006, Bliss 1990), or highly aggregated information on labor productivity, scale, and output prices (Baldwin and Gorecki 1986).

Problems with these claims of long-run, persistent industrial failure in Canada become evident when we interpret this view through a cliometric lens, employing the tools of microeconomic theory and careful quantitative analysis. Since the late nineteenth and early twentieth centuries, labor and power, more specifically electricity, have been relatively cheap in Canada compared to the USA, while capital has been expensive (Wylie 1990, Keay 2000a, Inwood and Keay 2008, Harris et al. 2015). Because these input price differences reflect differences in geographic endowments, institutional differences, and differences in policy, we can reasonably assume that they have been taken as given and hence considered exogenous by individual decision-makers within Canada's industrial establishments. Cost minimization in the presence of well-defined industrial technology suggests that optimal decision-making can lead to quite dramatically different input employment

combinations, technological choices, and output decisions when exogenous market conditions, particularly input market conditions, differ across production units. In other words, low labor productivity, low output levels, and the use of industrial technology deemed to be well below the technological frontier by business historians or economic nationalists, mainly because it differs from US technology, may not necessarily be evidence of poor decision-making and industrial failure. Better tests of industrial performance require the assessment of technological choice, input employment decisions, and the productivity of entire production processes, conditional on the presence of domestically unique input market conditions.

Tornqvist productivity indexes are geometric averages of relative labor, capital, and intermediate input partial factor productivities, with factor weights derived from cost shares (if we assume competitive market conditions) or econometrically estimated from flexible production function specifications. These total factor productivity indexes allow us to compare performance across production units – Canadian and US manufacturing industries, for example – and over time (Keay 2000b, Inwood and Keay 2008). Inwood and Keay (2008) construct these indexes for 1870, with carefully matched Canadian and US census manuscript micro-data on industrial establishments in Ontario and US border counties in New York, Ohio, Pennsylvania, and Michigan. They find that among establishments in the same industries, located along the Canada-US border, the average TFP difference favored US producers by less than 10%. This total factor productivity gap was so small despite clear evidence that in 1870 Canadian industrial establishments were producing nearly 28% less output per worker; they were smaller and more seasonal; they used less capital but more raw materials; they used less steam or water power; there were far fewer factories; and Canadian markets were considerably thinner than US border county markets (Inwood and Keay 2005).

The comparison of Canadian and US industrial productivity can only be extended over a longer time period for a much narrower cross section. Keay (2000b) draws firm-level evidence from *Moody's Industrial Manuals* for 78 firms representing 9 Canadian and US manufacturing industries. His productivity comparisons reveal remarkable persistence in the North American TFP gap, with an average difference of less than 10% through the decades from the 1910s until the 1990s. The industrial failure narrative in Canada is clearly not consistent with this evidence of near parity in Canadian and US TFP performance from as early as the 1870s until the very end of the twentieth century. Reconciling this evidence requires a detailed analysis of longer-run patterns in technological change and input employment.

Generalized Leontief and translog functional forms are flexible in the sense that they allow for the estimation and assessment of production technology that can evolve independently across production units and over time. Input employment, therefore, can be idiosyncratic, and technological change can be strongly non-neutral within these production and cost function specifications. This flexibility allows for the comparison of industrial performance – input employment decisions, technological choice, and total factor productivity – between production units that operate in widely varying economic environments, such as those characterized by Canadian and American input markets through the late nineteenth and twentieth centuries.

Wylie (1990) estimates systems of input demand functions for a sample of Canadian and US manufacturing industries derived from underlying translog production function specifications. They reveal that between 1900 and 1929, Canadian producers were using technologies that were strongly biased in favor of electricity use, and these biases were strongest relative to the USA, where Canadian electricity prices had fallen most steeply. Keay (2000a, b) finds the same tendency toward domestically unique technological biases that reflect relative input price differences, this time favoring labor use and capital saving among Canadian producers, when he estimates systems of input demand equations derived from generalized Leontief cost functions over the 1902–1990 period. Keay's input demand estimates also capture cross-border differences in input employment, with labor and raw material intensity rising among the Canadian establishments as the relative prices of these inputs declined, and capital intensity falling as relative capital costs in Canada increased over the twentieth century. Even cross-border differences in adoption patterns for new agricultural technologies during the first half of the twentieth century, including the tractor, can be confidently linked to changes in the price of fuel, labor, and capital in the Canadian prairie provinces relative to the US mid-western states (Lew and Cater 2018).

Guided by theoretical predictions of optimal decision-making under cost minimization, the estimation of flexible underlying production and cost function specifications, and the measurement of total factor productivity, a new understanding of the impact of policy choice on Canadian industrial performance has evolved that is at odds with the traditional economic nationalists' failure narrative. In response to idiosyncratic, potentially policy-induced price differences, manufacturers in Canada adopted and adapted their technologies, and employed their inputs in a way that favored relatively cheap labor and raw materials, while conserving on expensive capital. We cannot observe the impact of these decisions by looking narrowly at partial factor productivities or by relying on descriptive characterizations of "backward" technological choices. Cliometric tools allow us to search for evidence of sound decision-making and industrial success among Canadian manufacturers, and the evidence reveals that since at least 1870, technological and input employment choices fostered technical efficiency and total factor productivity growth that nearly matched the global technological leaders.

Concluding Remarks

Long-run economic and industrial development in Canada has been characterized by discontinuous changes in policy, technology, and industrial structure, which have triggered complex and often endogenously determined consequences. These discontinuities, combined with remarkably good historical evidence and the application of cliometric tools and techniques, allow us to ask a wide range of historical questions about the Canadian experience. The presence of close historical and contemporary parallels makes these questions broadly relevant across time and space. The coincidence of complexity, common experience, and exceptional evidence that characterizes Canadian economic history allows us to revisit accepted, long-held traditional

narratives. From the survey in this chapter, it is clear that the evidence, and economic theory, is often, but not always, inconsistent with these narratives.

Looking at resource-led development, we find that the presence of diversifying linkages can help to overcome the corrupting effects of resource rent and price volatility. In addition, the trade in beaver pelts reveals dynamic, responsive decision-making within indigenous peoples' commercial transactions, and population and export booms associated with the emergence of the prairie wheat economy triggered structural changes in Canada's underlying growth trends.

When we turn our attention to the impact of policy choice, we find that Canada's selective adoption of protectionism under the National Policy in 1879 was associated with dynamic productivity effects, infant industry protection, and possibly even welfare improvement. Overland transport costs played a key role in the first era of globalization's transport cost revolution, and the Canadian government's efforts to support domestic railway building appear to have been both effective and efficient. We can also see that the maintenance of a relatively open immigration policy into the twentieth century promoted the assimilation of foreign-born human capital into Canadian labor markets. Finally, the evidence suggests that there is little support for claims of policy-induced entrepreneurial failure among Canadian manufacturers.

While this review of the impact that cliometrics has had on Canadian economic history is admittedly incomplete, it does illustrate the value of structured analysis and the power of statistical rigor. By working to identify causal factors in the complex, fluid, and rapidly evolving economic environment that characterizes long-run Canadian development, cliometricians have transformed our understanding of Canada's economic history, which in turn has affected broader perspectives on development in resource-rich, rapidly globalizing economies.

Cross-References

- ▶ Agricliometrics and Agricultural Change in the Nineteenth and Twentieth Centuries
- ▶ Cliometric Approaches to International Trade
- ▶ International Migration in the Atlantic Economy 1850–1940
- ▶ Property Rights to Frontier Land and Minerals: US Exceptionalism
- ▶ Railroads
- ▶ The Census of Manufactures: An Overview
- ▶ The Cliometric Study of Innovations
- ▶ Trends, Cycles, and Structural Breaks in Cliometrics

References

Alexander P, Keay I (2018a) A general equilibrium analysis of Canada's national policy. Explor Econ Hist, forthcoming

Alexander P, Keay I (2018b) Responding to the First Era of Globalization: Canadian Trade Policy, 1870–1913. Unplublished manuscript at Journal of Economic History

Armstrong A, Lewis FD (2012) International migration with capital constraints: interpreting migration from the Netherlands to Canada in the 1920s. Can J Econ 45:732–754
Auty R (2001) The political economy of resource driven growth. Eur Econ Rev 45:839–846
Baldwin J, Gorecki P (1986) The role of scale in Canada/US productivity differences in the manufacturing sector, 1970–79, research study volume 6, Macdonald commission. U of T Press, Toronto
Beaulieu E, Cherniwchan J (2014) Tariff structure, trade expansion and Canadian protectionism from 1870-1910. Can J Econ 47:144–172
Bhattacharyya S, Williamson JG (2011) Commodity price shocks and the Australian economy since federation. Aust Econ Hist Rev 51:150–177
Bliss M (1990) Northern Enterprise: five centuries of Canadian business. McClelland-Stewart, Toronto
Boyce J, Emery JC (2011) Is a negative correlation between resource abundance and growth sufficient evidence that there is a resource curse? Resources Policy 36:1–13
Carlos AM, Lewis FD (1995) The creative financing of an unprofitable enterprise: the Grand Trunk Railway of Canada. Explor Econ Hist 32:273–301
Carlos AM, Lewis FD (1999) Property rights, competition and depletion in the eighteenth-century Canadian fur trade: the role of the European market. Can J Econ 32:705–728
Carlos AM, Lewis FD (2010) Commerce by a Frozen Sea: native Americans and the European Fur trade. University of Pennsylvania Press, Philadelphia
Carlos AM, Lewis FD (2012) Smallpox and native American mortality: the 1780s epidemic in the Hudson Bay region. Explor Econ Hist 49:277–290
Cuddington JT, Ludema R, Jayasuriya SA (2007) Prebisch-Singer redux. In: Lederman D, Maloney WF (eds) Natural resources: neither curse nor Destiny. Stanford University Press and The World Bank, Washington, DC, pp 103–140
Dean J, Dilmaghani M (2016) Economic integration of pre-WWI immigrants from the British Isles in the Canadian labor market. International Migration and Integration 17:55–76
Easterbrook WT, Aitken HGJ (1956) Canadian economic history. Macmillan, Toronto
Eastman H, Stykolt S (1967) The tariff and competition in Canada. Macmillan, Toronto
Easton ST, Gibson WA, Reed CG (1988) Tariffs and growth: the dales hypothesis. Explor Econ Hist 25:147–163
Emery JC, McKenzie KJ (1996) Damned if you do and damned if you don't: an option value approach to evaluating the subsidy of the CPR mainline. Can J Econ 29:256–270
Evans LT, Quigley NC (1995) What can univariate models tell us about Canadian economic growth, 1870-1985. Explor Econ Hist 32:236–252
George PJ (1968) Rates of return to railway investment in Canada and implications for government subsidization of the Canadian Pacific Railway: some preliminary results. Can J Econ 1:740–762
Green AG, Green DA (1993) Balanced growth and the geographical distribution of European immigrant arrivals to Canada, 1900-1912. Explor Econ Hist 30:31–59
Green AG, Green DA (2016) Immigration and the Canadian earnings distribution in the first half of the twentieth century. J Econ Hist 76:387–426
Green AG, MacKinnon M (2001) The slow assimilation of British immigrants in Canada: evidence from Montreal and Toronto, 1901. Explor Econ Hist 38:315–338
Green AG, Sparks GR (1999) Population growth and the dynamics of Canadian economic development: a multivariate time-series approach. Explor Econ Hist 36:56–71
Green AG, Urquhart MC (1994) New estimates of output growth in Canada: measurement and interpretation. In: McCalla D, Huberman M (eds) Perspectives on Canadian economic history, 2nd edn. Copp Clark Longman, Toronto, pp 158–175
Grossman GM, Helpman E (1994) Protection for sale. Am Econ Rev 84:833–850
Hamilton GC, Keay I, Lewis FD (2018) Contributions to Canadian economic history: the last 30 years. Can J Econ 50:1596–1632
Harris R, Keay I, Lewis FD (2015) Protecting infant industries: Canadian manufacturing and the National Policy, 1870-1913. Explor Econ Hist 56:15–31
Hotelling H (1931) The economics of exhaustible resources. J Polit Econ 39:137–175
Inwood K, Keay I (2005) Bigger establishments in thicker markets: can we explain early productivity differentials between Canada and the United States? Can J Econ 38:1327–1363

Inwood K, Keay I (2008) The devil is in the details: assessing early industrial performance across international borders using late 19th century North American manufacturers as a case study. Cliometrica 2:85–118
Inwood K, Keay I (2015) Transport costs and trade volumes: evidence from the trans-Atlantic iron trade, 1870-1913. J Econ Hist 75:95–124
Inwood K, Minns C, Summerfield F (2016) Reverse assimilation? Immigrants in the Canadian labor market during the great depression. Eur Rev Econ Hist 20:299–321
Inwood K, Stengos T (1991) Discontinuities in Canadian economic growth, 1870-1985. Explor Econ Hist 28:274–286
Inwood K, Stengos T (1995) Rejoinder: segmented trend models of Canadian economic growth. Explor Econ Hist 32:253–261
Irwin D (2010) Trade restrictiveness and deadweight losses from US tariffs. Am Econ J Econ Pol 2:111–133
Jacks D, Pendakar K (2010) Global trade and the maritime transport revolution. Review of Economics and Statistics 92:745–755
Keay I (2000a) Scapegoats or responsive entrepreneurs: Canadian manufacturers, 1907-1990. Explor Econ Hist 37:217–240
Keay I (2000b) Canadian manufacturers' relative productivity performance: 1907-1990. Can J Econ 33:1049–1068
Keay I (2007) The engine or the caboose? Resource industries and twentieth-century Canadian economic performance. J Econ Hist 67:1–32
Keay I (2018) Protection for maturing industries: Evidence from Canadian Trade Patterns and Trade Policy, 1870–1913. Unplublished manuscript at Canadian Journal of Economics
Lederman D, Maloney WF (2007) Neither curse nor destiny: introduction to natural resources and development. In: Lederman D, Maloney WF (eds) Natural resources: neither curse nor Destiny. Stanford University Press and The World Bank, Washington, DC, pp 1–14
Lew B, Cater B (2018) Farm mechanization on an otherwise featureless plain: tractor adoption on the northern great plains and immigration policy of the 1920s. Cliometrica. forthcoming
Lewis FD, MacKinnon M (1987) Government loan guarantees and the failure of the Canadian northern railway. J Econ Hist 47:175–196
Mackintosh WA (1923) Economic factors in Canadian economic history. Canadian Historical Review 4:12–25
Melitz MJ, Trefler D (2012) Gains from trade when firms matter. J Econ Perspect 26:91–118
Mercer LJ (1973) Rates of return and government subsidization of the Canadian Pacific Railway: an alternative view. Can J Econ 6:428–437
Mohammed S, Williamson JG (2004) Freight rates and productivity gains in British tramp shipping, 1869-1950. Explor Econ Hist 41:172–203
Naylor RT (2006) The history of Canadian business, 1867–1914. McGill-Queen's Press, Montreal (first published 1975)
Pomfret WT (1993) The economic development of Canada. Nelson, Toronto
Ray AJ, Freeman D (1978) Give us good measure: an economic analysis of relations between the Indians and the Hudson Bay company before 1763. University of Toronto Press, Toronto
Sachs J, Warner A (2001) Natural resources and development: the curse of natural resources. Eur Econ Rev 45:827–838
Urquhart MC (1993) Gross National Product Canada, 1870–1926: the derivation of the estimates. McGill-Queen's Press, Kingston
Williams G (1994) Not for Export (revised edition). McClelland-Stewart, Toronto (first published 1979).
Wylie PJ (1990) Scale-biased technological development in Canada's industrialization, 1900-1929. Rev Econ and Stat 72:219–227

Part II

Human Capital

Human Capital

Claudia Goldin

Contents

Abstract

Human capital is the stock of skills that the labor force possesses. The flow of these skills is forthcoming when the return to investment exceeds the cost (both direct and indirect). Returns to these skills are private in the sense that an individual's productive capacity increases with more of them. But there are

C. Goldin (✉)
Department of Economics, Harvard University and National Bureau of Economic Research, Cambridge, MA, USA
e-mail: cgoldin@harvard.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_23

often externalities that increase the productive capacity of others when human capital is increased. This essay discusses these concepts historically and focuses on two major components of human capital: education and training, and health. The institutions that encourage human capital investment are discussed, as is the role of human capital in economic growth. The notion that the study of human capital is inherently historical is emphasized and defended.

Keywords

Nutrition · Economic growth · Training · Education · Health · Hemographic transition · Human capital · Malthusian equilibrium · Institutions · Slavery · Indentured servitude · Formal schooling · School enrollment · Return to schooling · Compulsory education · High school · Academy · School district · rate bill · High school movement · Health human capital · Antibiotics · Age of modern medicine · Public health interventions

Human Capital and History

For much of recorded history, income levels were low, lives were short, and there was little or no economic growth. We now have healthier, longer, richer, and hopefully happier lives. The regime shift involved increased knowledge and its diffusion, greater levels of training and education, improved health, more migration, fertility change, and the demographic transition. In short, the process involved advances in *human capital*.

What Is Human Capital?

Human capital is defined in the *Oxford English Dictionary* as "the skills the labor force possesses and is regarded as a resource or asset." It encompasses the notion that there are investments in people (e.g., education, training, health) and that these investments increase an individual's productivity.

We use the term today as if it were always part of our lingua franca. But it wasn't. Not that long ago, even economists scoffed at the notion of "*human* capital." As Theodore Schultz noted in his American Economic Association presidential address in 1961, many thought that free people were not to be equated with property and marketable assets (Schultz 1961). To them, that implied slavery.

But the concept of human capital goes back at least to Adam Smith. In his fourth definition of capital, he noted: "The acquisition of . . . talents during . . . education, study, or apprenticeship, costs a real expense, which is capital in [a] person. Those talents [are] part of his fortune [and] likewise that of society" (Smith 2003, orig. publ. 1776).

The earliest formal use of the term "human capital" in economics is probably by Irving Fisher in 1897.[1] It was later adopted by various writers but did not become a

[1]Fisher cites J.S. Nicholson, "The Living Capital of the United Kingdom," for the term "living capital" as opposed to "dead capital."

serious part of the economists' lingua franca until the late 1950s. It became considerably more popular after Jacob Mincer's 1958 *Journal of Political Economy* article "Investment in Human Capital and Personal Income Distribution." In Gary Becker's *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, published in 1964 (and preceded by his 1962 *Journal of Political Economy* article, "Investment in Human Capital"), Becker notes that he hesitated to use the term "human capital" in the title of his book and employed a long subtitle to guard against criticism[2] (Becker 1962, 1964).

Schultz's article (1961) demonstrates the importance of the concept of human capital in explaining various economic anomalies. Some are easy to figure out, such as why both migrants and students are disproportionately young persons. Some are more difficult, such as why the ratio of capital to income has decreased over time, what explains the growth "residual," and why Europe recovered so rapidly after World War II. Some are even more difficult, such as why labor earnings have risen over time and why they did not for much of human history. As is clear from most of these issues, the study of human capital is inherently historical.

Why the Study of Human Capital Is Inherently Historical

Robert Solow's pioneering work on economic growth in the 1950s led to the formulation of growth accounting and the discovery (or uncovering) of the "residual."[3] Solow (1957), working with data from 1909 to 1949, demonstrated that the residual was 87.5% of total growth in per capita terms. The residual is that portion of economic growth that the researcher cannot explain by the increase in physical productive factors such as the capital stock, the number of workers, and their hours and weeks of work.

The size of the residual during much of the twentieth century relative to economic growth in per capita or per worker terms demonstrated that physical capital accumulation did not explain much of growth and that something else did. That something else is knowledge creation and the augmentation of the labor input through education and training. In other words, much of the residual was due to the increase in human capital.

Some researchers devised methods to close the "residual" gap by adding human capital growth to the Solow model (Mankiw et al. 1992). Others demonstrated that the growth of knowledge and other "non-rival" goods meant that some of the implications of the Solow model were violated (Jones and Romer 2010).

Among the most important findings regarding economic growth over the long run, and the one most relevant to the study of human capital in history, is that the

[2]A Google "N Gram" of the term "human capital" reveals that there was virtually no usage in the English language until the late 1950s. After the 1950s the usage of the term increased until today, with a somewhat greater uptick in the 1990s than previously.

[3]For an understanding of the "residual" in economic growth, see the original Solow (1957) article or an economic growth theory textbook such as Barro and Sala-i-Martin (2003).

residual has greatly increased over time. Physical capital accumulation and land clearing explain a substantial fraction of economic growth in the past. But they do far less well in the more modern era. As a fraction of the growth of income per capita in US history, the residual has increased from about 57% for the 1840–1900 period to around 85% for the 1900–1980s period.[4]

The residual can be reduced by about 20% for the 1900–1980s period by accounting for the growth in human capital embodied in individuals.[5] But growth in human capital does little to reduce the residual for the earlier period. In large measure the reason that human capital advances explain more economic growth in the twentieth century than the nineteenth century is because education advances were slower. That is, there simply was not a lot of human capital formation in the earlier period. Exactly why schooling levels advanced in the late nineteenth century is discussed in the section on education below. But another reason is because the productivity increase from higher levels of education was probably less.[6]

The inclusion of human capital in growth accounting treats increases in education as enhancing the productivity of individuals. Differential productivity is measured by how much higher earnings are for workers of different levels of education. That is, earning ratios by education (e.g., college/high school graduates) are held constant and the fractions of workers with different levels of education are allowed to change over time. These relative "prices" can be updated in the same way that prices are changed in chain-weighted prices for commodities.[7]

The impact of education would be considerably larger, and the residual smaller, if non-private aspects of human capital accumulation were included. These non-private aspects of human capital include spillovers across firms from increased knowledge, lower amounts of criminal activity in society, and greater innovation because there are more smart and informed people.

Another way in which the study of human capital is inherently historical concerns the origins of the "knowledge economy" (Mokyr 2004). Knowledge evolved historically beginning with observations about *natural phenomenon* – the elemental discoveries or the "what" of knowledge. These include "my headache goes away when I chew on the bark of the willow tree."[8] Knowledge then shifted from "*what* is it?" to answering "*how* does it work?" This knowledge involved generalizations and scientific findings. The willow tree contains (acetyl) salicylic acid, which is an anti-inflammatory and anticlotting drug. Aspirin was made out of this substance in the early 1900s. Knowledge then advanced to a deeper understanding of the "how" and

[4]See the calculations in Robert Gallman's chapter in Davis et al. (1972) and those in Denison (1962).

[5]The calculation is larger in Denison's work than in Goldin and Katz (2008). But both of these are a lower bound for a host of reasons including the endogenous nature of capital and, most importantly, the externalities from having a more educated workforce and population.

[6]The data needed to assess this point are very thin and consist of earnings for various occupations.

[7]See Goldin and Katz (2008, Table 1.3).

[8]Hippocrates left records of this finding.

only in 1971 did researchers figure out that the anti-inflammatory response occurred because of suppression of prostaglandins.

An important part of the creation of knowledge is diffusion of the initial "what." In premodern periods, the existence of large numbers of people living in close proximity was important to the maintenance of knowledge. Widely dispersed settlements, on the other hand, meant that chance discoveries would be less likely to spread and to be built upon. Later advances, such as the printing press, books, scholarly societies, and formal schools, helped preserve knowledge and spread discoveries. The notion that denser populations enhance the spread of knowledge and heighten innovation is important to understanding how humans escaped the Malthusian trap and why investments in human capital were worthwhile.

Human Capital and Economic Growth

Human Capital and Economic Performance in the Long Run: Escaping Malthus

According to many economic historians, real wages in Europe were stagnant from at least 1200 to about 1800 (Allen 2001; Clark 2005, 2007a, b). As can be seen in Fig. 1, real wages may have been stagnant, but they were not unchanging during those centuries. The real wages of both agricultural laborers and building craftsmen rose when population decreased, as during the Black Death (peaking around 1350), and they fell as populations rebounded. They varied, as well, due to agricultural vicissitudes. But, on average, they changed little. World population increased, but only slightly from around −5000 BC until around 1800 AD (see Fig. 2).

By and large, the data series in Figs. 1 and 2 point to a classic Malthusian equilibrium – stagnant real wages during long periods, small increases in population, and occasional periods of real wage growth followed by increased population and subsequent decreased wages. The Malthusian problem was twofold: a fixed amount of resources in the form of land and no fertility controls.

But sustained growth in real income per capita and in real wages is apparent in mid-nineteenth century Europe (see Figs. 2 and 3) and somewhat earlier in North America. Population growth had been extremely low but increased enormously in the period just after the "industrial revolution." The demographic transition set in at various moments in Europe and North America. It occurred in the United States and France in the early 1800s, in parts of Europe later in the nineteenth century, and in other parts of Europe as late as the early twentieth century.

By the nineteenth century many parts of Europe, the Western Hemisphere, and elsewhere had entered the modern era of economic growth and had escaped the Malthusian trap. How the regime change came about is one the most important issues in economic history. The answer mainly concerns technological change and the fertility transition. Underlying both of these transformations is the concept of human capital. Without knowledge embodied in people, there can be no

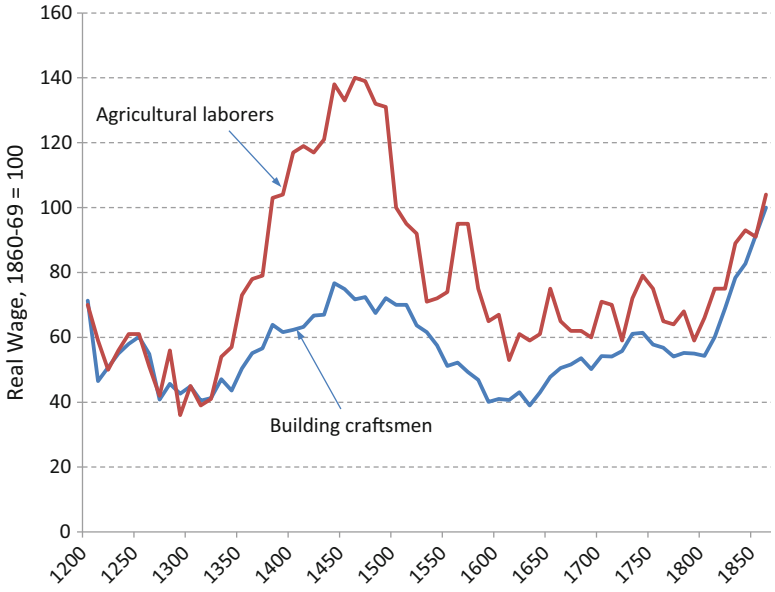


Fig. 1 English laborer's real wages, 1209–1865 (Sources and Notes: Clark (2007b), Table A2 for agricultural laborer real day wages and Clark (2005), Tables A2 for building craftsman real wages. Both series set 1860–69 = 100. The building craftsman series is for decades, the midpoint of which is used here. Agricultural wages are given at approximately annual intervals, and the midpoint numbers for the decades are used. The higher real wage for agricultural wages during much of the period shown may be due to the greater uncertainty during the year and the fact that these are daily wages)

technological change. Without an increase in the value of each child, parents will opt for quantity over quality.

One way of reconciling the historical facts is through an insightful endogenous growth model, pioneered by Galor and Weil (2000) and expanded in Galor (2011). Central to their model is the emerging role of human capital.

The model contains three regimes, and the decision makers are parents who determine how many kids to have and how much to invest in each. At the outset there are low levels of income, no schooling, no income growth, and a very low increase in population. As population increases, technology advances (recall that the "what" of knowledge diffuses with larger, denser populations). Even small levels of technological change increase incomes and induce parents to allocate some of their resources to school their children. Education increases, which in turn boosts technological change, income, and population. At some point intensive growth, a demographic transition, and sustained growth per capita, all become possible, and the world escapes the Malthusian trap.

Human capital is fundamental to the Galor-Weil model. A greater and denser population increases technological change because of the notions about knowledge creation, discussed above. Technology complements skill and increases the returns

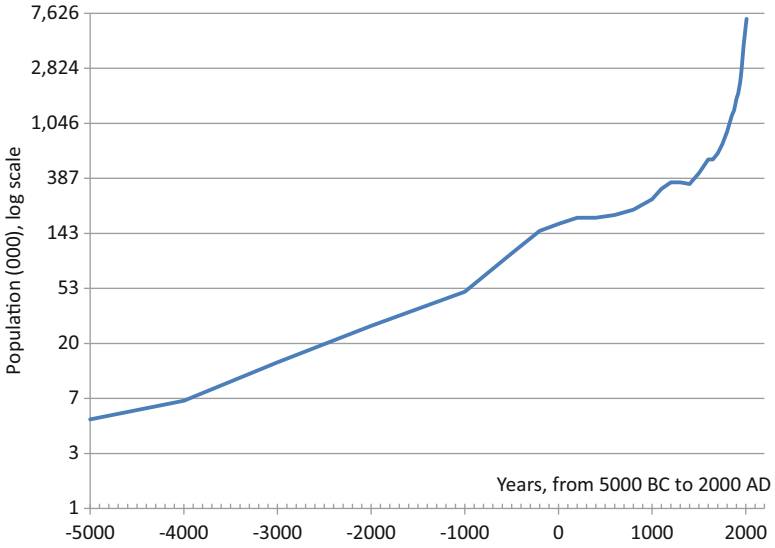


Fig. 2 World population (000) from the Neolithic era to 2010 (Source: Kremer (1993), Table 1 and recent world population estimates after 1980)

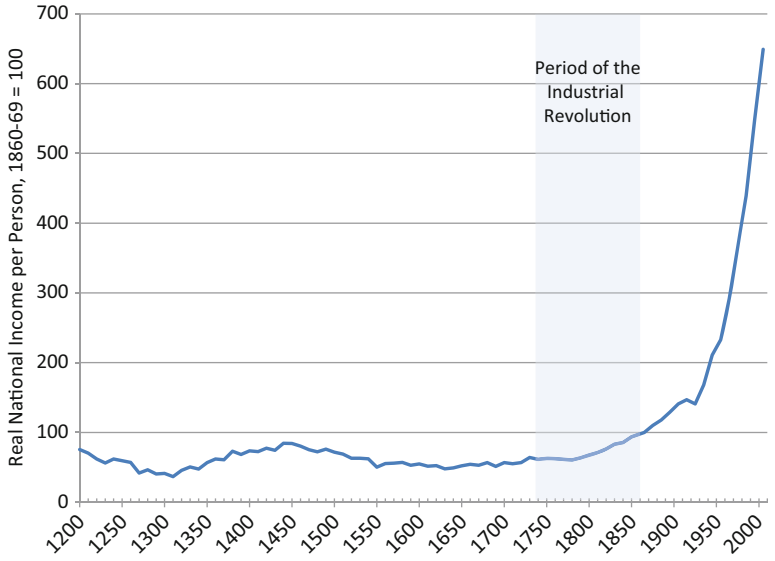


Fig. 3 Real national income per person, England: 1200–2000 (1860–69 = 100) (Source and Notes: Clark (2009), Table 28, column labeled Real National Income/N (PNDP) for 1200–1850, and Table 34 for 1860–2000, where the midpoint of the decade given is graphed. The approximate dating of the industrial revolution is from Clark (2007a), Fig. 10.2)

Table 1 Private versus public provision and funding of K-12 schooling

	Provision	
Funding	Public	Private
Public	Public schools (also called common schools, graded or grammar schools, high schools)	Vouchers (U.S., Sweden 21st century) Pauper schools (U.S. 19th century)
Private	Rate-bills (early to mid-19th century U.S.) Tuition bills (early 20th century)	Private schools (any century) Academies (U.S. 19th century)

to investments in education. Education, in turn, induces more technical change. Finally, families are induced to have fewer and more highly educated children than a greater number of lower-educated children, and the crucial demographic transition can eventually set in.

Human Capital, Institutions, and Economic Growth

The ability of nations to foster human capital accumulation depends on the existence of enabling institutions. One set of these enabling institutions is the legal and extralegal rules that define property rights in man. Another set includes a host of related institutions such as the franchise, form of government (due process, rule of law), and religion.

Optimal human capital investment depends on various factors such as the degree to which capital markets are well functioning and the level of certainty in the economy and polity. When political power is unequally held, human capital accumulation is likely to be suboptimal since groups cannot make credible, long-term commitments to the "elites." Even though everyone could be better off, one can get stuck in a bad equilibrium.

If the key to economic success is good institutions, then "why isn't the whole world developed?" as Richard Easterlin aptly questioned in his 1981 Economic History Association presidential address (Easterlin 1981). A compelling answer is provided in a series of papers.[9]

Acemoglu et al. (2002) reveal the origins of growth-dependent institutions in the colonized parts of the world. As Europeans arrived, places that were dense in existing populations and rich in resources were exploited and given "bad" institutions that allowed Europeans to tax and extract rents. The bleaker, poorer places, on the other hand, were given enabling institutions to encourage European migration.

[9]Mark Twain provides a similar answer in *A Connecticut Yankee in King Arthur's Court*.

These institutional differences persisted and produced "*reversals of fortune*." The poorer places, like North America, became richer and the richer places, like the Caribbean, stagnated.

Engerman and Sokoloff (2012) and Sokoloff and Engerman (2000) contain similar logic and underscore the fact that the same European powers that brought bad institutions to some places brought good institutions to others. The British settled much of North America, but they also settled parts of the Caribbean. Engerman and Sokoloff emphasize particular institutions such as those relating to property rights in man, educational institutions, and the franchise.

A spectrum of labor and human capital institutions has existed historically. Starting with the least free, these institutions include slavery, indentured servitude, labor contracts of various types including apprenticeships, and, ultimately, free labor with its associated educational institutions. If free labor is at one end of the spectrum, then slavery is at the other, and indentured servitude and contract labor are somewhere in between.

Slavery is an ancient labor system. Slaves are mentioned in the Bible, and the majority of Athenians were enslaved in some respects. But slavery in the New World was different. It was not a temporary state. Rather, it was in perpetuity. And in the Americas slavery was mainly based on *race*. Whites could be indentured servants and convict laborers, and they could be coerced and duped, but they could not be slaves.

The slave trade from Africa began in the 1500s with the vast majority brought to Brazil and the Caribbean (60%). Just 7% went to North America. Slaves in the Western Hemisphere were mainly used in tropical and warmer areas to produce sugar, rice, tobacco, indigo, and later (and most consequentially) cotton. But they also produced a large amount of the food consumed in the South. Slavery had once existed in many northern states but changed in the 1790s through a series of gradual and then immediate emancipation laws and state constitutions.[10]

After the US and British slave trade closed in 1808, the market for slaves, which had previously existed in various ways, rapidly developed into a market for hires (rentals) and a market for sales (prices). Slavery provides the most extreme form of the market for human capital. Human beings were rented and they were sold. But did this market mean that there was optimal human capital investment in slaves? Did masters have the right incentives to invest in formal schooling and training in various trades such as carpentry, shoemaking, and mechanics?

It would appear that slavery reduced two barriers to optimal human capital investment. The first concerns capital market constraints since most masters would have been wealthier than ordinary laborers. The second is having an employer invest in the general skills of his employee. In this case the employee was bound to the master. But two larger problems arose.

The first concerns the alignment of private incentives. If a master invested in his slave, would he be able to obtain optimal effort, and could the slave escape? In

[10]Fogel (1989) provides a definitive treatment of the subject.

antebellum southern towns and cities and even in the farm areas, slaves were often hired out.[11] Some of the more trusted and skilled slaves hired themselves out, and master and slave had an implicit or explicit contracts regarding how the income would be shared. These agreements, however, were not commonly found possibly because of issues of trust but most likely for another reason. That reason concerned the public sphere.

Reading and writing, it was believed, would provide slaves with a greater ability to communicate with each other and revolt. Around the 1820s all southern states made the teaching of slaves any literacy skills illegal. The second reason, therefore, is that endowing slaves with much human capital became prohibited.

Indentured servitude was another labor market form that existed to solve a capital market problem. Whereas slavery was for life and for all future generations, indentured servitude was for a given period to pay back a loan generally for passage to America but also to care for orphaned children (e.g., Oliver Twist). In the eighteenth century many who came to North America were "indentured servants" (Galenson 1984). Indentures appear to have enhanced capital markets and enabled geographic mobility. They declined as transport costs decreased and as incomes in the sending nations rose, thereby obviating the need for loans.

Producing Human Capital: Education and Training

The Rise of Formal Education and the Role of the State

A fundamental difference between humans and other species is the extensive transmission and preservation of knowledge among humans. This transmission and preservation is what had led to modern economic growth. But the transmission could not have been broad based and could not have reached the "masses" of people if not for institutions called schools.

Knowledge was, and still is, transmitted without a formal and extensive school system. Socrates taught Plato; Plato taught Aristotle; private tutors taught the Confucian classics to hundreds of thousands of Chinese from the Sung to the Qing so they could take part in the "exam system"; apprentices were taught skills by their masters; parents have always taught their children. But only with schools, in which training begins with young children, could the system reach large numbers of ordinary people.

Formal Schooling in Europe and America

The transition to mass primary education in much of Europe began sometime in the late nineteenth century but occurred much earlier in North America. According to

[11]See Goldin (1976) on slavery in US cities from 1820 to 1860.

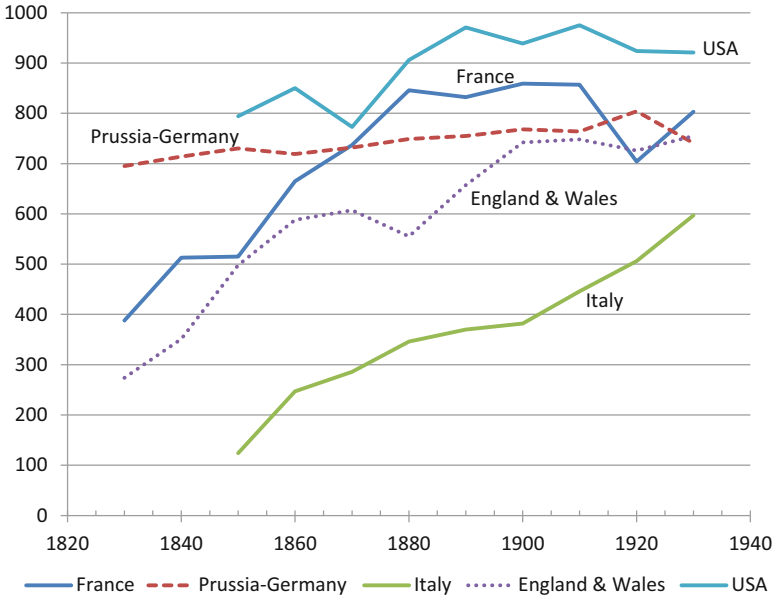


Fig. 4 Public and private primary school students per thousand 5- to 14-year-olds (Sources: All data other than United States, 1850–1870: Lindert (2004a), Table 5.1; Lindert (2004b), Table A1 contains the raw data for the numerator. United States 1850–1870: Carter et al. (2006) Bc 438–446 for enrollment rates and Aa 185–286 to convert 5- to 19-year-olds to 5- to 14-year-olds) (Notes: Data for Prussia after 1910 are extrapolated on Germany's data. Public and private schools are included for Prussia, but public only are reported for Germany. US data is listed in Lindert (2004a) as public and private for 1880–1930 and includes all races; 1850–1870 data are for whites only obtained from Carter et al. (2006) where the denominator is 5- to 19-year-olds and is converted to 5- to 14-year-olds using population data for whites. The 1850–1870 data are from the US Census rather than from administrative records. Lindert cites Carter et al. for 1880–1930, which are from census data and thus could not differentiate public from private enrollments)

the data in Fig. 4, the United States and Prussia, leading nations in education, had primary schooling rates of about 70% by 1860 for 5- to 14-year-olds.[12] The United States retained its lead and surpassed (unified) Germany in the late nineteenth century and had a primary schooling rate of exceeding 90%. But France, Germany, and Britain all had primary school rates in excess of 70% by the start of the twentieth century.

Although the main contours of educational change at the elementary level from around 1840 to 1940 in Europe and the United States are probably well captured in these data, they must be used with some caution. School data are often gleaned from census records and do not account for the number of months children are in school.

[12]The figure for the United States beginning with 1880 includes the South and all races. Thus, the underlying data are even higher for the white population and that outside the US South, which had and still has lower schooling rates than the North and the West.

Especially when school data come from administrative records, it is not always clear that the numerator is for youths between certain ages even though the denominator is. In many places where secondary schools did not yet exist, older youths attended the common or primary schools that held the elementary grades. Therefore the numerator could be inflated by the older children. Another difficulty is that comparisons across nations must account for a variety of institutional details since schools are almost always at least in part in the public sector.

The educational lead of the United States that is apparent in the primary school data for the nineteenth century expanded enormously in the twentieth century with the beginnings of the "high school movement." Although many of the richer nations of Europe had broad-based primary education for youth by the early part of the twentieth century, they did not have mass education at the secondary and tertiary levels. But the United States did.

The United States greatly increased the number of its youth graduating from secondary schools, as can be seen in Fig. 5, so that by the 1950s the median youth in the United States had graduated from a secondary school. In contrast, carefully assembled data comparing 13 OECD nations in the mid-1950s show that secondary school enrollment rates for teenagers in full-time general schools were low in all of Europe. Many of the nations in northern Europe had technical programs for teenaged youth. But even adding these, the enrollment numbers, as seen in Fig. 6, do not demonstrate a broad-based system of secondary education and thus not an open system for tertiary education.

Europe eventually caught up in mass education to the United States and has, in more recent decades, leaped ahead in terms of both the quantity of secondary

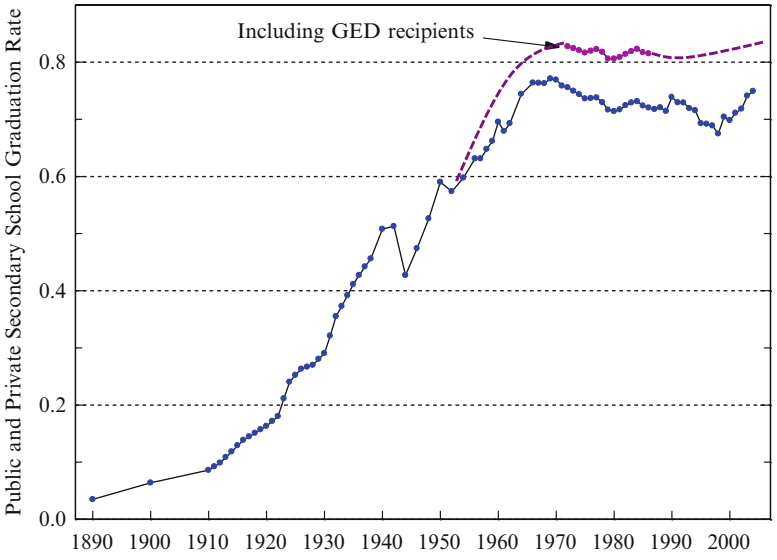


Fig. 5 Public and private secondary school enrollment and graduation rates: United States, 1890–2005 (Source: Goldin and Katz (2008), Fig. 9.2)

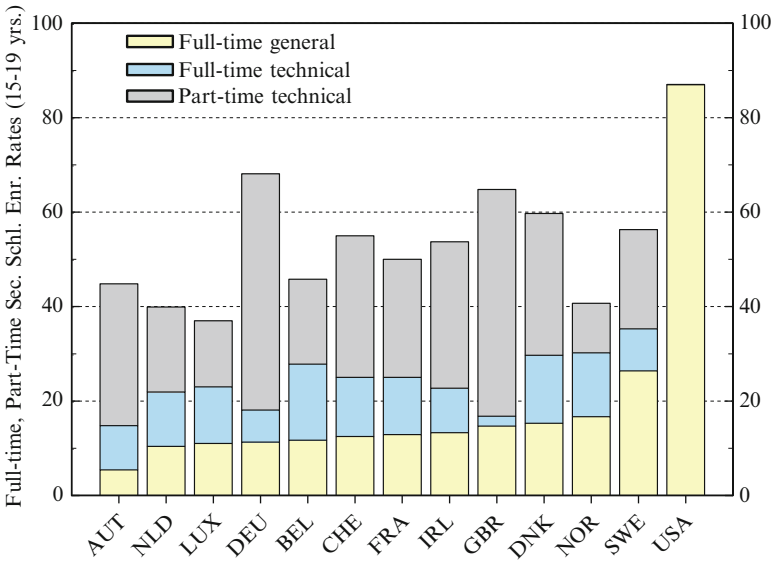


Fig. 6 Secondary school enrollment rates: OECD, 1955/56 (Source: Goldin and Katz (2008), Fig. 1.7)

schooling and its quality. But at this point it is instructive to understand why the United States took the initial lead and what roles were played by individual families and by local, state, and federal governments in terms of funding and regulations.

The twentieth century clearly became the human capital century. It began first in North America but later spread to the rest of the world. How and why did that occur? Mass education in the United States was achieved early because of several characteristics, emphasized in my previous writings (Goldin 2001; Goldin and Katz 2008). These characteristics were "virtuous" at the time and for some time after. Many remained in place, even as some lost their virtuous characteristics.

Education in the United States has generally been open and forgiving in nature. Openness means that schools, by and large, allowed all children to enter. The openness of US schools is related to that fact that ever since the mid-nineteenth century, elementary and secondary schools were (fully) publicly funded by local and state governments. Forgiving means that students who did poorly in one grade were generally allowed to advance to the next. The forgiving nature is related to the fact that until recently there were few standardized tests that were required by law.

The funding of schools, moreover, was provided by small, fiscally independent districts. Because the provision was largely by small districts, rather than the state or the federal government, it was difficult to impose uniform testing. Even more important to the issue of funding is that the districts were so small that there were thousands of them. At their peak in the 1920s, there were around 130,000 school

districts.[13] School districts could compete for families, and families could move to areas that had better schools or less expensive schools or different types of schools.

Another characteristic was that US education was academic, yet it was also practical. Unlike many European nations there were few "tracks" that shunted youth into industrial and vocational programs. All children were to be given the chance to advance to higher grades, even if financial and intellectual limitations often prevented that ideal.

Related to that ideal is that by the early part of the nineteenth century, most primary schools were gender neutral, and during the high school movement, the same was true of the nascent secondary schools. Colleges became gender neutral somewhat later although those in the public sector were generally coeducational from their initial opening (Goldin and Katz 2011a). US education was, as well, secular. Not only did the United States have no established religion (prohibited by the US Constitution Bill of Rights, Amendment 1); state constitutions in the nineteenth century were rewritten to forbid the use of state and municipal funds for religious schools.

These characteristics were virtuous because, in a variety of ways, they increased secondary school enrollment when it was low. The most obvious reason is that openness to most groups meant no exclusions. Small, fiscally independent districts allowed groups of families to determine the amount spent on and taxed for education. By not tracking children at young ages, all children had a chance to rise to the next level. By being forgiving, the errors of one's youth had less impact on one's future.

When levels of education are low, these characteristics are virtuous. Even if a small fraction of the districts want to increase taxes to fund a secondary school, they can do just that and do not have to wait until the majority in a state wants to do so. Families that want to increase public schooling expenditures can migrate to districts that have them, or they can send their children across school district borders and pay tuition.

But these characteristics are not necessarily virtuous in all times. They might increase the quantity of schooling but not necessarily the quality. When enrollment and graduation rates increase, quality not quantity becomes important. Small districts will increase quantity but may also lead to large differences in expenditures per pupil.

Exactly what these characteristics did manage to accomplish in the United States can be seen in Fig. 7 part A. Educational attainment rose by about 1 year per decade, a feat that could happen only with a broad-based educational system. Each generation could look forward to being more educated than its parents and to having children more educated than they.

Interestingly, females had more years of education than males until cohorts born around 1920, and they again did with cohorts born after 1950 when the female lead emerged because of their increase among college students. But after the cohorts born

[13] Although most were "common" school districts, a large fraction was fiscally independent. There are about 16,000 largely independent school districts today. See Goldin and Katz (2008), Chaps. 3 and 4.

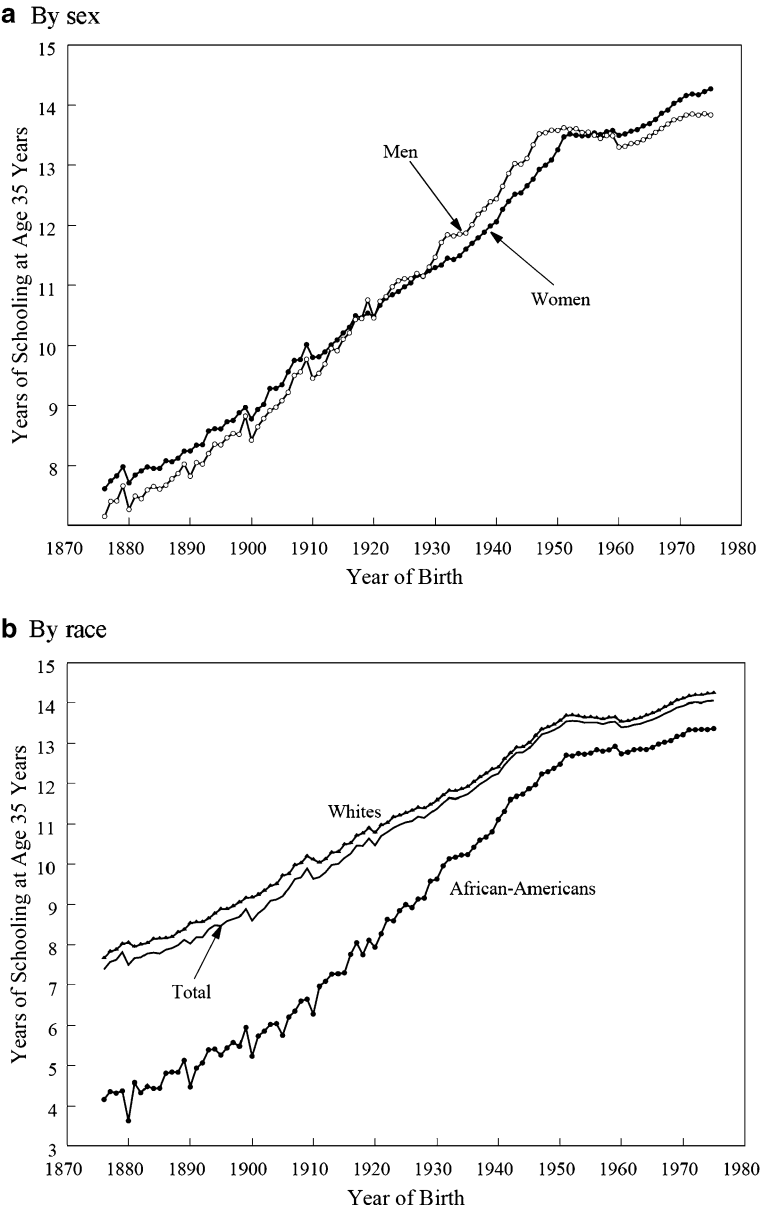


Fig. 7 Years of schooling at age 35 years by year of birth, US cohorts born 1875–1975 (Source: Goldin and Katz (2008), Figs. 1.5 and 1.6)

around 1950, a great slowdown ensued in educational attainment, which has only just begun to pick up again, especially for females.

African-Americans, as can be seen in Fig. 7 part B, had extremely low levels of education early in the twentieth century. Those born around 1880 would have

completed just 4 years of formal schooling, whereas the average white would have completed around 8 years. The quality of schooling for African-Americans was considerably worse than it was for whites, and the number of actual months attended was far less (Card and Krueger 1992a).[14] The levels increased considerably in the twentieth century but did not converge.

Why Invest in Education or Training?

The discussion of educational attainment has not confronted the most basic question in human capital. Why invest at all in education or training? Much can be learned from the simplest of frameworks. Assume a two-period model of human capital investment in which an individual can work or can invest in human capital during the first period. If work is chosen, then w_1 is the first period non-investment wage and w_2 is the second period non-investment wage. But if investment is chosen that costs C, then E_2 is the second period investment wage ($>w_2$). The individual can borrow at rate r. The individual should invest if and only if the following relationship holds,

$$\frac{\left({}^{E_2}/_{w_2}\right) - 1}{1 + r} > \frac{C + w_1}{w_2} \tag{1}$$

which is equivalent to saying that the individual should invest if the discounted returns, expressed as a fraction of the second period non-investment wage, exceed the costs (direct costs of C plus the opportunity cost of the first period non-investment wage, w_1), also expressed as a fraction of the second period non-investment wage.

The simple human capital investment model says that investments are more likely when the returns are higher, the costs are lower (possibly lower with economies of scale provided by schools), and the discount rate (possibly a function of parental income and greater certainty) is lower. But the simple framework does not address several important factors such as where the training takes place (school, on the job, at home), who provides for the training and pays for it, and what role the "state" or collective plays in these matters. These topics are addressed here.

First off, what determines where the training takes place? Does it occur in a formal school or as training on the job or informally at home? It is well known that America in the nineteenth century had far fewer formal apprenticeships than in England. What does this have to do with the investment decision? The answer is that when technological change is rapid and geographic mobility is high, a general, flexible education is more valuable than one that is specific to a particular occupation or place and is relatively inflexible. When the opposite is the case, specific training is

[14]On the quantity and quality of education for African-Americans and whites in US history, see Card and Krueger (1992a). A related article of theirs (Card and Krueger 1992b) shows that the quality of schools, as measured by pupil/teacher ratios, average term length, and teacher salaries, positively affects rates of return to education at the state level.

much better. America had greater geographic mobility and, for some time, greater technological dynamism than in Europe. Both factors made general, flexible education more valuable and occupation-specific apprenticeships and industrial training less valuable.

The second item that is omitted in the simple framework concerns who pays for human capital investment and the role played by the "state." By the "state" I mean a collective of individuals. The collective can be involved in the provision of schooling (e.g., capital investment in the building, hiring of teachers, and selection of curriculum), and it can also be involved in its funding. One can think of the possibilities as a two by two matrix, as in Table 1, where the horizontal headings are the provision (public, private) and the vertical headings are the funding (public, private).

From the nineteenth century to the present, there have been examples of each of the forms contained in the matrix. Cases along the main (positive) diagonal are the most common. They include privately provided schools that are privately funded and publicly provided schools that are publicly funded.

But the minor diagonal elements, which may seem like oddities, exist today and have existed historically. In the nineteenth century, many school districts had schools that were publicly provided but were privately funded. Families received rate bills, also known as tuition bills, for the attendance of their children. The rate bill in some places was for the full amount, but in other districts parents were assessed for the child's attendance only above some maximum number of days.

There are also cases of schools that are privately provided but publicly funded. In some cities in the early nineteenth century, schools were provided for the children of impoverished families by private groups, including religious orders, but were funded by the municipality. In more recent periods, the United States and Sweden, for example, have been using vouchers to fund private schools out of taxes. Therefore, there are a multitude of possibilities that history provides.

Another issue is whether individuals who invest in training have greater ability and, therefore, whether estimates of the return to education and training are biased upward because of selection. The best recent analysis of the magnitude of ability bias shows that it is not very large (Card 1999). For historical estimates of the return to years of schooling given below, because secondary schools in the early twentieth century were just spreading, youths who did not attend them were not necessarily less able than those who did.

Role of the State in Education

In almost all places and during most historical periods, education has been publicly provided and publicly funded. There have been times when the private sector has been larger, but the public sector has almost always increased in relative importance compared with the private sector. The reasons for the increasing government involvement in education are many.

The state has various interests in education that increase demand for schools and, in turn, lead the state to subsidize education. A main interest of the state is that

education provides public goods of various types including endowing citizens with a set of common values. The state also has interests in correcting market failures concerning schooling.

Democracies require literate citizens and educated leaders; nondemocratic governments often restrict education (see, e.g., Sokoloff and Engerman 2000). States have a multiplicity of needs for educated individuals including teachers, engineers, military personnel, clerical staff, and bureaucrats. Education creates positive externalities of many types, such as lower crime rates and better health. In places with low population density, schools are often natural monopolies, and state provision or regulation can be justified on efficiency grounds to increase the quantity available to the public and decrease the price.

Another reason for state involvement in education is that parents often face capital market constraints. Some parents may be insufficiently altruistic, and because children cannot write binding contracts with their parents, they cannot borrow against their future human capital. To increase efficiency, the state might want to lower the interest rate faced by parents and children. A customary way of doing this is to have the schools funded by communities as in an "overlapping generations" framework. Young families with children are subsidized by older families whose children have grown.

If parents are insufficiently altruistic, the state might want to compel them to send their children to schools. If children are too myopic, the state might want to compel them to attend school. Compulsory schooling and child labor laws often accomplish these goals. But these laws were not often binding in the United States (Goldin and Katz 2011b). The reason is that the United States was already providing schools for the masses. In consequence, few families and youths were constrained by the compulsory education and labor laws. In contrast, these laws were often binding in other countries, such as Britain and Ireland. The main reason is that they "compelled" the state to provide schools and pay for teachers for broad-based education.

Most of the reasons just provided for government interference in the production of educated individuals need not involve the provision of schools but would involve the financing of education. The involvement of government in the provision of schools and in the hiring of teachers is often because it is more convenient for the collective to provide these than for a private organization to do the same.

I had previously noted that the United States had an enormous number of independent school districts and that more districts allow the sorting of families by the demand for education and the ability to pay. In many other countries school districts are far fewer in number relative to population and centralization is more the norm. In France, for example, there is just one school district.

Because education is not a pure public good and can be purchased from the private sector, parents can opt out of the public system even though they are still taxed to pay for it. If the parents in a school district have a sufficiently wide distribution of demands for the quantity and quality of education, the public sector can be stymied by what is known as the "ends against the middle problem" (Epple and Romano 1996). Parents with low demand for education (or low income) will not

want to vote for high spending, and parents with high demand for education (or high income) will also not want to vote for high spending since they will, most likely, opt out and use the private sector. One solution is to have small districts that will better match school expenditures to parental demands and result in higher levels of schooling. One of the initial virtues of US education was the existence of a large number of small, fiscally independent districts.

Why Education Levels Increased

Education in the United States, and in most other nations, advanced to mass education across the three transformations that are the three parts of schooling: primary, secondary, and tertiary.[15] The precise number of years of each of these portions and the ages at which youth make each of the transitions varies somewhat across nations. But there is considerable uniformity probably relating to the biology of child development.

In the United States the first transformation to mass primary schools occurred before the twentieth century. Schools for primary school students were often called "common" schools, in part because youth had a shared experience but also because they were generally one-room school houses. In their graded form, they were called grammar schools. The schools were common in the sense of including everyone and common in the sense of being ordinary and abundant. They were operated, most often, by small communities and in their early days were funded by parents through "rate bills," which were charges based on the number of days children attended school. In the mid-nineteenth century, various social movements led to the ending of the rate bills, and by the 1870s primary schools were free to parents and their children. The same laws and judicial interpretation that made primary schools free of marginal charges to parents also made secondary schools free of tuition charges for those living in the school district.

In the late nineteenth century, another educational movement emerged, particularly in the eastern states. It led first to the establishment of private academies of various types that trained youth beyond the limited courses of the common schools. The private academies were generally small, ephemeral institutions, and not much is known about them. Some were later converted into the public high school after the community voted to fund one. The fact that academies were private institutions and almost always funded by individual parents demonstrates that the movement was grass roots in origin. The academy movement morphed into the "high school movement," one that is better known and was national in scope.

Both the academy and high school movements were spurred by the increased demand for skills so that young people could be better prepared to enter the burgeoning world of business and commerce and the more mechanized, electrified

[15]Many places have added two other transitions: preschool to kindergarten and middle school or junior high school to high school.

world of industry. But why did the high school movement begin and expand when it did, around 1910?

One way to assess this question is to look at the earnings of young people with skills valuable to commercial establishment relative to those without these skills. In the pre-1920s, these ratios were exceptionally high, pointing to high rates of return to secondary education just as the high school movement was spreading.[16] The evidence on whether these rates of return were also high in the nineteenth century is less clear (see Goldin and Katz 2008, Chap. 4).

The first national US Census to ask years of completed schooling was in 1940. A few state censuses contained questions on education, and the best of these was done in Iowa. The Iowa state census of 1915 contains rich information on education and earnings for the precise period when greater education was exceptionally valuable in the factory, the counting house, and even the farm.

A set of individual-level earning functions, as provided in Table 2, reveals that years of high school greatly mattered at the start of the high school movement. The pecuniary return to each year of high school, for 18- to 34 year-olds, was around 12%. The return to more education was experienced even within blue collar and farming occupations and was not just because of a shift in the educated population to the white collar sector.

In places that did not have public secondary schools, some youths remained in the common schools for more years. But additional years in the common schools was far less valuable than years in an actual secondary school that could provide instruction in a host of separate disciplines and that could endow youths with various skills.[17] Technological changes were occurring in many of the economy's sectors, and education was a complement to it in 1915, as it is today.

The virtues of education discussed earlier also impacted the spread of higher education in the United States. US higher education was academic yet practical. The enormous number of higher education institutions in the United States produced enormous variety and competition among schools for students and faculty. In 1900, England had just one-seventeenth the numbers of higher education institutions per capita. And even in 1950, England had one-eighth the numbers per capita that existed in the United States.

US higher education was relatively open and forgiving, just as was the case for the lower grades. Students who did not do well enough in high school to enter a university could go to a community college and then transfer to a better institution. The institutions of higher education were geographically close to the people, enabling even rural families to send their children to college. The outcome was that sometime in the twentieth century American colleges and universities became the finest in the world.[18]

[16]Goldin and Katz (2008), Chaps. 4 and 5

[17]This result is given in Goldin and Katz (2008), Table 2.5.

[18]For a discussion of higher education in the United States, see Goldin and Katz (1999, 2008).

Table 2 Returns to a year of education by type of schooling, occupational grouping, age, and sex, Iowa 1915

	18–34 years old					
	Males					Females[a]
Years in school	All occupations	Nonfarm	Farm	Blue collar	White collar	All occupations
Common school	0.0483 (0.00395)	0.0375 (0.00442)	0.0637 (0.00837)	0.0229 (0.00450)	0.0438 (0.00889)	0.00714 (0.00877)
Grammar school	0.0693 (0.00421)	0.0671 (0.00443)	0.0568 (0.0110)	0.0634 (0.00458)	0.0679 (0.00909)	0.0454 (0.00913)
High school	0.120 (0.00564)	0.114 (0.00516)	0.132 (0.0176)	0.0908 (0.00738)	0.0826 (0.00747)	0.101 (0.00760)
College	0.146 (0.00915)	0.143 (0.00799)	0.166 (0.0381)	0.0575 (0.0195)	0.131 (0.00849)	0.151 (0.0122)
Business school, dummy	0.284 (0.0988)	0.273 (0.0831)		0.452 (0.180)	0.0825 (0.0886)	0.508 (0.0969)
R^2	0.251	0.296	0.241	0.256	0.313	0.273
Number of observations	7,145	5,249	1,784	4,021	1,744	2,001

Source: Goldin and Katz (2008), Table 2.1

Notes: Regressions also contain a quartic in potential experience, a race dummy, and a dummy variable for those missing "years in the United States." Potential experience is defined as min (age – 15, age – years of schooling – 7). Blue collar includes craft, operative, service, and laborer occupations. White collar includes professional, semiprofessional, managerial (but not farming), clerical, and sales occupations. Standard errors are given in parentheses below the coefficients

[a]Includes only unmarried women

Race Between Education and Technology

The rate of return to secondary school was high in the period just preceding the high school movement. As secondary school enrollment and graduation increased, the high school premium, meaning the return to graduating high school relative to eighth grade, plummeted.

Because the high school movement shifted some individuals into the college ranks and since there was some substitution of the skills of high school graduates for those of the college educated, the premium to college relative to high school also fell. By the 1950s the wage distribution was far more compressed than it was in the 1910s and 1920s (Goldin and Margo 1992).

But the relative demand for skilled and educated workers continued to advance. The college premium rose in the 1970s, and it has continued to increase. The premium to a year of education today, as seen in Fig. 8, is even somewhat higher than it was in 1910 at the dawn of the high school movement.

The point is that ever since the late nineteenth century at the latest, there has been a race between education, on the one hand, and technology, on the other. That is, there is a race between the supply of skills and the demand for skills with the return

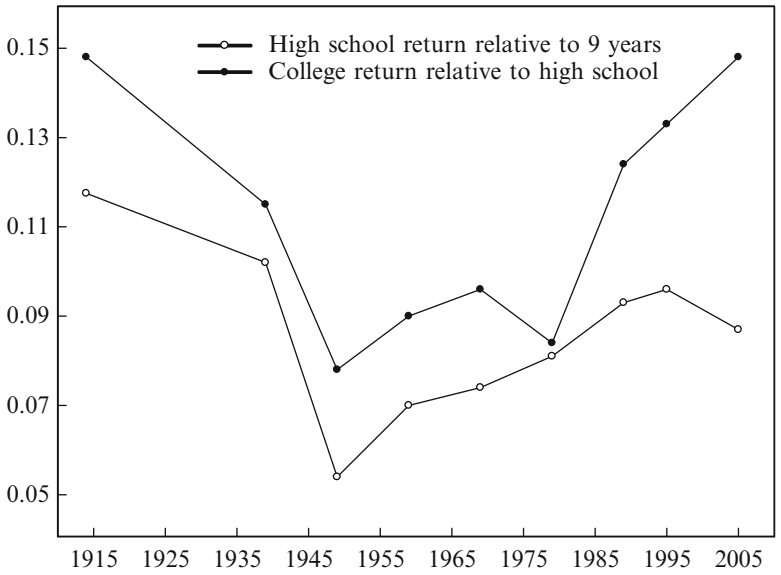


Fig. 8 Returns to a year of school for young men: 1914–2005 (Source: Goldin and Katz (2008), Fig. 2.9 and Table 2.7 for notes)

to education as the equilibrating price. When the return is high, the supply of new skills will be greater, and when it is low, the supply of new skills will be smaller.

New technologies increase the demand for superior skills. The technologies of the late nineteenth and early twentieth centuries increased the demand for workers who could read blueprints, knew a bit about electricity, and were numerate and sufficiently literate to type from scribbled notes and hastily dictated letters. Technological advances throughout the last century increased demands for yet more human capital.

The large increase in the rate of return to education and training in the United States during the last several decades occurred largely because the supply of human capital did not increase sufficiently not because the demand for skills accelerated (Goldin and Katz 2008). But the supply of human capital has recently begun to increase again.

Human Capital and Education: Concluding Remarks

Human capital, in the form of schooling embodied in the labor force, increased in the United States from the beginnings of the nation. It greatly changed in content as the demands for skills in the economy shifted. The increase in years of schooling from the nineteenth century was fairly continuous until the past three decades when it slowed down. The increase followed the three transformations and was often a grassroots movement with the cooperation of communities, states, and, at times,

the federal government. Compulsion had little effect in the United States but had a greater impact in other nations where it often constrained governments to build and maintain schools.

The several virtues of education discussed previously aided the spread of human capital in terms of years of education. But, in recent decades, these characteristics may have slowed progress particularly in terms of the quality of education. Publicly funded education by small, fiscally independent districts increased years of education but produced large differences in per student resources. An open and forgiving system helped spread education to the masses, but such a system often has few promotion and graduation standards at even the state level. Many of these defects of the initial virtues are currently being reassessed by states and by the federal government.[19]

Producing Human Capital: Health

Health Human Capital and Income

In 1650 Thomas Hobbes famously wrote in the *Leviathan* that life was "[solitary], nasty, brutish, and short." He meant that without strong government, civil society would disintegrate into war of every man against every man. But in 1650 life *was* "nasty, brutish, and short," with or without strong government. It was filled with infectious disease and pestilential maladies. And people really were "short." They were 5 in. shorter in Great Britain and France than today and 7 in. shorter in Denmark than currently.

People eventually became healthier and taller. They live a lot longer now and have less nasty lives with less pain and suffering. People now die mainly of chronic diseases, far less from infectious maladies. During the period from the 1600s to the present, the human body changed in a multitude of ways and in a time frame that defies the usual rules of Darwinian evolution.

Increased resources allow people to invest more in their health human capital. But, in addition, more health human capital allows people to be more productive. In the discussion that follows, the causation will mainly go from increased resources to advances in health human capital. There is also an important historical literature in which the causation goes from improvements in health to increases in income.

Improvements to health for most of history are the result of increased resources, not the cause. More resources allow people to consume more calories and protein and to eat more nutritious foods. Investments in improved nutrition enhance health human capital.

[19]For example, state equalization plans have restricted the degree to which separate districts can raise funds, and states have transferred resources to poorer districts. States have passed more stringent high school graduation standards, and "No Child Left Behind," passed in 2002, has forced states to have higher standards at all grades.

For the more recent historical period, however, health improvements have served to increase income. The channel is generally through improvements in health for the young that enable children to attend school for more days and to learn more. Bleakley (2007) shows the effect for hookworm eradication in the US South in the early twentieth century. Almond (2006) investigates the long-term consequences of the 1918 influenza epidemic for those in utero at the time. Health improvements also allow adults to work more days and years over their lifetime and to labor more intensively. The direction of causality here is from an exogenous improvement in health human capital to income.[20]

Measures of Health Human Capital

Mortality is the clearest indicator of health status and one that exists across long periods and for many places. A large number of related measures of health exist historically. Heights and weights for adults and for children during the growth spurt, infant weights, body mass index (BMI), and chronic and infectious disease rates also exist historically. Quality of life measures generally do not.

Expectation of life at birth is given in Fig. 9 for four European nations from the eighteenth century to the present and in Fig. 10 for white males and white females

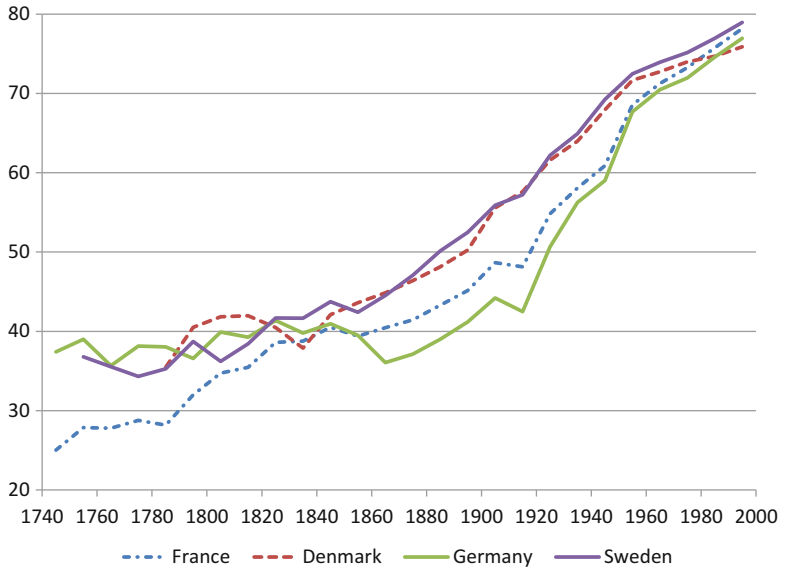


Fig. 9 Expectation of life at birth (period rates) for four European nations, 1745–1995 (Source: Floud et al. (2011), Table 5.1)

[20]See Weil (2007) for a clever way to separate the effects of health on income from the reverse causality.

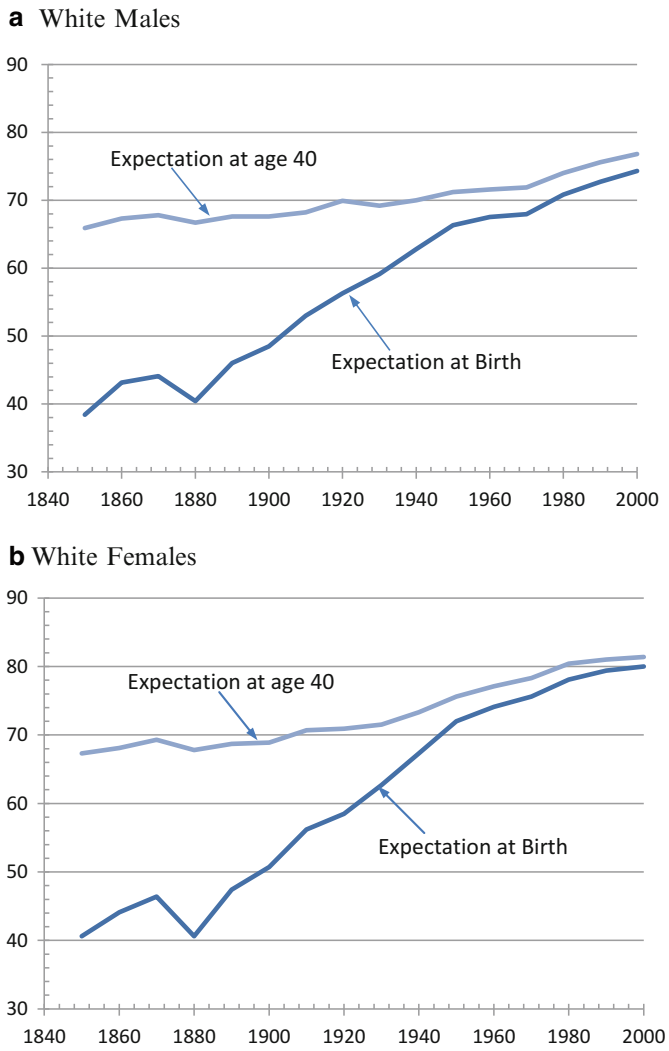


Fig. 10 Expectation of life at birth and at age 40 (period rates) for US whites by sex: 1850–2000 (Source: Carter et al. (2006), Tables Ab644-655)

separately for the United States from the mid-nineteenth century. Life spans at birth were not very long in much of Europe until the mid-twentieth century. The average citizen of France in the late eighteenth century had a life expectation that was less than 30 years at birth, and an individual in Sweden or Germany could expect a lifetime at birth of less than 40 years. Even by 1900 a German or French baby could expect to live to just 40 years or so and one in Sweden, Demark, or the United States to 50 years. But by 1980 all life expectations at birth converged to around 75–80 years old.

Much of the rise in life expectation to the mid-twentieth century in these nations was due to the decrease in infant and child mortality since life expectation conditional on reaching adulthood does not change much until fairly modern periods. That fact can be seen in Fig. 10 for white males and females in the United States. The largest decrease in infant and child mortality occurred from around 1880 to 1920s, although decreases continued.

Life expectation conditional on reaching age 40 changes little until the early twentieth century when it slowly begins to increase. But from the mid-twentieth century onward, the increase in life expectation was primarily from decreased mortality conditional on reaching adulthood. Expectation of life at age 40 increases, and the distance between it and life expectation at birth changes far less than it had up to that point.

Americans lived longer relative to those in other rich countries to mid-nineteenth century. Relative to the English, life expectation at birth before 1850 was better in the United States. But post-1850, life expectations for the two populations were about the same. Relative to France, the expectation of life at birth was better in the United States to around 1900. Americans were abundantly well nourished at least from the start of the nation and were the tallest people in the world to the mid- to late twentieth century (Floud et al. 2011).

In the United States from 1800 to 2000, there was a gain of about 35 years, 40–74 years for men and 44–80 for women. In England from 1750 to 2000, there was a gain of 38 years from 35 to 77 years. In France from 1750 to 2000, the gain in life expectation was 43 years, 25–78 years. The increase in life expectation in each of the countries can be divided into three phases. The first concerns improvements in nutrition, the second involves improvements in public health interventions, and the third phase encompasses a host of medical discoveries such as antibiotics.

Increased Life Expectation: The Three Historical Phases

Phase I: Improvements in Nutrition

Phase I, described by Fogel (2004) as the escape from hunger and malnutrition in Europe, occurred from 1700 to the late nineteenth century.[21] Fogel and his coauthors have emphasized that increased income produced better nutrition and that better health, as children and as adults, allowed the population to fight off infectious disease (Fogel 1997, 2004; Floud et al. 2011).

The notion that health status improved around 1700 because of a marked decrease in chronic malnutrition goes back to Thomas McKeown, a medical historian who wrote *The Modern Rise of Population* (1976). McKeown's goal was to eliminate from consideration two competing factors – public health and medical treatments.

Fogel extended McKeown and gave his ideas considerable force. Fogel noted that before 1700 chronic malnutrition, not crisis-year famine, was an ever-present

[21]The division among the three phases is the author's, not necessarily that of the various contributors to the literature.

problem that limited the health of the population. Sometime around 1700 the second agricultural revolution with its enclosures, plow, seed drill, threshing machine, crop rotation, and selective breeding, brought about a marked increase in caloric intake. In England, for example, calories per capita increased by 300 and in France by a whopping 1,000 in the century after 1750.

Nutrition not only allowed populations to be healthier. More calories also led, over generations, to changes in the human body. Greater food consumption first brought about heavier adults and then produced taller people. The upshot was a higher BMI, healthier individuals, and decreased mortality.

The interpretations offered by Fogel and McKeown have been criticized by Preston (1975, 1996) who notes that the disease environment worsened when urban populations polluted their drinking water, were more distant from food sources, and lived in packed quarters. Some of the overall gain in health status that would have been achieved from increased resources was clearly eaten away by increased population. But much remained that led to an increase in weights, heights, and life span.

Phase II: Public Health Interventions

The next period, Phase II, occurred from the late nineteenth century to the 1930s and was characterized by public health campaigns and interventions. The era could only have begun in the late nineteenth century because of the necessity for the "how" of disease to be discovered and for scientific discoveries concerning the germ theory of disease to be widely accepted.

Little could be accomplished before the understanding of the germ theory of disease. And even after the mechanism for infectious disease was known, water filtration, chlorination, proper sewage disposal, vaccination, quarantine, and food quality regulations had to await public measures and expenditures. Thus, greater public acceptance of the channel through which disease spread was essential. Without that municipalities could not have gained the support to spend large sums on projects to provide clean water and to separate sewage from drinking water.

The mode of disease transmission began to be discovered around 1850 by Semmelweiss who observed that puerperal fever decreased when physicians washed their hands using chlorinated lime. The precise causal agents were not known until around 1870s, first with anthrax, then typhoid and tuberculosis. Robert Koch's work on anthrax in the 1870s proved the germ theory. The foundations had been set by Leeuwenhoek (1600s) who first saw the germs, Pasteur (1860s) who discovered the bacterial basis of decay in foods and later the cause of disease in living organisms, and Lister (1860s, 1870s) who used carbolic acid as an antiseptic in surgery. The understanding of the causal agents was later advanced by Paul Ehrlich known for using salvarsan chemotherapy, also known as the "Magic Bullet," to treat syphilis.

In the case of the United States, Cutler and Miller (2005) have demonstrated the impact of cleaner water on the decrease in infectious disease, particularly typhoid. They estimate the "treatment" effect of filtration and chlorination for 13 cities using plausibly exogenous variation. According to their estimates, water filtration and chlorine treatment account for half of reduced urban mortality in the period.

Some cities (e.g., Philadelphia, Pittsburgh) experienced large effects, but the impacts were small or nonexistent in other cities. Across all cities, filtration and chlorination reduced typhoid fever mortality by 25%, total mortality by 13%, infant mortality by 46%, and child mortality by 50%. Since total mortality declined by 30%, clean water accounted for 43% of the total, 74% of the reduction in infant mortality, and the complete elimination of typhoid (all from 1900 to 1936). The rates of return to investments in clean water technologies were huge.

Life expectation in the United States during the period of public health interventions increased from 45 to 62 years or by 50% of the total change experienced from 1850 to 2000. No other period is as great. Most of the decrease in the period came from the reduction in infectious disease as a cause of death. The period also saw the elimination of the "urban health penalty."[22]

Phase III: The Age of Modern Medicine

The third phase began with the introduction of sulfa drugs in 1935. It was preceded by other medical advances such as the small pox vaccine and salvarsan, an arsenic compound to treat syphilis. The first antibiotics, penicillin in 1941 and streptomycin in 1944, were followed by a multitude of broad-spectrum drugs and antivirals.

Jayachandran et al. (2010) show that from 1937 to 1943, before the discovery and diffusion of penicillin, substantial decreases occurred in deaths from infectious diseases such as scarlet fever, pneumonia, and flu. Maternal mortality, in particular, decreased considerably. The reason for the decline in particular infectious diseases was because of the discovery of sulfa drugs, the precursor to antibiotics.

Infectious diseases were responsible in 1900 for 30% of all deaths in the United States but just 17.5% in 1936 and only 4% in 2000.[23] A combination of public health measures and modern medicines has all but eliminated infectious disease as a cause of death. Not only have life spans been lengthened, a host of modern medical procedures and medications have improved the quality of the years remaining.

In sum, the majority of the gains in longevity in the United States and elsewhere in the rich world came about before the spread of modern medicine. But modern medicine is probably responsible for most of the increase from 65 to 75 or 80 years in the expected age at death from 1936 to 2000 for US men and women. And because of modern medicines and treatments, chronic disease no longer incapacitates large numbers of individuals in their older years.

Human Capital: Summary

Human capital is the stock of productive skills, talents, health, and expertise of the labor force, just as physical capital is the stock of plant, equipment, machines, and tools. Within each type of capital, the performance, vintage, and efficiency can vary. The stocks of human and physical capital are produced through a set of investment

[22]On the changing relationship between health and economic development, see Preston (1975).

[23]See Cutler et al. (2006), Fig. 3 for US data and Floud et al. (2011), Fig. 4.5 for England and Wales.

decisions, where the investment is costly in terms of direct costs and, for human capital investment, in terms of the opportunity cost of the individual's time.

In this essay I have explored human capital in terms of its use and production. Human capital (E) enters the aggregate production function given by Eq. 2 by augmenting labor, which is a function of the level of population (P) and the aggregate labor force participation rate (λ). In practice, human capital is measured as an index of efficiency units of labor. Aggregate output (Q) is altered as well by other inputs such as the stock of capital (K), resources (X), and the level of technology (A).

$$Q = f(A, [E \cdot P\lambda], K, X) \tag{2}$$

This methodology was employed to understand how human capital affects income levels and economic growth. I mentioned that individual well-being could also be impacted in ways that do not necessarily get reflected in aggregate output. Quality of life measures, as they are called, have become an important research area but are difficult to produce historically.

This essay has discussed how human capital is augmented and the rules that are often employed in making human capital investment decisions. Two main types of human capital have been considered here – education and training and health. Both are produced in schools, families, firms, and a variety of other facilities. Both types of investments require good information. Knowledge regarding the cause of disease was important in making investments in health human capital, particularly expensive ones determined by governments, such as water purification. Information regarding the effectiveness of education is required for public investments in schools.

This essay has not emphasized the forces that alter population growth and labor force participation, both of which are related to aggregate measures of human capital. These subjects are covered in other essays, as is the effective use of human capital that can be hampered by discrimination and insufficient geographic mobility.

I have stressed that the subject of human capital is inherently historical. There is much that remains to be explored historically. Why do governments expand formal schooling, and why is informal training more important in certain places and during certain periods? What has been the interplay between grassroots demands for schooling and top-down provision of education? What is the interaction between education and health? Currently, more educated people are healthier. But has that always been the case? The history of schooling across the globe, particularly outside Europe and North America, is still in its infancy. The list of questions and topics in the study of human capital and history is long.

References

Acemoglu D, Johnson S, Robinson J et al (2002) Reversal of fortune: geography and institutions in the making of the modern world income distribution. Quart J Econ 117:1231–1294

Allen R (2001) The great divergence in European wages and prices from the middle ages to the first world war. Explorat Econ Hist 38:411–447

Almond D (2006) Is the 1918 influenza pandemic over? Long-term effects of in utero influenza exposure in the post-1940 U.S. Population. J Polit Econ 114:672–712

Barro R, Sala-i-Martin X (2003) Economic growth, 2nd edn. MIT Press, Cambridge, MA
Becker G (1962) Investment in human capital: a theoretical analysis. In: NBER special conference 15, supplement to J Polit Econ 70(5), part 2, pp 9–49
Becker G (1964) Human capital: a theoretical and empirical analysis, with special reference to education. Harvard University Press, Cambridge, MA
Bleakley H (2007) Disease and development: evidence from hookworm eradication in the American South. Quart J Econ 122:73–117
Card D (1999) The causal effect of education on earnings. In: Ashenfelter O, Card D (eds) Handbook of labor economics, vol 3A. Elsevier/North Holland, Amsterdam
Card D, Krueger A (1992a) School quality and black-white relative earnings: a direct assessment. Quart J Econ 107:151–200
Card D, Krueger A (1992b) Does school quality matter? Returns to education and characteristics of public schools in the United States. J Polit Econ 100:1–40
Carter SB, Gartner SS, Haines MR, Olmstead AL, Sutch R, Wright G (2006) Historical statistics of the United States, Millennialth edn. Cambridge University Press, Cambridge
Clark G (2005) The condition of the working-class in England, 1209–2004. J Polit Econ 113:1307–1340
Clark G (2007a) A farewell to alms: a brief economic history of the world. Princeton Press, Princeton
Clark G (2007b) The long march of history: farm wages, population and economic growth, England 1209–1869. Econ Hist Rev 60:97–136
Clark G (2009) The macroeconomic aggregates for England, 1209–2008. University of California, Davis, Economics WP, 09-19
Cutler D, Miller G (2005) The role of public health improvements in health advances: the twentieth-century United States. Demography 42:1–22
Cutler D, Deaton A, Lleras-Muney A et al (2006) The determinants of mortality. J Econ Perspect 20:97–120
Davis LE, Easterlin RA, Parker WN et al (1972) American economic growth: an economist's history of the United States. Harper and Row, New York
Denison EF (1962) The sources of economic growth in the United States and the alternatives before us. Committee for Economic Development, New York
Easterlin R (1981) Why isn't the whole world developed? J Econ Hist 51:1–19
Engerman SL, Sokoloff KL (2012) Economic development in the Americas since 1500: endowments and institutions. Cambridge University Press, Cambridge
Epple D, Romano RE (1996) Ends against the middle: determining public service provision when there are private alternatives. J Public Econ 62:297–325
Fisher I (1897) Senses of 'Capital'. Econ J 7:199–213
Floud R, Fogel RW, Harris B, Hong SC et al (2011) The changing body: health, nutrition, and human development in the Western World since 1700. Cambridge University Press, Cambridge
Fogel RW (1989) Without consent or contract. W.W. Norton, New York
Fogel RW (1997) New findings on secular trends in nutrition and mortality: some implications for population theory. In: Rosenzweig MR, Stark O (eds) Handbook of population and family economics. Elsevier/North Holland, Amsterdam, pp 433–481
Fogel R (2004) The escape from hunger and premature death: 1700–2100: Europe, America, and the third world, Cambridge studies in population, economy and society in past time. Cambridge University Press, Cambridge
Galenson D (1984) The rise and fall of indentured servitude in the Americas: an economic analysis. J Econ Hist 44:1–26
Galor O (2011) Unified growth theory. Princeton University Press, Princeton
Galor O, Weil D (2000) Population, technology, and growth: from the Malthusian regime to the demographic transition. Am Econ Rev 90:806–828
Goldin C (1976) Urban slavery in the American South, 1820 to 1860: a quantitative history. University of Chicago Press, Chicago

Goldin C (2001) The human capital century and American leadership: virtues of the past. J Econ Hist 61:263–291
Goldin C, Katz LF (1999) The shaping of higher education: the formative years in the United States, 1890 to 1940. J Econ Perspect 13:37–62
Goldin C, Katz LF (2008) The race between education and technology. Belknap, Cambridge, MA
Goldin C, Katz LF (2011a) Putting the 'Co' in education: timing, reasons, and consequences of college coeducation from 1835 to the present. J Hum Cap 5:377–417
Goldin C, Katz LF (2011b) Mass education and the state: the role of state compulsion in the high school movement. In: Costa D, Lamoreaux N (eds) Understanding long run economic growth. University of Chicago Press, Chicago, pp 275–311
Goldin C, Margo RA (1992) The great compression: the wage structure in the United States at mid-century. Q J Econ 107:1–34
Jayachandran S, Lleras-Muney A, Smith KV et al (2010) Modern medicine and the twentieth century decline in mortality: new evidence on the impact of sulfa drugs. Am Econ J Appl 2:118–146
Jones CI, Romer P (2010) The new kaldor facts: ideas, institutions, population, and human capital. Am Econ J Macroecon 2:224–245
Kremer M (1993) Population growth and technological change: one million B.C. to 1990. Q J Econ 108:681–716
Lindert P (2004a) Growing public: social spending and economic growth since the eighteenth century. The story, vol 1. Cambridge University Press, Cambridge
Lindert P (2004b) Growing public: social spending and economic growth since the eighteenth century. Further evidence, vol 2. Cambridge University Press, Cambridge
Mankiw G, Romer D, Weil D (1992) A contribution to the empirics of economic growth. Q J Econ 107:407–438
McKeown T (1976) The modern rise of population. Academic, New York
Mincer J (1958) Investment in human capital and personal income distribution. J Polit Econ 66:281–302
Mokyr J (2004) Gifts of athena: historical origins of the knowledge economy. Princeton University Press, Princeton
Preston SH (1975) The changing relation between mortality and level of economic development. Popul Stud 29:231–248
Preston SH (1996) American longevity, past, present, and future, Policy brief no. 7. Center for policy research. Maxwell School, Syracuse University, Syracuse
Schultz TW (1961) Investment in human capital. Am Econ Rev 51:1–17
Smith A (2003; orig. publ. 1776) An inquiry into the nature and causes of the wealth of nations, Book 2. Bantam Classic, New York
Sokoloff KL, Engerman SL (2000) History lessons: institutions, factor endowments, and paths of development in the new world. J Econ Perspect 14:217–232
Solow R (1957) Technical change and the aggregate production function. Rev Econ Statist 39:312–320
Weil D (2007) Accounting for the effect of health on economic growth. Q J Econ 122:1265–1306

Labor Markets

Robert A. Margo

Contents

Abstract

This chapter presents a brief historical overview of labor and labor markets, using the United States as a case study. Topics include the concepts of the labor force and the labor market; sources of information for historical study; basic features of change over time in the size and composition of the labor force, hours worked, occupations, and skills; changes in real wages over time and in the structure of wages; the emergence of a national market for labor; and the evolution of racial differences.

R. A. Margo (✉)
Boston University and National Bureau of Economic Research, Boston, MA, USA
e-mail: margora@bu.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_27

Introduction

This chapter presents an overview of issues in the economic history of labor and labor markets, using the United States as a case study. My overview is highly selective in method and topics. In terms of method, I focus on research in the "cliometric" tradition. Cliometricians are economic historians who use the tools of modern academic economics – formal theoretical models of economic behavior and econometric models used to test and refine the theory – to study long-term economic development. In their use of the theoretical and statistical tools of modern economics, cliometricians are generally like other economists, and their work is judged by the same standards. However, cliometricians differ from other economists in two key respects.

First, a critical component of the cliometric research agenda is the documenting of long-term change. This requires the collection and analysis of primary historical economic data, often from archives and related sources. Second, when cliometricians use the tools of modern economics, it is primarily to contribute to scholarly understanding of an important issue in economic history, not to validate (or disprove) a particular economic theory (although this can also be a goal of the research). To do both properly – that is, the collection of primary historical data and their analysis – requires deep immersion in the historical context. Good cliometrics, in other words, requires good history, not just good economics.

Cliometricians have made fundamental and lasting contributions to scholarly understanding of the evolution of labor and labor markets. My chapter touches on many of these contributions, although it is far from a complete review. Broadly speaking, I focus on topics involving the measurement of aggregate economic quantities – for example, the unemployment rate – and the demand and supply of labor. The chapter begins by first discussing what economic historians mean by the "labor force" and by a "labor market." I then turn my attention to sources of information for historical study and to the basic features of historical change in the labor force – size, composition, the "intensive margin" (e.g., hours worked), occupations, and skills.

I follow the discussion of the labor force with a discussion of the price of labor – namely, the wage. I present terms, sources, and summarize change over time in real wages and in the "structure" of wages – for example, differences by education level. I also discuss how wages differed across regions in the United States historically and that changes in these differences over time speak to the emergence of an integrated, national market for labor. The chapter concludes with a brief road map of suggestions for further research.

Definition of the Labor Force

An organizing principle in modern economics is the aggregate production function:

$$Y = F(L, K, T) \quad (1)$$

In this equation, Y refers to some measure of aggregate output, L is the aggregate labor input, K is the capital, and T is natural resources ("land"). F is the production function or "technology" linking the use of productive factors (L, K, and T) to output. Output is a "flow" variable – that is, measured over some period of time – and the inputs are also flows over the same period.

Changes in Y between two time periods of production reflect changes in the use of inputs or in the technology (or both). Letting d(ln X)/dt represent the rate of change in a variable over time, we can summarize this point quantitatively in the following equation:

$$d(\ln Y)/dt = dA/dt + \alpha_L d(\ln L)/dT + \alpha_K d(\ln K)/dt + \alpha_T d(\ln T)/dt \quad (2)$$

The "α's" in the above equation are output elasticities – the percentage change in Y for a given percentage change in the relevant factor of production, holding other factors constant. For quantitative purposes, it is generally assumed that the sum of the output elasticities (all of which are positive by definition) is one and that, for computational purposes, each elasticity can be identified with its respective factor share.[1]

In terms of the above equation, L represents the total amount of labor supplied in the economy. Here "amount" has two components – the number of people supplying labor (the extensive margin) and how much time is spent working (the intensive margin).

To measure the first of these components, economists define the "labor force" to consist of individuals who are actively contributing their time and skills to the production of national income. If national income is defined broadly to include production in the home, the majority of the adult population would be in the labor force and the concept would not have much analytic usefulness.[2] But if we take a narrower view, that of market production, there have been significant changes over time in the size and composition of the labor force – that is, how many are working and who they are.

The fundamental source for long-run information on the labor force for the United States is the federal census. Historically the census provides the basis for two different definitions of the labor force. The first definition, used prior to 1940, relies on the "gainful worker" concept as a bright line – if a person reports a "gainful

[1]The assumption that the factor shares sum to one is equivalent to assuming that the aggregate production function is constant returns to scale. However, a strong case can be made for increasing returns in the aggregate – so-called endogenous growth; see PM Romer (1986). The output elasticities will equal their respective factor shares if the factor markets are competitive, so that each factor is paid the value of its marginal product.

[2]The distinction is important historically in the case of the United States because historically much production took place within households. Household production, however, declined as transportation costs fell and more economic activity took place within markets.

occupation" to the census, the individual was part of the labor force. Examples of gainful occupations include farmer, carpenter, domestic servant, and clerk.[3]

According to the second definition, in use today, a person is in the labor force depending on their activities during the census week – that is, a particular window of time. If the person has a job at which he/she is working or would be working except for a temporary hiatus (e.g., a vacation) or works for himself/herself (self-employed), the person is in the labor force. If he/she is without such work but is actively looking for it, he/she is unemployed and still considered part of the labor force. Although modern survey methods are sufficiently refined to measure job-seeking activity, the concept is still fuzzy in practice and especially so during economic downturns in which people turn into discouraged workers, convinced that there is no point looking for work because there is no work to be had.

What Is a Labor Market?

In an idealized market, the goods that are exchanged between buyers and sellers are assumed to be homogenous in quality, and one unit is equivalent as another as far as buyers are concerned – that is, the units are perfect substitutes. For better or worse, this off-the-shelf model is often applied to labor markets – that is, a market in which the good in question being exchanged is the quantity of labor services. Holding the supply curve fixed, an increase in demand will drive up the equilibrium wage and quantity of labor services. Conversely, holding constant the demand curve, an increase in the supply of labor will drive down the equilibrium wage (and increase the equilibrium quantity).

Most markets exist in geographic space, and this is certainly true of most (if not all) labor markets. In a typical labor market, buyers and sellers of labor services are located in physical proximity to each other, and the buyer commutes to her place of employment. This notion of commuting to the workplace gives rise to a fundamental geographic construct in the modern United States – the standard metropolitan statistical area, or SMSA. An SMSA is defined as a collection of counties such that the substantial majority of individuals living in the SMSA also work within its boundaries.

Given a set of labor markets that are geographically delineated, it is natural to ask if they operate distinctly from one another or else are linked together. Imagine that there are two such labor markets, A and B, and suppose that, at the moment, the equilibrium wage in A exceeds that in B. If the cost of migrating between A and B exceeds the difference in wages, there will be no tendency for the situation to change. However, in the long run, if the difference in wages is expected to persist, the net benefit of labor to migrate from B to A may be positive. If this occurs, the

[3]It is possible to adjust the figures to make the pre- and post-1940 figures comparable because information was collected at the time using both questions allowing adjustment factors to be computed; see Durand (1948).

supply of labor will increase relative to demand in A, causing the wage there to fall. Economists speak of this process as the integration of geographically separated labor markets and the narrowing of the wage gap as convergence in wages between A and B. Market integration of this sort occurs if the costs of transporting people between A and B were to fall due to technological change.[4] Alternatively, even if people are not free to migrate between A and B for some reason, the wage difference may narrow if the goods produced in both regions are traded between them; this is called "factor price convergence."

The view that a national market for labor eventually emerged in the United States as the knitting together of geographically distinct labor markets is one that is embraced by many economic historians (see, e.g., Rosenbloom 2002). Later in the chapter, I argue that this view has merit, but it must be supplemented by the consideration of the gradual extension of the frontier – in other words, labor markets expanded outward as the country was settled from east to west.

As noted above, the analogy between goods and labor markets is highly useful but presumes that each unit of labor services is a perfect substitute to one another. A more sophisticated approach to labor market equilibrium invokes the notion of hedonic prices (Rosen 1974). In this model, individuals arrive at the labor market with a set of characteristics which are valued by employers – for example, education or skill – and each employer also comes with a distinct set of characteristics. In this setting, there is not a single equilibrium wage but rather an equilibrium wage function that is comprised of a set of equilibrium prices for worker and employer characteristics. This model is highly useful, for example, in describing how wages vary with education or skill or employer characteristics such as plant safety or the likelihood of job termination or layoff.

Labor markets do not exist in a theoretical vacuum but rather in a specific historical and institutional context. Generally, modern economists have in mind a so-called "free" labor market in which individuals are presumed to have the right to sell their labor services to the highest bidder.[5] However, the model per se does not rest upon a particular set of property rights invested in individuals – the right to trade labor services could be allocated to a third party. Such was the case with indentured servitude and slavery.

In the case of indentured servitude, former individuals were willing to give up their freedom for a period of time in exchange for transit across the Atlantic. Indentured servitude made economic sense because transportation costs from Europe to the New World were very large relative to the productivity (in Europe) of potential servants, and there was no means by which servants could finance the

[4]It can also occur if information flows between A and B improve so that workers have more accurate knowledge of labor market conditions.

[5]That said, labor markets exist in a continuum between truly free labor and slave labor. Historically much labor was restricted in ways that limited labor mobility and even today labor markets are not truly free in the economic sense. For example, many workers in the United States today sign so-called "noncompete" clauses which prohibit them from working for a competitor for some period of time if they terminate their employment with their current employer.

journey themselves. Ship captains were middlemen in this market, arranging for transportation in Europe and then selling servant contracts in the New World. The length of indenture varied with the expected productivity of the servant – shorter, if productivity was higher, longer otherwise – and also with location characteristics. Servants who went to the Caribbean had shorter periods of indenture arguably because health conditions were very poor (Galenson 1984; Grubb 1985).

Chattel slavery, such as was practiced in the United States before the Civil War among other New World economies (e.g., Brazil), was very different from indentured servitude. Slaves did not willingly enter into a contract – they were forcibly abducted or were spoils of war and sold on an international market. The New World, including the United States, was an eager recipient of slave labor from Africa. In the case of the United States, the international slave trade was vigorously active until banned by the act of Congress in 1808. Within the United States, however, slaves were traded more or less freely, until the peculiar institution ended with the defeat of the Confederacy in the American Civil War. These were rental and asset markets – that is, markets in which individuals could transact for the use of slave labor for a specified period of time (rental) or for trading slaves as capital goods (asset). Because of this, there is substantial historical information on both asset and rental prices of slaves, which enables historians to quantitatively assess various key features of the workings of the slave economy (Fogel and Engerman 1974; Fogel 1989).[6]

Documenting the American Labor Force

The historical documentation of the American labor force rests fundamentally on the federal census of population. A census of population was mandated by the United States constitution to be taken every 10 years for the purpose of determining representation in Congress, with the first such census occurring in 1790.

The censuses taken during the first half of the nineteenth century contain relatively limited economic information (the 1840 census is an exception), but there are sufficient data such that, with judicious assumptions, reasonable accurate estimates of the labor force can be made. Starting in 1850, additional information was collected, most importantly on occupation. As the economy shifted out of agriculture and experienced occasional bouts of distress – "panics" in nineteenth-century parlance or business cycles today – "unemployed" workers made their appearance, and it became evident that this, too, became an economic outcome worth documenting, beginning with the 1880 census. States also got into the act of collecting information on the labor force. Starting with Massachusetts, state governments created divisions which monitored and, eventually, regulated various features of labor markets, such as maximum hours that children were permitted to work, or

[6]For example, because both asset and rental prices are known, it is possible to estimate the internal rate of return to owning a slave; see Fogel and Engerman (1974).

plant safety. Established in the late nineteenth century, the US Bureau of Labor Statistics (BLS) also monitored and surveyed the labor market, often at the request of Congress. Many of the documents prepared by the US BLS and its state counterparts in the late nineteenth and early twentieth centuries contain vast quantities of information on individual workers or establishment-level data. The technology to process the data did not exist much less the economic theory to interpret any findings, but the agencies still published the information anyway – perhaps with the belief that, in the not-too-distant future, both the theory and technology (i.e., computers) would be available.

Although the collection of economic data on labor increased steadily after 1900, it became obvious in the early years of the Great Depression that the information was neither comprehensive nor especially timely enough to be of use to policymakers. From the labor statistics point of view, the 1940 census marks a watershed moment – the census was the first to collect comprehensive national data on wages, week works, and educational attainment, all mainstays of modern labor market analysis. But this information was still taken too infrequently to be of use for short- or even medium-term policy.

Economic historians have used the available historical data to construct long-run statistics on the size and composition of the American labor force. The pioneering estimates were undertaken by Lebergott (1964). Lebergott's estimates for the nineteenth century have been revised and updated by Weiss (1992, 1999).

Although the census is indispensable for establishing long-run trends, it provides no evidence on short-run movements. To be useful, such information must be timely – quarterly, say – but the costs of taking a full census every 3 months are obviously prohibitive. Enter the Current Population Survey or CPS for short. The CPS was first taken in 1940 and again in 1944; in the late 1940s, it became a monthly survey. Today, government, business, and academic economists rely heavily on the CPS to give current information about earnings, employment, and unemployment. From time to time, the CPS includes additional questions in its survey, and these have proven invaluable in shedding light on specific topics of current interest.

Since the establishment of the CPS, there have been innumerable specialized surveys by government and private agencies aimed at eliciting labor market information. Of these surveys, arguably the most useful – and certainly the most frequently used – are the Panel Study of Income Dynamics (PSID) and the National Longitudinal Surveys (NLS). These surveys track individuals over time for many years and, indeed, across generations.

Most of the data described above is readily available to the general public via the Internet, such as the websites of the United States Census Bureau (2014) and BLS (United States Department of Labor and Labor Statistics 2014a). Some of the most useful data, such as the CPS or the more recent American Community Surveys (ACS), are samples of individual and household level information. A very convenient source for these samples is the IPUMS (Integrated Public Use Microdata Series) project at the University of Minnesota (Minnesota Population Data Center, University of Minnesota (2014)). The IPUMS site is regularly updated when new

samples become available; among the most interesting in recent years are those in which individuals are linked across census years forming a panel (such as 1880–1910).

The literature on methods for analyzing labor force data is vast and far too complex to discuss here. Excellent background information on survey methods, both historical and contemporary, can be found in the BLS *Handbook of Methods* (United States Department of Labor and Labor Statistics 2014b) (http://www.bls.gov/opub/hom/). The publishing firm Elsevier produces an economics handbook series in which distinguished authors survey the literature on topics of interest at a level useful for professional economists and graduate students; currently there are four volumes in their *Handbook of Labor Economics* series (Ashenfelter and Layard 1986a, b; Ashenfelter and Card 1999a, b, c; Ashenfelter and Card 2011a, b). For tables giving long-term time series on a wide array of labor statistics (e.g., the size of the labor force, unemployment, and so on), a very convenient source is the most recent edition of *Historical Statistics of the United States* (Carter et al. 2006).

Size and Composition of the American Labor Force

Table 1, taken from Margo (2015), displays the aggregate labor force and the labor force per capita (labor force divided by population) from 1800 to 2010. In 1800 there were 1.7 million workers in the labor force, or 320 per 1,000 persons – an aggregate labor force participation rate of 32%. By 1900 the labor force had grown by a factor of 17, and the aggregate labor force participation rate was 38%, 6 percentage points higher than in 1800. The labor force continued to grow in the twentieth century. In 2010, the latest year for which census data are available, there were 154 million workers in the American labor force, and the aggregate participation rate was 50%, 18 percentage points higher than in 1800. As the aggregate labor force participation rate increases in the long run, so does per capita income – implying that rising labor force participation has contributed to rising living standards over the past two centuries of American economic growth.

Changes in the aggregate size of the labor force and in per capita terms reflect complex shifts in population composition, as well as fundamental economic and social change driven by technology, economic growth and development, cultural norms, and government regulation. Children were much more likely to be in the labor force in the nineteenth century than in the twentieth century. The decline in child labor reflects a secular rise in the relative demand for educated workers coupled with the fact that investment in education is sensibly "front-loaded" – that is, undertaken by the young – in the life cycle. It also reflects, to a lesser extent, the passage of laws requiring that individuals remain in school until a certain age – compulsory schooling laws – or which restrict the employment of children – child labor laws (Margo and Finegan 1996). Another long-run trend of enormous importance is the rise of "retirement." Retirement refers to the phenomenon of individuals leaving the labor force at older ages, usually permanently. Retirement was uncommon in the nineteenth century but begins to be observed in the late nineteenth century and accelerates in the twentieth century with the advent of private pensions,

Table 1 The labor force in the United States, 1800–2010

	Labor force (in 1000s)	Per 1,000 population, all ages
1800	1,713	323
1810	2,337	323
1820	3,163	328
1830	4,272	332
1840	5,778	338
1850	8,193	353
1860	11,293	359
1870	13,752	345
1880	18,089	361
1890	23,701	376
1900	29,483	387
1910	37,873	411
1920	42,345	399
1930	49,343	401
1940	56,168	425
1950	62,208	411
1960	69,628	388
1970	82,771	405 [604]
1980	106,940	472 [638]
1990	125,840	506 [665]
2000	140,863	501 [671]
2010	153,889	497 [647]
Average annual rate of growth, 1800–2010	2.11%	0.21%
Average annual rate of growth, 1800–1900	2.88%	0.18%
Average annual rate of growth, 1900–2010	1.50%	0.24%

[] Per 1,000 people, civilian noninstitutionalized population, ages 16 and over (Source: see Margo (2015))

Social Security, and Medicare (Ransom and Sutch 1986; Costa 1998). For detailed discussions of these issues, see Margo (2000a) and Goldin (2000).

The shifts in child labor and labor force participation among the elderly tended to reduce the labor force per capita, and yet the ratio of workers to population rose substantially while these trends were occurring. Some of the upward trend can be attributed to immigration; historically, the foreign-born tend to have higher labor force participation than native-born Americans. But the primary trend offsetting decreases in child labor and older workers is the long term, very substantial rise in the labor force participation rate of married women. The long-term increase in participation among married women reflects shifts in the structure of the economy toward sectors in which women were closer substitutes for men; growth in the relative demand for educated labor, coupled with a largely gender-neutral education system; shifts in cultural norms that enabled women to enter occupations that were formerly closed to them, along with associated anti-discrimination legislation; and

improvements in contraceptive technology, which enabled younger women to more readily invest in schooling and other skills that paid off later in the life cycle (Goldin 1990).

The Intensive Margin

The number of people in the labor force is a very imprecise measure of the labor input into the aggregate production. Among the reasons for this imprecision are changes over time in hours worked. This refers to changes in hours among employed persons as well as, potentially, changes in the incidence and duration of unemployment.

It is frequently assumed that each hour worked by an employed worker is a perfect substitute for the other, so that hours in the aggregate is simply the sum across workers. However, this ignores much evidence that reducing hours per day but keeping days of work per week constant may have a different effect on output than holding hours per day constant but reducing days per week. My discussion below ignores such subtleties (see Atack et al. 2003; Sundstrom 2006).

For the nineteenth-century United States, most of what is known about hours worked pertains to the manufacturing sector. In the early 1830s, the average workweek in manufacturing was about 69 h; this declined to about 62 h on the eve of the Civil War. Weekly hours continued to trend downward for the remainder of the century but slowly – in 1900, the typical workweek was 59 h. This decrease occurred not because of fewer days of work per week but rather fewer hours per day, perhaps a fall of about 90 min per day from the early 1830s–1880s. The 1880 census provides detailed information on variation in hours worked; these data reveal substantial differences across industries and geography, as well as substantial seasonality (Atack and Bateman 1992).

It seems likely that the decrease in weekly hours was offset by an increase in annual weeks worked. Over time, an increasing share of manufacturing establishments operated on a full-year rather than part-year basis. The increase in full-year operation has a multitude of causes – improvements in indoor heating and lighting, enabling firms to operate continuously in colder climates; improvements in transportation networks, which lessened the frequency and severity of supply chain interruptions; and a greater use of fixed capital (machinery), which created incentives for more continuous production (for further discussion, see Atack and Bateman 1992 and Atack et al. 2002).

The decline in weekly hours continued after the turn of the twentieth century, falling from around 60 h per week in 1900 to about 50 h per week in 1920. Further decreases continued in the 1920s and, especially in the 1930s, when employers turned to so-called work sharing as an alternative to layoffs. Not surprisingly, weekly hours rose temporarily during World War Two, but resumed a modest decline after, settling eventually on the norm today, just shy of 40 h per week.[7]

[7]For additional analyses of historical data on the length of the work data, see Whaples (1990), Costa (2000), and Vandenbroucke (2009).

The long-run decline in hours is often interpreted using a simple labor-leisure choice model. In this model, individuals choose between time not spent working – leisure – and goods, which are purchased using the income from work. The outcome of this choice is a labor supply curve relating hours of work supplied to the real wage. If leisure is a normal good, that is, if the income elasticity of the demand for leisure is positive, it is possible for the labor supply curve to be "backward bending," that is, hours of labor supplied will be a negative function of the real wage. For workers who are highly attached to the labor force, it is generally believed that the wage elasticity of labor supply is close to zero; in this case, for a given increase in real wages, the income effect on leisure demand(fewer hours supplied) is just offset by the substitution effect – the worker substitutes away from leisure when its price relative to consumption goods increases. Although the model is useful, it abstracts from changes over time in preferences for leisure which may have been important (see Hunnicutt 1980; Maoz 2010).

As noted earlier, today's concept of the labor force includes individuals who are not currently employed but who are seeking work – the unemployed. In the early nineteenth century in which the majority of the labor force was engaged in self-employed agricultural production, the modern notion does not have much meaning. However, as development progressed in the nineteenth century, labor shifted out of agriculture, and workers increasingly became the employees of someone else.

An important feature of the evolution of labor markets has been the development of laws to deal with contractual relationships between employers and employees. Broadly speaking, in the United States, this evolution produced the notion of "employment at will." With few exceptions, employees are free to leave their job with little or no notice to the employer – that is, the employee is free to quit her job – and the employer has no legal recourse to prevent this from occurring. The flip side is that, again subject to restrictions (more on the employer in today's labor market than on the employee), employers can "divorce" their employees by terminating their jobs. The termination can be with the expectation of being recalled, a separation, or it can be permanent with no expectation of recall – a divorce. It is this two-sided freedom that accounts for two of the three ways in which an individual can enter the status of unemployment.

As noted earlier, unemployment was first recorded in the 1880 census but the data are poor in quality. Much better quality data was collected in 1900 and 1910; the individual level responses have been analyzed by Margo (1990) and reveal significant differences from patterns prevailing in the late twentieth century. Specifically, in the early twentieth century, the odds that a worker would enter into unemployment appear much higher than today (here I am speaking in comparison to similar points in the business cycle), but the length of unemployment was shorter. The labor market, in other words, of the early twentieth century seems closer to the proverbial spot market with more churning than was typical in the late twentieth century.

The time series characteristics of aggregate unemployment have been of central importance to macroeconomics because these are thought to reveal crucial features of the impact of policy. A tenet of post-World War Two macroeconomics for a long time was that policy was effective in taming the business cycle. This was allegedly

revealed by changes in the dynamics of aggregate unemployment – in particular, unemployment was supposedly less volatile once activist policy was adopted.

This tenet was challenged in a celebrated debate originating in Romer (1986a). Pre-1940 unemployment rates were constructed in an entirely different manner from postwar rates. In particular, the pre-1940 rates were inferred as residuals from estimates of the labor force and employment. Romer argued that the assumptions in this method tended to overstate true volatility relative to a time series in which unemployment was collected directly (as is the case after 1940 using the survey week method described earlier). When this bias is corrected, there is no clear evidence that unemployment after WW2 was less volatile than before. Since the publication of the original article, Romer (1999) modified her criticism somewhat, allowing for some dampening of volatility in the 1990s. However, the recent financial crisis may have changed this interpretation of the long run – the jury is still out.

While the debate continues over the second moment of the unemployment series (its variance), there has been little debate – again, excepting the recent period of financial crisis – that in the very long run the unemployment rate has shown little tendency to drift upward or downward. This belies, however, some stubborn cross-sectional differences. The most notable of these is a persistent racial gap in unemployment; this gap, along with associated differences in labor force participation overall, is addressed later in the chapter.

Occupations and Skills

The American labor force changed in the long run not just in terms of numbers or the amount of time spent laboring. There have also been vast changes in the type of work performed. A useful, if imperfect way, to capture these changes is to examine the structure of occupations.

In Table 2, I present estimates of the occupation distribution at 50-year intervals between 1850 and 2000. The distributions are derived initially from the IPUMS samples and subsequently adjusted to be as comprehensive as possible. The broad contours of change are adequately revealed by using so-called "one digit" categories as shown in the table. Details on the construction of the estimates can be found in appendix B of Katz and Margo (2013).

In the top half (Panel A) of Table 2, I show the percentages of the labor force in the various occupation categories. In terms of change, the most substantial are the secular decrease in the share in agriculture and the secular increase in the share in white collar. According to Weiss (1999), about three-quarters of the labor force was engaged in agriculture in 1800, so the long-term shift away from farming began quite early in American history.

The shift of labor out of agriculture can be readily explained by a standard two-sector general equilibrium model with specific factors, although for quantitative purposes more complex versions of the model would be required (Lewis 1979). In the standard two-sector model, agricultural output is a function of labor and land, and

Table 2 Occupation and skill distributions, the United States, 1850–2000

Panel A: by occupation				
	1850	1900	1950	2000
White collar	**6.9**%	**17.1**%	**37.5**%	**61.8**%
Professional-technical	2.3	4.3	8.9	23.4
Manager	3.1	5.7	9.0	14.2
Clerical/sales	1.5	7.2	19.6	24.2
Skilled blue collar	**11.6**	**11.0**	**14.0**	**9.8**
Operative/unskilled/service	**28.7**	**36.4**	**36.8**	**27.1**
Agriculture	**52.7**	**35.3**	**11.7**	**1.2**
Operator	23.9	20.0	7.7	0.6
Farm laborer	28.8	15.5	4.1	0.6
Panel B: by skill group				
	1850	1900	1950	2000
High skill (prof/tech/man)	5.4%	10.0%	17.9%	37.6%
Middle skill 2 (clerical/sales/farm operator/craft)	37.1	38.3	41.3	34.6
% Low skill (oper/unsk/serv/farm lab)	57.5	51.1	40.8	27.7

Source: Computed from Tables 4 and 6 of Katz and Margo (2013). See Katz and Margo (2013, Appendix B) for details on the construction of the figures in this table

nonagricultural output, a function of capital and labor. Labor is allocated between the two sectors so as to equate the value of its marginal product. If the demand for agricultural output is price and income inelastic, an increase in total-factor productivity in agriculture would "push" workers off the farm into the city (i.e., nonfarm occupations), whereas an increase in total-factor productivity in nonagriculture would "pull" workers into the nonfarm sector. It is clear that total-factor productivity increased in both agriculture and nonagriculture, so the effect of technical progress was to shift labor away from farming.

An interesting feature of the estimates is that the share of skilled blue-collar labor remained more or less constant over the nineteenth century, while the shares of white collar and of operative/unskilled/service workers increased. The rough stability in the blue-collar share is the outcome of competing forces. On the one hand, the economy experienced its own industrial revolution, leading to the emergence of a growing and highly productive manufacturing sector. A key feature of manufacturing development in the nineteenth-century United States is the growth of the factory system and the concomitant displacement of the artisan shop. Labor historians refer to this process as one of "de-skilling," in which the share of artisans in manufacturing decreased, while the shares of operatives/unskilled and white-collar workers increased – or as Katz and Margo (2013) put it, the occupation distribution in manufacturing "hollowed out." But manufacturing was more intensive in the use of artisans than the economy overall, and, in addition, demand for artisan labor increased because the construction sector expanded. De-skilling in manufacturing, therefore, reduced the demand for artisans, while manufacturing and construction growth overall increased the demand, leaving the overall share more or less constant (see also Chandler 2006).

During the first half of the twentieth century, there was a steady and substantial move out of agriculture, an equally steady and substantial increase in the share of white collar, stability in the unskilled/operative/service share, and a modest rise in the share of blue collar. Since World War Two, the rise in the white-collar share has been inexorable, while the other groups have declined. In the year 2000, slightly more than 1% of the American labor force was engaged in agriculture, a vast decline over the previous two centuries.

A somewhat different take on the same evidence is provided in Panel B, which classifies occupations by broad skill categories – high, middle, and low. Low-skill jobs require relatively little or no training or education; high-skill jobs require substantial (for the time period) human capital investment; middle-skill jobs are in between. The most salient changes in Panel B are the long-term rise in the share of high-skill jobs and corresponding decrease in low-skill jobs. Middle-skill jobs expanded their share from the late nineteenth century through the first half of the twentieth century but have decreased since 1950.

To make sense of these shifts, they need to be combined with shifts in the relative wages by job category. Based on the available wage data, Katz and Margo (2013) argue that the relative demand for high-skill jobs appears to have increased more or less continuously throughout American history. Here the basic idea is the complementarity between new technologies and skills: as technology advances, much of which is embodied in new capital goods, the demand for high-skilled workers increases relative to the other groups. The increase in relative demand can be met, or not, by shifts in relative supply, or what Goldin and Katz (2008) refer to as the "race" between technology and skills. In the nineteenth century, the relative wage data suggest that the demand for high-skill workers grew slightly faster than supply because the relative wage of high-skilled workers was slightly higher at the end of the century than ca. 1820. In the twentieth century, the relative wages of high-skill workers declined over the first half of the twentieth century but rose over the second half. Goldin and Katz (2008) show that these twentieth-century shifts are explained by shifts in relative supply – the relative supply of highly skilled (educated) workers increased faster than demand over the first half of the twentieth century, but the reverse was true over the second half. The rise in the relative wages of highly skilled workers in recent decades is an important component of increasing income inequality in the United States (see Goldin and Katz 2008).

Wages: The Price of Labor

The wage is the price of labor – a payment per some unit (e.g., per hour) for the rental of a person's labor services. This payment could be in money or "in kind" – for example, housing or food.

Economists distinguish between nominal and real wages. Nominal wages are expressed in terms of current monetary values, whereas real wages are adjusted to reflect changes in purchasing power over time. A real wage, in other words, needs to be deflated (divided) by an index of prices. The price index could be an index of producer prices, in which case the real wage is called the "product wage" and is

isomorphic (or dual) to an index of labor productivity. The price deflator could be an index of consumer prices, in which case the real wage measures the extent to which workers over time can command more goods and services for a given quantity of labor services provided to the market.

Real wages increase over time for two primary reasons. First, individuals may have more complementary inputs to work with – more capital per worker, say. Increases in complementary inputs per worker will raise labor productivity and therefore real wages as well. Second, the economy will experience technical progress, which will raise labor productivity even if there are no corresponding increases in complementary inputs per worker. Historically, both factors have been in play more or less continuously over the course of American economic history.

Sources of Information About Wages in American Economic History

At the start of the nineteenth century, the vast majority of workers were self-employed in agriculture and not working for wages. Information on wages for the colonial and early national period, therefore, tends to come from occasional transactions recorded in account books of farmers or craftsmen. As the nineteenth century progressed, more persons worked for wages and more information is available. For the census years 1850–1870, for example, the Federal Census of Social Statistics recorded average daily wages (with and without board) for common labor and carpenters, the weekly wages of female domestics, and the average monthly wages (with board) of farm labor. Extensive wage data survive for the construction and maintenance of the Erie Canal; company records provide evidence in certain industries, such as textiles. For a more extensive discussion of available sources, see Margo (2000b).

By far the most extensive (and comparable) data pertain to civilian employees of the US Army, who were hired by quartermasters at the various army posts throughout the nineteenth century in a wide array of unskilled, artisanal, and white-collar jobs. Margo (2000b) provides a comprehensive analysis of the extant data for the antebellum period for this source, which yields regional and aggregate time series for unskilled labor, artisans, and white-collar workers. After the Civil War, the available information increases sharply as governments at all levels began collecting wage information on a regular basis. In the late nineteenth century, the US Bureau of Labor Statistics became the primary federal source of routine information on wages, which continues to the present day, supplemented since 1940 by wage information collected by the US population censuses and the CPS. Generally speaking, the BLS data is collected from employers, whereas the census (and CPS) data derive from self-reports by individuals.

Long-Run Growth in Real Wages

Standard long-run series of nominal and real wages can be found in Margo (2006). A useful approximation is that, in the aggregate, real wages have increased at a long-

run rate of about 1.5% per year, implying a fourfold rise every century. There has been acceleration in real wage growth comparing the twentieth to the nineteenth century, and volatility – the standard deviation of the real wage – is also lower in the twentieth century.

The growth in the aggregate real wage, however, masks important and sometimes dramatic shifts in the structure of wages. Economists refer to wage structure as the measures of the distribution – for example, its variance or the difference between wages at the 10th and the 90th percentiles – or closely associated differences by level of skill, such as the difference in wages between white-collar and unskilled workers or the difference in wages between high school and college graduates.

Economic historians have worked hard to measure shifts in wage structure over the course of American history, with a fair degree of success at documentation. In the nineteenth century, these shifts appear to be relatively modest, tending to show that over the century the relative wages of white-collar workers increased compared with unskilled labor or artisans. In conjunction with the evidence on occupations discussed earlier, this suggests that the relative demand for white-collar skills grew more quickly in the nineteenth century than did relative supply, although any difference between two trends was fairly modest (Margo 2000b; Katz and Margo 2013).

In the twentieth century, measures of wage structure follow a U-shaped pattern (Goldin and Katz 2008). Specifically, the relative wage of skilled or educated workers appears to have decreased during the first half of the twentieth century but increased during the second half, leaving the level of the skill or education premium approximately the same. The U-shaped pattern was not due to shifts in the relative demand for skills – these appear to have been more or less constant across decades, with the exception of the 1940s (Goldin and Margo 1992). Rather, the shifts in wage structure are due to supply. During the first half of the twentieth century, the supply of skilled, or educated, workers increased relative to demand, whereas in the second half of the century, supply lagged significantly behind demand (Goldin and Katz 2008).

Regional Differences: The Emergence of a National Labor Market in the Nineteenth Century

A major theme in American economic history is the emergence of national markets in goods and mobile factors of production. This process began in the nineteenth century and occurred in conjunction with the settlement of the country from east to west (Rosenbloom 2002). The process was greatly facilitated by the so-called transportation revolution – canals, inland waterways, and, most importantly, railroads (Taylor 1951; Slaughter 1995; Atack et al. 2010).

The evolution of a national market in labor begins with a consideration of a paradox. In 1840, per capita income was highest in the Northeast and lower, on average, in the South than in the North. Within the North, per capita income was much higher in the Northeast than in the Midwest. The direction of population

movement within the North, however, was from east to west – that is, from the region where per capita income was highest to where it was lowest (Easterlin 1960).

A variety of explanations have been offered to explain east-west migration given the regional income gradient. One prominent hypothesis views the west as a "safety valve" for disaffected eastern labor. The idea here is that migration to the western frontier may have been selective – those who left the East were low-wage workers whose wages were in fact higher in the West than they were in the East but, on average, were still lower than average wages in the East. This would happen if migrants to the west were "negatively selected," but the extent of negative selection cannot fully resolve the paradox (Ferrie 1997).

A complementary explanation focuses on the possibility of capital gains to land (Galenson and Pope 1992). Migrants to the frontier could not immediately begin farming – the land had to undergo extensive improvement. Moreover, the value of the land was heavily dependent on its proximity to transportation, which itself was a function of settlement (Craig et al. 1998; Coffman and Gregson 1998; Atack and Margo 2011). Capital gains, nonetheless, did occur, and it is also worth noting that per capita incomes in the Midwest rose substantially relative to the Northeast between 1860 and 1880. The key point is that migration was expected to be permanent rather than transitory, implying that the present discounted value of migration is the relevant gross benefit of moving, not the current difference in income.

Another argument is that, for a variety of reasons, the per capita income estimates do not capture the actual geographic pattern of differences in the marginal product of labor. That is, the marginal product of labor may have been higher in the West than in the East, and yet measured per capita income was actually lower. The simplest models of labor market integration posit that labor should move from A to B if the value of the marginal product of labor is higher in B than in A, allowing for the costs of migration.

If the value of the marginal product of labor was higher on the frontier, this should be evident in real wages. Margo (2000b) provides annual time series of nominal and real wages for the United States from 1820 to 1860 for three occupations, common labor, skilled artisans, and clerks (i.e., white-collar workers), for four census regions – the Northeast, Midwest, South Atlantic, and South Central regions. In addition, he also provides regional estimates of the number of workers in three occupations over the same period.

Margo uses these data to study shifts in relative wages and employment before the Civil War. The first pattern that emerges is that, for all three occupations, real wages within the North and within the South were higher on the frontier – the Midwest in the case of the North and the South Central in the South – than in the settled region, the Northeast and the South Atlantic.

Second, within the North, there was a general tendency for the regional wage gap to decline over time. For example, in the case of common labor, Margo estimates that real wages were about 32% higher in the Midwest than in the Northeast in the 1820s, but the gap had fallen to 17% in the 1850s. Over the same period, the share of common labor in the North residing in the Midwest rose substantially – that is,

relative (Midwest-Northeast) wages moved inversely with relative employment. This suggests a process of market integration in which labor moved from east to west, increasing the relative supply of labor in the west and causing the relative wage to fall.

An analogous process of convergence also occurred for skilled artisans and white-collar workers in the North. Interestingly, the initial gap was much larger for skilled artisans and white-collar workers than for common labor, indicating a skill "shortage" and therefore a relatively high initial skill premium on the frontier. Again, skilled blue- and white-collar labor responded by moving from east to west, causing the wage gap to narrow over time.

Within the South, there is less evidence of regional convergence in wages, regardless of occupation. However, it is also the case that, in absolute terms, the regional gaps were generally smaller than in the North, suggesting the possibility that the southern labor markets before the Civil War may have been more efficient in terms of regional allocation than the northern labor market.

In addition to regional gaps, Margo (2000b) also provides evidence on wage convergence using data from the 1850 and 1860 censuses of social statistics. These censuses recorded the average daily wages of common labor and other occupations, with and without board, at the level of minor civil divisions, a geographic aggregate smaller than a county. It is possible to use these data to construct a proxy for real wages and, therefore, estimate a regression of the change in real wages between 1850 and 1860 on the initial level. Margo finds that the coefficient on the initial level is significantly negative, consistent with a market integration process in which labor migrated generally from low to high real wage areas.

If state dummy variables are added to Margo's regressions, these show that the extent of wage convergence was less complete across states rather than within. This is not surprising because the average distance within states between counties (the unit of observation in the regressions) is shorter than the average distance across states. We expect that distance will matter – less accurate information about job opportunities and wage differences and higher costs of migration.

Both because the shortest distance between two points is a straight line and because much human capital in agriculture in the nineteenth century was latitude-specific, the settlement process in nineteenth-century America was generally due West and, importantly, mostly incremental (Steckel 1983). Occasionally, however, vast amounts of intermediate settlement were sidetracked in favor of direct movement to a very distant location. This occurred because of very large "shocks" to labor demand in the distant locations, invariably in response to the discovery of natural resources.

By far the most famous example of such a discovery in the nineteenth century was the California Gold Rush. Close study of the Gold Rush reveals much of interest to the student of historical labor markets.

The Gold Rush commenced with the discovery of gold in California in January 1848 and was for all practical purposes complete by the middle of the 1850s. Although obviously part of the land mass of North America, it would be incorrect to call California part of the American economy in the early nineteenth

century – if anything, it was part of the Mexican economy. But American fur traders had arrived at the start of the nineteenth century, and Russia also had its eyes of California, establishing a fort in 1812 in the northern part of the region. Slowly, Americans began arriving, agreeing to become Mexican citizens in exchange for land grants. Conflicts between the settlers and the Mexican government escalated into war in 1846 which ended with California (and other lands) being ceded to the United States by treaty in 1848. Ironically, the 1848 treaty was signed shortly after gold was discovered. When the news of the discovery reached the east coast, it set off a frenzy of activity as "49ers" made their way arduously to the gold camps.

Margo (2000b) provides a model to assess the impact of the Gold Rush on the labor market in California at the time. The model is inspired by similar frameworks used to assess "Dutch disease" – so-called in reference to the effects of the discovery of oil on the Netherlands in the 1970s. In the model, there are two sectors, one of which is gold mining. The discovery dramatically increases the demand for labor in gold mining. Some labor responds by shifting out of the other sector, but this is not nearly enough to prevent wages in gold mining from rising. The increase in wages draws in migrants from the rest of the country, prompting wages to decline. If the Rush is fully temporary – that is, over when all the gold is mined – both labor supply and wages should return to their pre-gold equilibrium.

Based on archival wage data, Margo provides estimates of nominal and real wages for artisans, white-collar workers, and common laborers in California from 1847 to 1860. He finds, as the model predicts, a sharp rise in wages after gold is discovered, but the rise is abated and to some extent reversed as in-migration occurs. However, he also finds that real wages in California settle at a level that is significantly higher than before the Rush. Since not all of the labor returned, this suggests that what the discovery really did was speed up the exploration (and exploitation) of the Pacific coast. In fact, California entered the union as a state long before many of the other territories in the West, and to this day it remains far more settled than much of the land between it and the Midwest.

Diversity in the Labor Market: Racial Differences

Race is central to the economic history of the United States – one cannot truly understand American economic development without understanding the role of slavery, nor can one understand the post-Civil War history of the country without appreciation for the role of race. Race, too, plays a key role in the evolution of American labor markets and in the income distribution.

Comprehensive data on income by race for the nation as a whole are not available until after World War Two. For the years after the Civil War but prior to World War Two, race-specific information on earnings (income from wages) is available in the 1940 census. Prior to 1940 there are scattered surveys of wages by race and other information on income sufficient to allow economic historians to piece together a plausible timeline on racial income differences.

The earliest black-white income estimate in the aftermath of the Civil War is that by Robert Higgs (1977) for approximately 1870. The underlying data is more extensive and reliable for agriculture than for nonagriculture, but this is a good thing, because just after the Civil War, the vast majority of African-Americans were in the South, engaged in agriculture. Higgs' estimate of the black-white per capita income ratio in 1870 is 0.25 – for every dollar of income accruing to a white person, blacks received 25 cents. Higgs has also made an estimate for 1900, and here the ratio is 0.35. The implication is that racial convergence occurred in the three decades after the Civil War – the black-to-white income ratio increased.

There are several reasons to believe that the direction of change is plausible even if one quibbles about the magnitudes. In the aftermath of the Civil War, schools were set up for African-Americans in the South, and, for the vast majority, these were the first schools any had ever attended. As a consequence, the racial gap in literacy – a chasm in 1870 – had narrowed substantially by 1900. Literacy had an economic payoff in the postbellum south, and thus the narrowing literacy gap promoted convergence in incomes (Collins and Margo 2006).

A second reason to believe that convergence plausibly occurred is that there is evidence of convergence in wealth. The evidence comes in two forms. The first form refers to assessed wealth for tax purposes, which was reported by a number of southern states. These data show a narrowing of black-white differences in per capita wealth from after the Civil War to about World War I (Higgs 1982; Margo 1984). Unless the wealth to income ratio increased quite significantly for blacks relative to whites over the same period, the narrowing black-white wealth gap would imply a narrowing black-white income gap. Race-specific estimates of homeownership recently compiled by Collins and Margo (2011) also show a narrowing gap between 1870 and 1910, consistent with the wealth data and also with racial income convergence.

What about the twentieth century? Here the pattern is mixed – periods of stability and, on occasion, retrogression interspersed with periods of significant convergence.

Smith (1984; see also Smith and Welch 1989) is a well-known article that provides estimates of black-white income ratios for adult males from 1890 to 1980. Over this period, the income ratio increased from 0.44 to 0.62. Little change occurred, however, between 1890 and 1940; all of the long-run increase happened after World War Two. According to Smith's estimates, the increase between 1940 and 1980 was split evenly between 1940 and 1960, and 1960 and 1980. Smith argues that the primary factors behind the convergence were racial narrowing in educational attainment and African-American migration from the South. Margo (1986, 1990), however, argues that the census data on educational attainment, properly interpreted, do not support Smith's argument and that racial divergence in incomes likely took place in the South before World War Two, impeding convergence at the national level.

Schooling and migration are supply-side factors in racial convergence. Further research suggests that racial convergence took place in two distinct episodes, both of which reflected significant increases in the demand for black labor relative to whites. The first episode was the 1940s. In the 1940s, blacks gained relative to whites

because of shifts in demand that favored less-educated workers, the only such period in the twentieth century, and also because large number of blacks left the rural south in response to wartime shifts in production (Goldin and Margo 1992; Margo 1995; Goldin and Katz 2008). In the second period, shifts in demand associated with the Civil Rights Movement were critical, particularly in the South (Donahue and Heckman 1991).

Since 1980 there has been limited racial convergence in incomes (Neal and Rick 2014). The absence of convergence reflects many factors. On the supply side, black-white differences in skills and education have not narrowed significantly in recent years, and this lack of narrowing is an important reason why labor market differences by race remain large (Neal 2006). The growth of incarceration, which has disproportionately affected African-Americans, may play a role; employers are reluctant to hire ex-convicts (Neal and Rick 2014). Since 1980 there has been a substantial increase in income inequality in the United States, a portion of which can be attributed to rising relative demand for better-educated workers. Because African-Americans continue to lag behind whites in educational attainment, these shifts in demand have impeded racial convergence (Juhn et al. 1991). African-American incomes have also been harmed by globalization trends that reduce the demand for manufacturing labor in the United States and also by the growing share of foreign-born workers, who are closer substitutes for African-American workers than for white workers (Borjas et al. 2010).

Directions for Future Research

This chapter has presented an overview of issues in the economic history of labor and labor markets, using the United States as a case study. The overview is both brief and highly selective in topics and method. I have concentrated on topics associated with the measurement of aggregate quantities and those involving the demand and supply of labor, rather than the institutions of the labor market (slavery is an exception). Aside from being selective in topic, my review is also selective in method – I have focused largely on research in the cliometric tradition.

Although cliometricians have made important contributions to our understanding of the long-term evolution of the American labor force, there is still much to be learned, even about the "bread-and-butter" topics surveyed in my chapter. In keeping with the supply-and-demand architecture of the chapter, I group my suggestions for further research into those pertaining to the quantities (e.g., unemployment or hours worked) versus those pertaining to wages and labor compensation.

Cliometricians such as Stanley Lebergott (1964) and Thomas Weiss (1992, 1999) have developed excellent estimates of the labor force and its components going back well into American history. Broadly speaking, these estimates are on a very solid footing for the census years beginning in 1850 but are less secure for earlier years. Further research on this topic, perhaps using archival records such as diaries documenting labor force activity for specific population groups (children, young women), would be helpful. Even more valuable would be improvements in the

reliability of annual estimates of employment and unemployment prior to World War Two, which would greatly enhance our understanding of economic behavior over the business cycle.

As in other advanced industrialized economies, the shift of labor out of agriculture was the defining feature of long-term economic development in the United States. New investments in human capital across generations figure prominently in this shift. Successive generations of children growing up on the farm realized – or rather, their parents did – that their future was not in agriculture, and to secure this future required learning new skills and, typically, going to school for more years. How this process played out across space and time in the nineteenth-century United States, when the frontier was still expanding westward, is very poorly understood. Census micro-data linked across generations (such samples are available from the IPUMS project at the University of Minnesota mentioned earlier in the chapter) might provide essential insights into this topic.

Although cliometricians have done much in recent years to map out the history of wages in the United States (see Margo 2000; Goldin and Katz 2008), there is still much to learn. In particular, more needs to be done to understand the relationship between wages and human capital, especially the "returns to schooling" – the change in wages associated with an additional year of formal schooling. For the twentieth century, estimates of the returns to schooling can be made at a national level for period after 1940 and for one state, Iowa, in 1915 (see Goldin and Katz 2008). However, for the nineteenth century, all that is known at present is how wages differed by occupation – for example, the difference in wages between carpenters and common laborers. While it may prove impossible to find suitable direct micro-data on schooling and wages for the nineteenth century, further documentation of the differences in schooling across occupations could still be useful in understanding shifts in the demand for labor relative to supply for different skill levels (see Katz and Margo 2013).

References

Ashenfelter O, Card D (1999a) Handbook of labor economics, vol 3a. North-Holland, Amsterdam
Ashenfelter O, Card D (1999b) Handbook of labor economics, vol 3b. North-Holland, Amsterdam
Ashenfelter O, Card D (1999c) Handbook of labor economics, vol 4c. North-Holland, Amsterdam
Ashenfelter O, Card D (2011a) Handbook of labor economics, vol 4a. North-Holland, Amsterdam
Ashenfelter O, Card D (2011b) Handbook of labor economics, vol 4b. North-Holland, Amsterdam
Ashenfelter O, Layard R (1986a) Handbook of labor economics, vol 1. North-Holland, Amsterdam
Ashenfelter O, Layard R (1986b) Handbook of labor economics, vol 2. North-Holland, Amsterdam
Atack J, Bateman F (1992) How long was the workday in 1880? J Econ Hist 52:129–160
Atack J, Margo RA (2011) The impact of access to rail on agricultural improvement: the midwest as a test case. J Trans Land Use 4:5–18
Atack J, Bateman F, Margo RA (2002) Part-year operation in nineteenth century american manufacturing: evidence from the 1870 and 1880 censuses. J Econ Hist 62:792–809
Atack J, Bateman F, Margo RA (2003) Productivity in manufacturing and the length of the working day: evidence from the 1880 census of manufactures. Exp Econ Hist 40:170–194

Atack J, Bateman F, Haines M, Margo RA (2010) Did railroads induce or follow economic growth? Urbanization and population growth in the American midwest, 1840–1860. Soc Sci Hist 34:171–197
Borjas GJ, Grogger J, Hanson GH (2010) Immigration and the economic status of African-American men. Economica 77:255–282
Carter SB et al (2006) Historical statistics of the United States, earliest times to the present, millennial edition vol 2/Part B, work and welfare. Cambridge University Press, New York
Chandler A (2006) How high technology industries transformed work and life worldwide from the 1880s to the 1990s. Capital Soc 1:1–55
Coffman C, Gregson ME (1998) Railroad development and land values. J Real Est Fin Econ 16:191–204
Collins WJ, Margo RA (2006) Historical perspectives on racial differences in schooling in the United States. In: Hanushek E, Welch F (eds) Handbook on the economics of education, vol 1. North-Holland, Amsterdam, pp 107–154
Collins WJ, Margo RA (2011) Race and home ownership from the end of the civil war to the present. Am Econ Rev Pap Pro 2011:355–359
Costa D (1998) The evolution of retirement: an American economic history, 1880–1990. University of Chicago Press, Chicago
Costa D (2000) The wage and the length of the work day: from the 1890s to 1991. J Lab Econ 18:156–81
Craig LA, Palmquist RB, Weiss T (1998) Internal improvements and land values in the antebellum United States. J Real Est Fin Econ 16:173–189
Donahue J, Heckman JJ (1991) Continuous versus episodic change: the impact of civil rights policy on the economic status of blacks. J Econom Lit 29:1603–1643
Durand J (1948) The labor force in the United States, 1890–1960. SSRC, New York
Easterlin R (1960) Interregional differences in per capita income, population, and total income, 1840–1950. In: Committee on research on income and wealth (ed) Trends in the American economy in the nineteenth century. Princeton University Press, Princeton, pp 73–140
Ferrie J (1997) Migration to the frontier in mid-nineteenth century America: a re-examination of turner's safety valve hypothesis. Department of Economics, Northwestern University, Unpublished Working Paper
Fogel RW (1989) Without consent or contract: the rise and fall of American slavery. Norton, New York
Fogel RW, Engerman SL (1974) Time on the cross: the economics of American negro slavery. Little Brown, New York
Galenson DW (1984) White servitude in colonial British America: an economic analysis. Cambridge University Press, New York
Galenson DW, Pope C (1992) Precedence and wealth: evidence from nineteenth century Utah. In: Goldin C, Rockoff H (eds) Strategic factors in nineteenth century American economic history: a volume to honor Robert W. Fogel. University of Chicago Press, Chicago, pp 225–241
Goldin C (1990) Understanding the gender gap: an economic history of American women. Oxford University Press, New York
Goldin C (2000) Labor markets in the twentieth century. In: Engerman S, Galllman R (eds) Cambridge economic history of the United States, vol 3. Cambridge University Press, New York, pp 549–624
Goldin C, Katz LF (2008) The race between education and technology. Harvard University Press, Cambridge
Goldin C, Margo RA (1992) The great compression: the wage structure in the United States at mid-century. Q J Econ 107:1–34
Grubb F (1985) The market for indentured immigrants: evidence on the efficiency of forward-looking contracting in Philadelphia, 1745–1773. J Econom Hist 45:855–868
Higgs R (1977) Competition and coercion: blacks in the American economy, 1865–1914. Cambridge University Press, New York

Higgs R (1982) Accumulation of property by southern blacks before world war one. Am Econ Rev 72:725–737
Hunnicutt B (1980) Historical attitudes toward the increase of free time in the twentieth century: time for work, for leisure, or as unemployment. Soc Leis 3:195–218
Juhn C, Murphy KM, Pierce B (1991) Accounting for the slowdown in black-white wage convergence. In: Kosters MH (ed) Workers and their wages: changing patterns in the United States. Am Enter Inst, Washington, pp 107–143
Katz LF, Margo RA (2013) Technical change and the relative demand for skilled labor: the United States in historical perspective. Working paper 18752, NBER, Cambridge
Lebergott S (1964) Manpower in economic growth: the American record since 1800. McGraw-Hill, New York
Lewis FD (1979) Explaining the shift of labor from agriculture to industry in the United States: 1869 to 1899. J Econom Hist 39:681–698
Maoz YD (2010) Labor hours in the United States and Europe: the role of different leisure preferences. Macro Dyn 14:231–41
Margo RA (1984) Accumulation of property by Southern blacks before world war one: comment and further evidence. Am Econ Rev 74:768–776
Margo RA (1986) Race and human capital: comment. Am Econ Rev 76:1221–1224
Margo RA (1990) Race and schooling in the South, 1880–1950: an economic history. University of Chicago Press, Chicago
Margo RA (1995) Explaining black-white wage convergence, 1940–1950. Ind Labor Relat Rev 48:470–481
Margo RA (2000a) The labor force in the nineteenth century. In: Engerman S, Gallman R (eds) Cambridge economic history of the United States, vol 2. Cambridge University Press, New York, pp 207–243
Margo RA (2000b) Wages and labor markets in the United States, 1820 to 1860. University of Chicago Press, Chicago
Margo RA (2006) Wages and wage inequality. In: Carter S (ed) Historical statistics of the United States, millennial edition, Part B: work and welfare. Cambridge University Press, New York, pp 40–46
Margo RA (2015) The American labor force in historical perspective. In: Cain L, Fishback F, Rhode P (eds) Oxford handbook of American economic history. Oxford University Press, New York
Margo RA, Finegan TA (1996) Compulsory schooling legislation and school attendance in turn-of-the-century America: a 'natural experiments' approach. Econom Lett 53:103–110
Minnesota Population Data Center, University of Minnesota (2014) Integrated public use microdata series. www.ipums.umn.edu. Accessed 2 Sept 2014
Neal D (2006) Why has black-white skill convergence stopped? In: Hanushek E, Welch F (eds) Handbook of the economics of education, vol 1. North-Holland, Amsterdam, pp 512–576
Neal D, Rick A (2014) The prison boom and the lack of black progress after Smith and Welch. Working paper 20283, NBER, Cambridge
Ransom R, Sutch R (1986) The labor of older Americans: retirement on and off the Job, 1870–1937. J Econom Hist 46:1–30
Romer C (1986a) Spurious volatility in historical unemployment data. J Pol Econ 94:1–37
Romer PM (1986) Increasing returns and long run growth. J Pol Econ 94:1002–1037
Romer C (1999) Changes in business cycles: evidence and explanations. J Econ Persp 13:24–44
Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. J Pol Econ 82:34–55
Rosenbloom J (2002) Looking for work, search for workers: American labor markets during industrialization. Camb University Press, New York
Slaughter M (1995) The antebellum transportation revolution and factor price convergence. Working paper 5303, NBER, Cambridge
Smith J (1984) Race and human capital. Am Econ Rev 74:685–698
Smith J, Welch F (1989) Black economic progress after Myrdal. J Econ Lit 27:519–564

Steckel R (1983) The economic foundations of East–West migration during the nineteenth century. Exp Econ Hist 20:14–36
Sundstrom W (2006) Hours and working conditions. In: Carte S (ed) Historical statistics of the United States millennial edition, 2nd edn. Cambridge University Press, Cambridge, pp 46–54, 301–330
Taylor GR (1951) The transportation revolution, 1815–1860. Rinehart, New York
United States Census Bureau (2014) www.census.gov. Accessed 2 Sept 2014
United States Department of Labor, Bureau of Labor Statistics (2014a) www.bls.gov. Accessed Sept 2, 2014
United States Department of Labor, Bureau of Labor Statistics (2014b) BLS handbook of methods. http://www.bls.gov/opub/hom/. Accessed 2 Sept 2014
Vandenbroucke G (2009) Trends in hours: the U.S. from 1900 to 1950. J Econ Dyn Cont 33:237–49
Weiss T (1992) US labor force estimates and economic growth. In: Gallman R, Engerman S (eds) American economic growth and standards of living before the civil war. University of Chicago Press, Chicago
Weiss T (1999) Estimates of white and nonwhite gainful workers in the United States by age group, race, and sex: decennial census years, 1800–1900. Hist Meth 32:21–35
Whaples R (1990) Winning the eight-hour day, 1909–1919. J Econ Hist 50:393–406

The Human Capital Transition and the Role of Policy

Ralph Hippe and Roger Fouquet

Contents

R. Hippe (✉)
European Commission, Joint Research Centre (JRC), Seville, Spain
e-mail: ralph.hippe@ec.europa.eu

R. Fouquet
Grantham Research Institute of Climate Change and the Environment, London School of Economics and Political Science (LSE), London, UK
e-mail: r.fouquet@lse.ac.uk

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_79

Abstract

Along with information and communication technology, infrastructure, and the innovation system, human capital is a key pillar of the knowledge economy with its scope for increasing returns. With this in mind, the purpose of this chapter is to investigate how industrialized economies managed to achieve the transition from low to high levels of human capital. The first phase of the human capital transition was the result of the interaction of supply and demand, triggered by technological change and boosted by the demands for (immaterial) services. The second phase of the human capital transition (i.e., mass education) resulted from enforced legislation and major public investment. The state's aim to influence children's beliefs appears to have been a key driver in public investment. Nevertheless, the roles governments played differed according to the developmental status and inherent socioeconomic and political characteristics of their countries. These features of the human capital transition highlight the importance of understanding governments' incentives and roles in transitions.

Keywords

Human capital · Education · Economic history · Economic policy

Introduction

The transition in human capital is arguably the most important economic and social transformation to have occurred. After all, human capital accounts for two-thirds of overall wealth (i.e., including natural capital) in most industrialized economies (Hamilton and Liu 2014) and, based on rough estimates, has over the last 200 years (McLaughlin et al. 2014). Reflecting its importance, the Council of the European Union states that "[e]ducation and training have a crucial role to play in meeting the many socio-economic, demographic, environmental and technological challenges facing Europe and its citizens today and in the years ahead" (Council of the European Union 2009, p. C 119/2). Given the fundamental role human capital plays in economic growth and development, and especially for the "knowledge economy," research related to growth needs to include human capital in its analysis (e.g., Hippe and Fouquet 2018).

The particular aim of this chapter is to better understand the following questions: What factors have driven the human capital transition since before the Industrial Revolution? Was there a market failure in the area of human capital such that governments had to step in to foster increased investment? What role did the state play in the provision of mass schooling? What drove the public's willingness to sacrifice scarce resources for this public good? What role did other stakeholders play?

The human capital transition over the last 500 years, triggered by the development and diffusion of the Gutenberg Press, and, particularly, in the 200 years, with expansion of public schooling, offers numerous lessons for growth. First, it tells us (part of the story of) how we got here – how one of the pillars of the "knowledge

economy" developed. Second, it offers an example of a major transition of a key factor of production (hence, crucial for understanding economic development and growth). Third, it provides a lesson about a major transition where market failures were prevalent and government stepped in and provided education. In all cases, understanding past transitions will offer insights into how to promote future transformations.

The history of the human capital transition highlights how the state can have a major role to play in future transitions related to investment in education. Even though demand for a knowledge economy in all its dimensions might be increasing in the most industrialized countries, an effective transition is likely to require further state action. Where demand is especially lacking, such as in less developed countries, the state's role in achieving a transition to a knowledge economy will be particularly acute. A crucial factor in determining whether the transition moves beyond a first phase may well be the extent to which government directly benefits from the transition.

The chapter is organized as follows: First, we highlight economic development in the long run and the contribution of human capital. Second, we review some of the basic principles of human capital theory. Then, we consider the characteristics and the evolution of the apprenticeship system and the production of books. Subsequently, we emphasize the role of different origins of demand for education. Government policy and its effectiveness are more closely analyzed in the following section. The last section considers different indicators that show the evolution of human capital during the last 200 years and the actions taken by governments in education. Finally, a conclusion sums up the lessons from the human capital transition.

Long-Run Economic Development and Human Capital

The idea that human capital may be considered as a determinant of economic development has a long history (Demeulemeester and Diebolt 2011). Smith and Marshall had already incorporated the notion of something akin to human capital in their thinking. However, the first exogenous growth models developed by Solow and Swan (Solow 1956; Swan 1956) only incorporated capital, labor, and technological progress in the aggregate production function. Subsequently, the importance of human capital was put forward in particular by Becker (1981), an important founder of human capital theory. Furthermore, Nelson and Phelps (1966) emphasized that human capital is important for implementing and adopting new technologies. Schultz (1975) argued that workers are better able to cope with changes in the economic structure and handle new technologies if they have more human capital. An extension of the original Solow-Swan model, the human capital augmented Solow model, was presented by Mankiw et al. (1992). It explicitly included human capital as a factor in the production function.

Nevertheless, the prominence of human capital's role emerged following the development of endogenous growth models. Romer's (1986) work can be seen as

the first important contribution in this field. These "new growth models" aim to endogenize the different sources that lead to growth. In this way, the growth rate is not exogenously determined (as in the Solow-Swan models) but is established within the endogenous growth model itself. The overall category of endogenous growth models can be divided into two main approaches (Aghion and Howitt 1998). The first line of thought focuses on human capital accumulation as the main driver of growth (Lucas 1988). The second approach underlines the importance of technological change for the creation of economic growth (Romer 1990). Since the original contributions by Lucas (1988) and Romer (1990), the literature has steadily been advancing (see Ang and Madsen 2011). In fact, these initial models (and models by, e.g., Segerstrom et al. 1990; Grossman and Helpman 1991; Aghion and Howitt 1992) are now considered as the first generation of endogenous growth theory models. A second generation of models takes semi-endogenous (e.g., Jones 2002) or Schumpeterian approaches (e.g., Aghion and Howitt 1998; Peretto 1998; Young 1998). The appropriateness of these new types of models is still being debated and empirically tested (e.g., Madsen 2010; Madsen et al. 2010; Ang and Madsen 2011).

While endogenous growth models have been central to human capital's prominence in economic growth theory, they are limited – in the sense that they are not able to explain the process of economic growth since the beginning of human existence (Galor 2005). A new ambitious theory has been developed aiming at understanding the growth patterns of human kind in the very long run. This unified growth theory (UGT) has particularly been advanced by Galor (e.g., Galor and Weil 2000; Galor and Moav 2002). UGT classifies economic growth in the very long run in the following way. First, a long phase of hunting and gathering was typical for human development between one million BC and 8000 BC. After the Neolithic Revolution, the Malthusian growth regime began. It was characterized by a per capita income which always fluctuated around low levels of subsistence. Accordingly, Fig. 1 shows that GDP per capita in several European countries and China was rather low and fluctuating between 1300 and the takeoff during the Industrial Revolution. Thus, the Industrial Revolution allowed the European countries to escape the Malthusian growth regime. This post-Malthusian growth regime lasted until the demographic transition in Europe around 1870. Economies and populations grew substantively during this epoch.

The final growth pattern is the modern growth regime, which has characterized the world's advanced countries from the nineteenth century onward. This regime shows an acceleration of technological progress. Moreover, human capital is increasingly demanded. Thus, UGT postulates that human capital played a crucial role in the transition from the post-Malthusian regime to modern growth. Human capital and technological progress brought about the demographic transition, which has led to lower population growth. The overall result has been high and sustainable growth in per capita output in the most advanced countries. However, not all countries have reached the modern growth regime, and some countries reached it earlier than others.

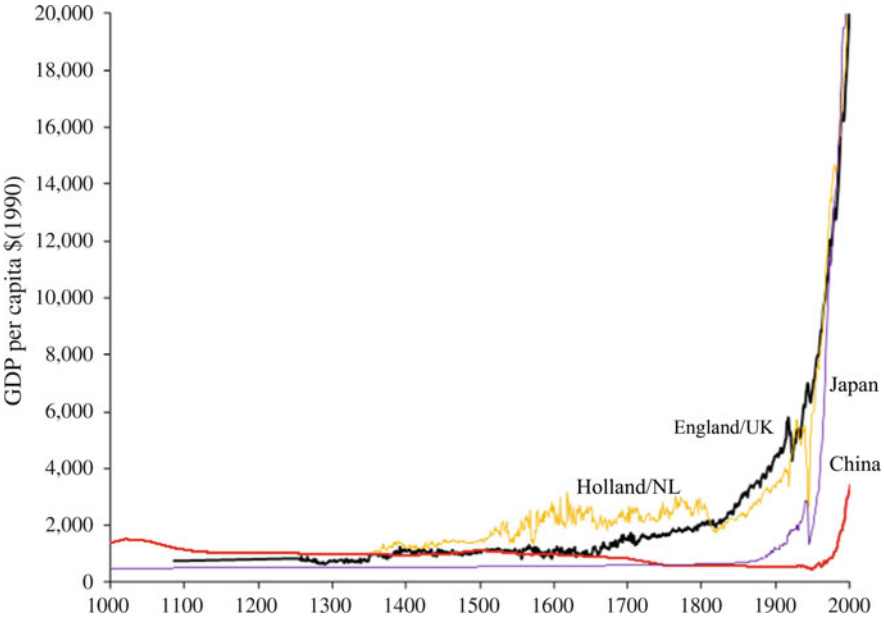


Fig. 1 GDP per capita in selected European and Asian countries, 1000–2000. (Source: Fouquet and Broadberry (2015), Bolt and van Zanden (2014))

Therefore, a Great Divergence has emerged between the most advanced countries and other countries.

These different theories show that human capital is an important driver for economic growth. Still, there has been some controversy about this issue over the last decades. In fact, Demeulemeester and Diebolt (2011) show that there have been several alternating waves of optimism and skepticism since the Second World War. The early contributions led to the consensus in the 1950s and 1960s that education makes an important contribution to economic growth. In contrast, the 1970s were characterized by skepticism in a time of economic downturn. The new important theoretical contributions of the 1990s once again reinvigorated the case for human capital. These optimistic ideas were supported by different empirical studies (e.g., Barro 1991; Mankiw et al. 1992; Barro and Lee 1993), but also more critical voices appeared, such as Benhabib and Spiegel (1994) and Pritchett (2001). However, measurement errors may account for some of these more pessimistic results (Krueger and Lindahl 2001). Thus, Sianesi and van Reenen conclude in their literature survey in 2003 that "as a whole we feel confident that there are important effects of education on growth" (Sianesi and van Reenen 2003, p. 197). Finally, more recent studies by, e.g., De La Fuente and Doménech (2006), Cohen and Soto (2007), Hanushek and Woessmann (2008), Ciccone and Papaioannou (2009), and Gennaioli et al. (2013) have further emphasized the crucial impact of human capital on growth and development.

Principles of Human Capital Theory

In the economics of education, the acquisition of human capital (through education or training) is assumed to give utility to an individual (see Brewer et al. 2010). It is an investment decision which enables an individual to obtain higher monetary returns in the future. The investment in human capital increases the knowledge and skills of an individual and thus his productivity (e.g., Schultz 1961; Becker 1964; Schultz 1971; Lucas 1988). This investment comes at an expense due to direct costs (e.g., schooling fees), foregone earnings, psychic costs related to the studies, and other related expenses. Still, the private returns of human capital acquisition can be numerous, e.g., higher earnings (Mincer 1958), a better job, and a lower probability of being unemployed. Further positive effects include a higher social status and social prestige, better health, higher social capital, higher cultural capital, and other benefits that are valued by an individual. The combined positive effects of human capital give incentives for individuals to invest in human capital to maximize their well-being during their life. Therefore, human capital theory suggests that a higher level of education is correlated with a higher level of earnings. This positive relationship has been found in many empirical studies, although the return to education can be influenced by many external factors. For example, 1 year of increased schooling gives a global average return of 10% (Psacharopoulos and Patrinos 2004).

Overall, an individual is part of a larger market which is characterized by demand for educational goods and its supply, leading to an equilibrium process whose outcomes can be observed. Thus, the consumption and investment characteristics of education make it similar to other durable goods. Thus, education cannot be described by the use of static models. In consequence, a longer-term view is inherent to the human capital concept. Therefore, the decision to invest in human capital has to be made in accordance with the expected discounted future returns and costs involved.

In addition, human capital may not only have important private returns but also significant social returns (see Lange and Topel 2006). These human capital externalities can be defined as "the sum of the private and external marginal benefits of a unit of human capital" (Lange and Topel 2006, p. 461). Human capital externalities have been analyzed from different perspectives. In particular, growth theorists such as Uzawa (1965) and Lucas (1988) have put forward the positive effects of interactions among individuals, leading to higher social than private returns of human capital. Individuals that have a higher endowment of human capital may increase the productivity of others, so that human capital accumulation leads to an increase in total factor productivity. On the other hand, it is also possible that positive externalities arise, which are not directly associated with productivity, such as lower criminality (Lochner and Moretti 2004), higher political participation (Friedman 1962), externalities connected to consumption, and higher economic growth (Sturm 1993; Hanushek and Kimko 2000; see Brewer et al. 2010).[1]

[1] In contrast, another strand of the literature suggests that there might also be negative externalities because human capital would not increase productivities but only waste valuable resources due to signaling effects (e.g. Spence 1973). However, Lange and Topel (2006) do not find important negative impacts; the positive effects largely dominate.

These social returns to human capital have been generally perceived by governments. Therefore, governments have spent an increased amount of public expenditures on education in the long run. One fundamental question that arises in this context is the allocation of these scarce resources to maximize human capital output. The state may not allocate these resources in an efficient manner because, in standard economic theory, markets provide an efficient method for allocating scarce resources. Still, it is possible that markets fail in their attributed task, so that a market failure occurs. Different reasons may lead to such a failure (Brewer et al. 2010). Firstly, it is possible that markets are not perfectly competitive. Thus, the market may become dominated by one or more agents who are able to set the prices. In the area of education, schooling markets are far from perfect. Secondly, information asymmetries may arise in a market. In particular, the quality of products or services may not always be perceivable by consumers due to a lack of information. In education, parents may not have sufficient information and may not perceive the true value of education. In addition, they may choose schools that may be socially suboptimal (e.g., preference for schools with a certain social or racial profile and not for the quality of a school). Although the market may satisfy the preferences of parents, the obtained solution may not be optimal for society as a whole. Thirdly, the existence of externalities suggests that consumption and production may have effects that are not included in prices. For example, the knowledge and learning of others can have a positive effect on an individual. Furthermore, the individual's decision to invest in education may have positive benefits but also costs for society. However, an individual does not take all of them into account when taking his investment decision. Even if the social benefits are higher than the costs, it is still possible that an individual may choose not to invest and consume a socially optimal amount of education. In this case, governments can intervene by making a certain minimum amount of education obligatory through compulsory schooling. Finally, public goods may not be sufficiently provided by markets because they are nonrival and nonexcludable. To some extent, the notion of public goods applies to education. For example, the consumption of services in the area of education does often not preclude others. A pupil obtains education together with other pupils by a teacher. Libraries give access to books to many individuals. Thus, education may not be sufficiently supplied in a perfectly competitive market, as is the case with other public goods.[2] For this reason, this has led to the "widespread belief that the market for educational services fails when left to its own devices" (Johnes 1993, p. 14).

Therefore, governments have intervened to avoid the consequences of market failures in this area. They have regulated the education sector, taken an important share of the financial burden, and operated education facilities. This involvement of the state may have led to suboptimal educational outcomes in some areas.[3] However,

[2]In addition, the assumption of perfect competitive markets implies that capital markets should be perfectly functioning. In this case, parents should always be able to find a way of financing the education of their children. This, however, is not the case either (Johnes 1993).

[3]For example, the existence of public schools may have led to a monopoly of these schools in certain geographical areas. Taking the point of view of the market, schools may thus not have been under market pressure to ensure quality standards and low operating costs. Some recent reforms have been aimed at improving the status quo and in part reorganizing the involvement of the state in this sector (Brewer et al. 2010).

the aim of the state may not always have been to maximize human capital levels. Schooling is never neutral but is a process of socialization that produces beliefs alongside human capital. Therefore, authors of the so-called class-conflict model suggest that there is a struggle between different classes of society in the area of education. The dominating elites of a country are assumed to use formal schooling to impose their social, cultural, and economic values and structures on other parts of society (Bowles and Gintis 1976; Carnoy and Levin 1985; Fuller and Rubinson 1992). Taking a more positive interpretation, one could also argue that "public schooling can promote social cohesion among disparate social groups and alleviate ethnic tensions by providing a core set of common norms that foster trust and promote interaction among individuals" (Gradstein et al. 2005, p. 5).

More generally, governments would not need to produce schooling but could sponsor the private sector to provide schooling. The simple production of skills could be contracted out, and private schools could be subsidized by the use of vouchers or mandates. Yet a government would not be able to control (cultural and ideological) beliefs, and private institutions (e.g., particular religious and social groups) could produce beliefs at school that could challenge the government and the political system.[4] Furthermore, egalitarian principles have given the state the moral responsibility to provide schooling to ensure that all children obtain a certain minimum level of education so that every child could have equal opportunities in later life. However, the external monitoring of schools is costly and difficult to implement effectively for the state. Therefore, it is argued that given the socializing and equalizing aspects inherent to education, a government needs to produce schooling in a direct way at arm's length (Pritchett 2003; Gradstein et al. 2005).

In addition, it should be mentioned that there are different types of schools and forms of training. More generally, according to Johnes (1993) one can distinguish between general and specific human capital. General human capital can be defined as skills and knowledge that can be used in any work context and increase the productivity of an individual. General human capital can take the form of skills such as numeracy and literacy. On the other hand, specific human capital can only be used in a specific work context to increase an individuals' productivity. For example, some of the specific skills of craftsmen could not be used in other professions. Similarly, if an individual works for a monopsonist, it cannot transfer its skills obtained through training to another employer either.

[4]For example, Milton Friedman also suggests that "a stable and democratic society is impossible without a minimum degree of literacy and knowledge on the part of most citizens and without widespread acceptance of some common set of values. Education can contribute to both" (Friedman 1962, p. 86; in Gradstein et al. 2005, p. 5). He further argues that "the major problem in the United States in the 19th and early twentieth century was not to promote diversity but to create the core of common values essential to a stable society… Immigrants were flooding the United States… speaking different languages and observing diverse customs … The public school had an important function in this task, not least by imposing English as a common language" (Friedman 1962, p. 96, in Gradstein et al. 2005, p. 9).

More generally, investments in human capital can occur (in the terminology of Becker 1993) in the form of on-the-job training in firms (distinguishing general and specific training analogously to the presentation above), schooling, information, and health.[5] Schools are considered to be specialized institutes that produce training, whereas firms additionally produce goods. Therefore, Becker argues that firms and schools are in many instances "substitute sources of particular skills. This substitution is evidenced by the shift over time, for instance, in law from apprenticeships in law firms to law schools and in engineering from on-the-job experience to engineering schools. [. . .] [T]here are complementary elements between learning and work and learning and time" (Becker 1993, p. 51). The level of complementarity depends partly on the quantity of formalized knowledge that is available. In some cases, schools can thus be treated as a special form of firms. Therefore, the effects of schooling (as mentioned above) are identical to those of general training.

In addition, there are still other possibilities than on-the-job training and schooling to accumulate human capital and increase one's productivity. These are particularly related to the investment in information. For example, more information about wages and employment possibilities may positively affect the future earnings of an employee. Thus, increased information and knowledge about the economic, social, and political system might increase productivity and incomes.

Finally, investments in health may also lead to improved productivity. Physical and emotional health has significant effects on earnings. While throughout most of history (and still today in some countries) physical strength has been an important factor for expected earnings (at least for an important share of the population), knowledge and skills have replaced physical strength in more developed countries. Similarly, emotional health has important effects on productivity and output (see e.g., Layard 2005).

Traditional Education and Skills Transmission

Apprenticeships

The transmission of technical skills from generation to generation has been central to maintaining productivity levels over centuries, and the transmission of new skills (often in conjunction with new technologies) spatially has been key to raising productivity levels from one decade to the next. Apprenticeships had been the most important source of training and technical knowledge acquisition for many centuries (Epstein 2004). In England, for instance, "domestic, agricultural or industrial apprentices and live-in servants made up 15–20% of the adult male population

[5]According to Sweetland (1996), the area of education can be further broken down: "[t]here is formalized education at primary, secondary, and higher levels (Cohn and Geske 1990), informal education at home and at work (Schultz 1981), on-the-job training and apprenticeships (Mincer 1974), and specialized vocational education at secondary and higher levels (Corazzini 1967)" (Sweetland 1996, p. 341).

in early modern England" (Humphries 2006, p. 79, referring to Stone 1966). "In the late sixteenth and seventeenth centuries roughly two-thirds of the English male labor force had at one time or another been apprenticed in one of the greater cities, primarily London" (Epstein 1998, p. 707; see Rappaport 1989). Even in the eighteenth century, apprentices in the nonagricultural sectors accounted for between 7.5% and 10% of the total labor force (Humphries 2006; Wallis 2008).

The apprenticeship was a contract (of between 3 and 7 years depending on the craft and the country (Epstein 1998; Wallis 2008)) between a master and his "student." Without a contract guaranteeing some return on the investment, few tradesmen would have been willing to share their skills with young men who were likely to leave once the training was complete and work for higher pay.

In addition, from the seventeenth century, masters required apprentices to pay an upfront premium. In eighteenth-century England, premiums were typically between 5 and 10 pounds (roughly £(2000)500 and £(2000)1,000) in the trades and about 50 pounds (£(2000)5,000) in the professional sector (Minns and Wallis 2013). In addition to the nature of the craft and prestige of the guild, premiums also varied according to family connections, experience, expected future income, and other factors that gave signals about the probability of attrition and the level of productivity of the apprentice.

The premium amounted to close to 1 year of a craftsman's salary, and the apprentice generally depended on financial help from parents and the wider family. For instance, it would take a youth 2 years of work in the agricultural sector to earn 5 pounds. Although a premium had to be paid in many instances, the apprentice received accommodation, boarding and subsistence wages (around 5 pounds per year in the eighteenth century), and a training. This training also offered important informal connections to colleagues and potential clients who might help recover the very expensive setup costs of one's own shop, which was often 10–20 times the initial cost of a premium (Campbell 1747, in Minns and Wallis 2013).

The Role of Guilds

Apprenticeships had their roots in the craft guilds, which began to exist from the eleventh century (Lauterbach 1994; Epstein 2004) – although the degree of control over apprenticeships by guilds varied and was greater in many German regions than in England. Apprenticeships were mostly left to the private decisions of masters and apprentices in some parts of Spain and France (Wallis 2008). The guilds were associations of craftsmen with tight regulations, creating conditions similar to cartels. They had many functions; in particular, they "supervised job performance, work conditions, and quality of instruction; enforced contracts through compulsory membership, statutory penalties, and blackballing; and protected apprentices against poor training in craft-specific skills within oligopsonistic labor markets" (Epstein 2004, p. 382). Most guilds sought to limit the poaching of trained apprentices among workshops, protect the apprentices from excessive abuse, and, therefore, promote proper training (Epstein 1998, pp. 691–692). By the sixteenth century, a national system of technical training was being introduced in England (Humphries 2006,

p. 75). Thus, along with their other roles, guilds were influential in the production of knowledge in the form of human capital.

Guilds provided an information pool and network for a particular supply of skills across regions. Although these networks were limited and far from efficient, the distribution of information helped signal changes in the demand for the skills and, therefore, provide a supply, if necessary (Epstein 1998, p. 694). Searchers wandered the country assessing practices and demands. Journeymen, either as part of a guild or as independents, travelled the country to provide their skills. In other words, the guilds acted as a means of disseminating knowledge (in the form of human capital) throughout the economy.

In fact, many of the European rulers tried to attract these skilled workers to their cities since the Middle Ages, and particularly during the times of the Renaissance and following the Reformation, especially if they were coming from the enemy. This development found a peak when mercantilist states and guilds tried to inhibit the emigration of skilled workers from the 1650s onward. However, these attempts were mostly unsuccessful due to the lack of administrative capacities and the incentives provided by other states (Harris 1998). Thus, Epstein believes that "technological leadership moved over time from southern to northwestern Europe [. . .] largely thanks to skilled migrants" (Epstein 2004, p. 385).

Apart from the opposition of governments and guilds, other factors contributed to the limitations of technical knowledge transfer: trade secrecy, information costs, transport costs, and the nonexistence of a sufficiently high skills base, which made it difficult to locally integrate the knowledge and inventions of migrants. However, where this knowledge base was sufficiently high, the accumulated knowledge of a country transferred by skilled migrants could be combined with the local knowledge, often giving a technological lead to this country. Information costs and transport costs were lowered due to urbanization and state competition over time, making it easier for technologies to diffuse (Epstein 2004).

Guilds have frequently been associated with stifling innovation. At times, they were probably a powerful force in encouraging knowledge to be embodied in human rather than physical capital. Yet, they were also responsible for the invention and diffusion of others. Evidence suggests that, in the face of growing competition and expanding markets, the guilds' barrier to technological innovation declined after the later Middle Ages (Epstein 1998 p. 694). Furthermore, craft guilds did "increase the supply of technology in three ways: by establishing a favourable environment for technical change; by promoting technical specialization through training and technical recombination through artisan mobility; and by providing inventors with monopoly rents" (Epstein 1998, p. 701).[6] Thus, Esptein (1998, p. 704) concludes, somewhat controversially (as more evidence is probably needed to conclude fully),

[6]An interesting question is whether there were also incentives for inventing technologies to sidestep the guilds. Epstein mentions that inventors had an incentive to keep their inventions secret from the guild. Yet, "although technical secrets were often kept within the craftsman's family, it is unlikely that significant breakthroughs could withstand a guild's scrutiny for long. On the other hand, an inventor had to weigh the guild's offer of a temporary quasi-monopoly rent against the possibility of obtaining a one-off royalty (net of migration costs) from a rival craft or government" (1998 p. 704).

the two opposing forces of a monopolistic support system for invention and of the demand for ever-wider competitive markets in skilled labor provided a healthy source of technological innovation and diffusion.

Yet the government increasingly saw the need to intervene in the transmission of these skills. A new structure was set up that defined the legal framework and regulated the apprenticeship training system. For instance, the guilds were formally abolished by decree in Prussia in 1869 – more generally, the state saw the guilds as a competing institution for power (Smits and Stromback 2001; Epstein 2004). Although the guilds may have formally been abolished, Smits and Stromback (2001) argue that they were only "transformed into the Chambers of Crafts and retained substantial control over the apprenticeship system" in Germany (Smits and Stromback 2001, p. 17). Therefore, the organization and ways of training were not determined by the state but were left to the self-administering chambers. In Germany, it may have been a "managed transformation of the highly restrictive guild system to a regulatory regime that allowed industrialisation to proceed" (Smits and Stromback 2001, p. 17).

The Decline of Apprenticeships

While the apprenticeship system and occupational skills played an important role before and during the Industrial Revolution (Mokyr 2009; Minns and Wallis 2013), the social changes it caused brought about a crisis in apprenticeship training during the late 18th and nineteenth century (Lauterbach 1994), arguing that, at least in Germany, the quality of apprenticeship training was being neglected. Meanwhile, the duration of apprenticeships in England fell to 4 years in the late eighteenth century (Wallis 2008).[7]

Structural shifts altered the demand for human capital during the Industrial Revolution (Weedon 2003, p. 62). Zeira (2009 p.63) argues that "industrialization changed the type of human capital required for production. While prior to the industrial revolution human capital was mainly specific to the profession, the industrial revolution created demand for more general human capital, which included the ability to read, to write, and to perform calculations. Hence, while the apprenticeship system could provide most of the required human capital prior to the industrial revolution, following it there was increasing demand for people who have a much broader and flexible human capital, which can be acquired only at schools." Similarly, Collins and Halverson (2010, p. 23) add that "[a]fter the Industrial Revolution, schools stressed the learning of basic skills that children would need to function as intelligent citizens and workers and on the knowledge in the different

[7]A comment from the early 1960s, after the decline in this form of training, summarizes what happened: "apprenticeship has all but disappeared, partly because it is now inefficient and partly because schools now perform many of its functions. Its disappearance has been hastened no doubt by the difficulty of enforcing apprenticeship agreements. Legally they have come to smack of indentured service" (Schultz 1961, p. 10).

disciplines." Thus, the apprenticeship partly lost its role because there was a changing demand for human capital.

Its decline was also partly due to changes in the supply of other forms of education. In principle, dissemination of this specialized technical knowledge could take different forms: through printed media, patents, or migrants. Epstein (2004) argues that, especially early on, texts were not a very successful method for the dissemination of specialized technical knowledge because the manuals were often incomplete and did not include some of the essential information to actually put a new technique into practice. Yet the costs of written education changed over time. In comparison with schooling, an apprenticeship was a much more expensive alternative. Minns and Wallis (2013) estimate that schooling costs were about 1 pound per year in the eighteenth century.[8] Eventually, in the eighteenth century, books on craft skills became the third most important category of books (the first being religion and the second being law), highlighting the increased importance of reading skills for tradesmen and craftsmen (Cook 2006).

Catalyzing the Human Capital Transition

The Gutenberg Revolution

Critical to the transformation of education in Europe was its ability to reduce the price of books and the written word. Before the invention of the printing press, book production involved the production of handwritten manuscripts. In addition to other costs, this was a very time-consuming exercise, as Clark (2007) illustrates (see also Clark and Levin 2001): Copyists were able to copy 3,000 words (of plaintext) per day, implying that one copy of the Bible took 126 man-days (Clark 2007). Because the copying of a text by hand needed more space than printing it, the area needed per word was also two times greater, further increasing the costs of the employed materials (Clark 2007). For this reason, books cost more than 1 year's salary for an average man and were, therefore, luxury goods (van Zanden 2009a).

Nevertheless, the supply and the demand of manuscripts were promoted by the growth of monastic institutions from the sixth century (see Fig. 2). Additional demand came from the growth of cities throughout Europe between the 12th and 15th centuries. Cities were characterized by an ever more elaborate division of labor, and they generated economic and military demands for individuals who were able to keep records, were trained in administration, and were able to communicate properly. Bureaucracy expanded continuously, and the oral tradition was slowly replaced

[8]They estimate this number in the following way: "Assuming that youths would earn a rising fraction of adult income with age (20% at age 14, 40% at age 15, 60% at age 16, 70% at age 17, 80% at age 18, 90% at age 19, and [100]% at age 20 – see Van Zanden (2009b), p. 160), a provincial adult unskilled wage of 1 s per day, that youths work 228 days per year (Voth 2001), and a discount rate of 7.5%, the present value of lost earnings during an apprenticeship, relative to a subsistence income of £5 per annum, was about 26 pounds" (Minns and Wallis 2013, p. 344).

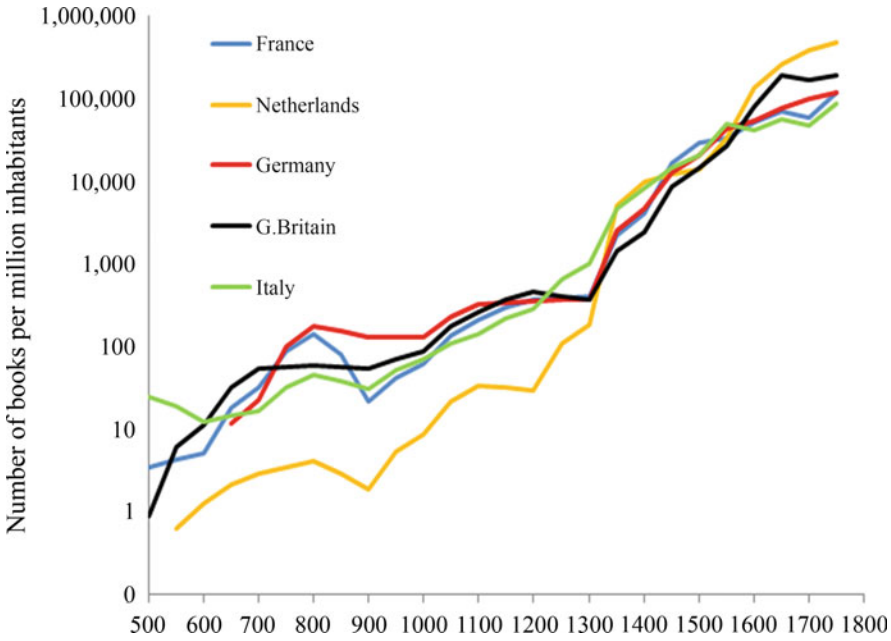


Fig. 2 Book production in Western Europe (per million inhabitants), 500–1750. (Source: Derived from Buringh and van Zanden (2009))

by the written word. For this reason more young people began to learn to read. In addition, universities were established for the purpose of advanced training and in turn created further demand for themselves (Venezky 1996, see also Hippe 2013a). In addition, the price of book materials was also lowered by economies of scale and learning effects associated with the increased use of paper (van Zanden 2009a). As a result of rising demand and declining costs, manuscript production increased until the eve of the invention of the printing press (see Fig. 2). It is estimated that more than 10,000 copyists worked in manuscript production in Paris and Orléans alone shortly before Gutenberg's printing press (Chassant 1846).

The invention of the printing press by Johannes Gutenberg in Mainz between 1446 and 1450 was, then, partly stimulated by this ongoing and increasing demand for written documents. In fact, Gutenberg was not the only one who tried to improve the existing manuscript production system. Different methods had been explored, including xylography and tabular printing. However, Gutenberg's printing press proved to be superior in quality and lower in cost for the production of books (Guellec 2004). Therefore, the printing press was not invented by accident but after many years of diligent pursuit carried out by Gutenberg.[9]

[9]In fact, the printing press was first invented in Korea. For more details, see Hippe (2015).

The fact that the printing press was invented in Mainz is itself not so much accidental, as the city was part of the Rhine valley, which was quite industrialized for its time and specialized particularly in metallurgy. In consequence, the broad timing of the printing press was the outcome of the several mentioned demand factors. These factors were again conditioned by the availability of paper in Europe in general, and in Germany in particular. Furthermore, the growing market made the development of the printing press increasingly competitive relative to manuscript production (Guellec 2004).

The printing press fundamentally changed the world of book production and knowledge diffusion. It enabled a much faster and less costly production of books. Cuijpers (1998) indicates that printed books were about 50–80% cheaper than their handwritten homologues in the 1460s. Furthermore, the printing press was a technology which was skill and capital intensive. Therefore, this high-tech innovation corresponded perfectly to the needs of the European economy, which was characterized by high labor costs (in contrast, e.g., to the Chinese economy) (van Zanden 2009a). Thus, the new technology spread quickly throughout Europe.

The spread of the printing press allowed a spectacular increase of book production, knowledge diffusion, and human capital accumulation. In the earliest centuries presented here (500–700), overall book production in Europe was restricted to only 12,000 manuscripts written every century, whereas in the eighteenth century, the number increased to more than 1,000,000,000 books (Buringh and van Zanden 2009). The decisive massive increase in book production clearly occurred during the period after the printing press was invented. The data show that more books (and manuscripts) were produced in Europe in the second half of the fifteenth century than in the entire 1000 years before the invention of the printing press. This radical change in the production of knowledge is also highlighted by the fundamental decrease in the price of books. The fall in book prices had the effect that a literate person was able to consume more books, which in turn increased the incentive (and reduced the cost) to learn to read. Furthermore, the increased output of books itself created economies of scale, lowering book production costs even further and leading to higher price reductions. Van Zanden (2009a) estimates the real price of books (i.e., deflating the book price by a cost of living index) for the Netherlands between 1460 and 1800 (see Fig. 3).

The dramatic fall in the real price of books is more than evident. Van Zanden (2009a, referring to Cuijpers 1998) illustrates this change by using the Gutenberg Bible as a point of reference. An original copy of the Gutenberg Bible was sold at a rate of about an annual wage gained by a laborer. A Bible of comparable quality could be obtained for a price of less than 12 days of a carpenter's wage before the end of the eighteenth century (and lower quality bibles were also available at prices equivalent to a daily wage).[10] Thus, van Zanden concludes that "within the span of one generation –

[10]This fall in prices also had a major impact on the spread of the ideas of Protestantism. For example, Luther's translation of the New Testament in 1522 was affordable even to laborers (Stöber 2004).

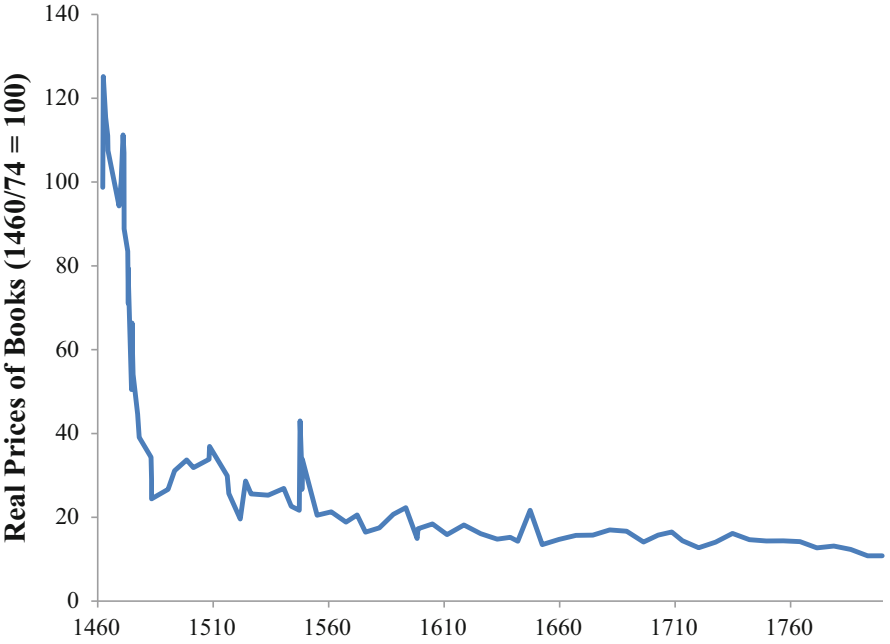


Fig. 3 Estimates of real prices of books in the Netherlands (1460/74 = 100), 1460–1800. (Source: Van Zanden (2009a))

from 1455 to 1485 – book prices may have declined (in real terms) by 85–90%, a revolution in the price of communication comparable to current developments in ICT [Information and Communication Technology]" (van Zanden 2009a, p. 182).[11]

Early Private Demand for Books and Literacy

Clearly, the dramatic decline in book prices following the diffusion of the Gutenberg Press was crucial to increasing the spread of the written word. At the beginning of the sixteenth century, literacy rates in most European countries were below 10%, possibly below 1% in German states (Engelsing 1973; Becker and Woessmann 2008). Then, over the next three centuries, particularly in North-Western Europe, such as the German states, Holland, Sweden, and England, greater access to books enabled literacy rates to increase (see Fig. 4 for England). However, literacy rates would not have reached 50% in the second half of the eighteenth century without a strong demand to read (Mitch 1992b).

[11]Similarly, Clark (2004, p. 8) calculates that "the estimated price of a standard page of text in the middle ages was 50 times the price in 1700–59."

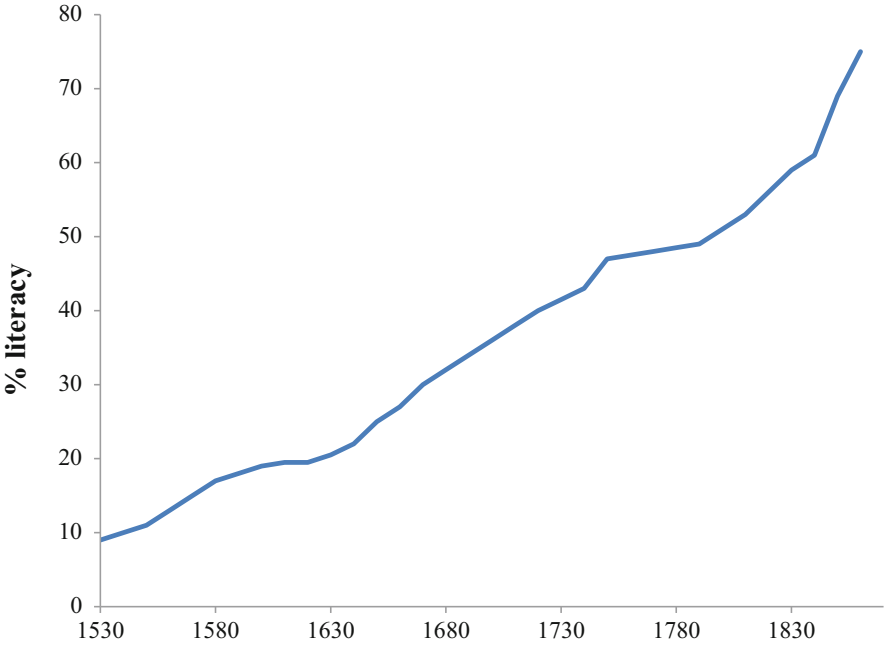


Fig. 4 Literacy in England, 1530–1870. (Source: Cressy (1980))

In fact, demand for literacy, numeracy, and a general education stemmed from many different private and social or public sources (e.g., Cipolla 1969; Fuller and Rubinson 1992; Mitch 1992a; Galor 2011). Commercial demand (i.e., coming from the trade sector), spiritual demand (from the Churches), military demand (to create a more efficient army), industrial demand (e.g., to increase labor productivity), parental demand (e.g., to improve the prospects of the future career of their children), status demand (to distinguish oneself as part of an elite), and "belief-formation" demand (to increase social control and advance nation building) were all private or social drivers to invest in education.

Mitch (1992b) has emphasized that early advances in literacy levels were brought about by private demand, rather than any direct actions by the state. "[T]he success of printing should probably be connected with more general changes in society. The end of the Middle Ages was marked by the rise of the bourgeoisie. Now that it controlled new economic and commercial sectors, the middle class intended to participate in political decisions that concerned them as well, and it signalled its social success by paying more attention to culture, which it adapted to its own interests" (Gilmont 1999, p. 215). Indeed, the percentage of households owning books increases from one in ten in 1560, one in four in 1580, and one in three in 1590 to nearly one in two by 1620 (Morgan 1997, p. 14). Similarly, the average size of personal libraries grew (Gilmont 1999). Increases in book collections offer an insight into the different income elasticities among different professions and social status.

"Between 1500 and 1525 the average physician's collection increased from 26 to 62 books; the average jurist's from 25 to 55; the average merchant's from four to ten and the average artisan's from one to four" (Morgan 1997, pp. 13–14).

Basic education in the form of literacy or numeracy had become increasingly important in economic affairs throughout history. Cook (2006, p. 71) suggests that "one of the earliest exceptions to the exclusion of literacy to all but social elites was its appearance among the trades," in particular due to the need of record keeping. For example, reading and writing ability became a prerequisite to obtaining an apprenticeship with English goldsmiths in 1478 (Anderson 1965; Venezky 1996). Increasing urbanization led to an economic environment that demanded more bookkeepers, administrators, and clerks, and more sophisticated forms of bureaucracy. More generally, "[l]iteracy continued [. . .] to be driven by practical utility, appearing first and to the highest degree in those trades most closely associated with the market economy (Thomas 1986), and in areas where new technologies were forcing change, such as navigation and warfare" (Venezky 1996, p. 59).

Meanwhile, demand for spiritual guidance, and the associated religious competition, was crucial ingredient for the explosion of printed works, particularly from the early sixteenth century. As already mentioned, the increasing number of monasteries throughout Europe during the Middle Ages reflected the growth in the supply of and demand for spiritual guidance.

With the cheaper costs of printing, Luther could dare to propose that everybody should be able to read the bible. Given that fewer than one German in one hundred was able to read at that time, his ambition must have seemed just a utopian vision (Engelsing 1973, Becker and Woessmann 2008). However, it is important to remember that for the first 70 years after the invention of the printing press, around 1450, the Bible had been printed in Latin. Most people did not read, but it made it harder to learn to read when the Bible was in Latin. Although translations had existed before (into Old French, Spanish, Catalan, and German in the thirteenth century and the Wycliffe Bible into English in 1383), these were usually banned and as a result were very hard to access and extremely expensive (Biller and Hudson 1996). The first printed non-Latin Bible was into Greek by Erasmus in 1516. However, it was Luther's translation into German of Erasmus' Greek Bible in 1521 that made the Bible accessible and at a low cost, electrifying the book market (see Fig. 5). In 1526, a Dutch Bible was printed; another German Bible was out of Zurich in 1530, in Italian in 1532, and in French in 1535 (Gilmont 1999). As a consequence, one of the most powerful drivers of European demand for books from the 1520s onward was the translation of the Bible into European languages.

Early Spiritual Demands for Mass Education

With the aim of ensuring that all should read the Bible, Luther argued for the establishment of compulsory schools and claimed that local rulers were responsible for their establishment. This would have effectively implied a shift of instruction from religious to secular authorities. In consequence, a range of ordinances

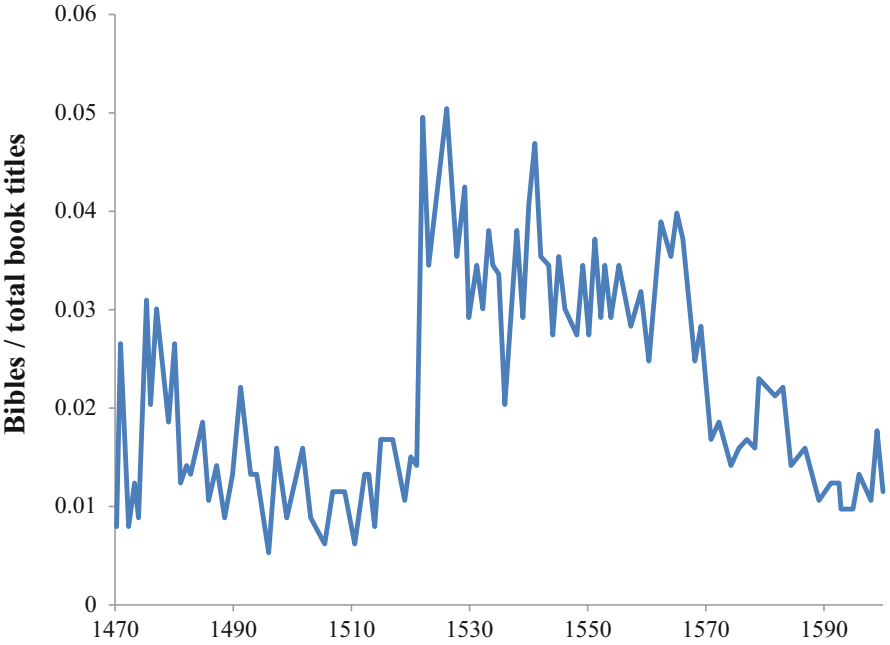


Fig. 5 Share of bibles in total book output in Europe, 1470–1600. (Source: Dittmar (2013))

that promoted schooling were passed in Protestant territories and their implementation was controlled (Becker and Woessmann 2009). For example, a first landmark church order was passed in Braunschweig in 1528, supporting the creation of new schools and defining the school curricula. It was soon followed by others (e.g., Saxony in 1557, Württemberg in 1559, Lüneburg 1564) (Green 1979). Yet Luther's preaching did not lead to universal mass schooling (Stone 1969; Ramirez and Ventraensca 1992).

Luther also urged parents to value the education of their children and send them to school. For example, he writes that "I see that the common people are dismissive to maintaining the schools and that they withdraw their children from instruction altogether and turn solely to the care for food and bellies, and besides they either will not or cannot consider what a horrible and un-Christian thing they are doing and what great and murderous harm they are doing everywhere in so serving the devil" (Luther 1909, p. 526, as cited in Becker and Woessmann 2009, p. 541). Thus, Luther's insistence over moral obligations might have altered the beliefs of parents and rulers about the benefits of educating children (Becker and Woessmann 2009).

At first, the Catholic Church attempted to ban Bible translations. Similarly, the state sought to censor publications. Henry VIII's experience highlights the difficulties of introducing policies to control information production and dissemination. In the early sixteenth century most publications read in England were printed on the continent. Following Henry VIII's break with the Catholic Church, Luther's

sympathizers were frustrated by Henry's lack of theological reforms and "bombarded England with highly aggressive pamphlets printed in Antwerp" (Gilmont 1999, p. 216). Unable to censor these publications, Henry VIII encouraged printers to be set up in England, which they did over the next two decades, and thus, he was better able to police their activities (Gilmont 1999).

Although the Catholic Church's Counter-Reformation[12] tried to undo many of Luther's teachings, it eventually embraced the ideal of improving literacy levels and the concept of schooling (Ramirez and Ventraensca 1992). Religious competition,[13] and the threat of losing an even greater share of the "market for spiritual and moral guidance," may have been an important factor. "In the face of Protestant competition, religious and secular authorities in Catholic areas also increased their efforts to provide popular schooling; it is notable that the parts of Europe where schooling lagged most seriously were those, like Spain and Italy, where the Counter-Reformation triumphed completely and eliminated religious competition" (Glenn 2012, p. 140).

Overall, the churches played a major role in the supply of schooling for the next 300 years. Their importance only declined with the rise of the national state. In the nineteenth century, religious authorities feared that state-provided education would be secular and educational content not in line with their own views (Boli 1992). This was the case of the Catholic Church in France. In France enrollment was relatively widespread before the central state developed a modern organized entity (Archer 1979) because the Catholic Church was interested in controlling the socialization of children. Later competition with the state further increased French schooling (Fuller and Rubinson 1992). Thus, enrollment rates of children were about one-third in the 1830s and reached 80% by the 1880s (see Galor 2011).

In other locations, particularly in Protestant regions, the state worked together with the Church. In particular, the early worldwide leader in reading, Sweden, passed a Church law in the seventeenth century with the intention of spreading literacy due to religious motivations. Thus, Venezky argues that "reading ability was pressed on the population by church and state [. . .] through regular parish examinations, fines for parents who failed to teach their children and denial of Communion and marriage rights to those adults who could not read and recite the catechism" leading to almost universal reading ability in the population until 1750

[12]See Ekelund et al. (2002) for an economic interpretation of the Protestant reformation and Ekelund et al. (2004) for an exploration of the economics of the Counter-Reformation.

[13]The question whether religious competition leads to more or less religious participation by individuals is still disputed. On the one hand, it is argued that the existence of various religions leads to a decrease in the plausibility of a given religion and thus less religious participation (Chaves and Gorski 2001). On the other hand, authors such as Adam Smith suggest that a non-state-sponsored religious group has to provide special care for its believers, raising the quality of and participation rate in religious activities (Iannaccone et al. 1997). For more information, see Höhener and Schaltegger (2012).

(Venezky 1996, p. 48). Similarly, the state cooperated with the Lutheran Church in Prussia (Soysal and Strang 1989).[14]

In general, where the state was weak, it had to rely more on the infrastructure of the Church, giving the Church more power and influence (Vincent 2000). The USA is another example where enrollment rates were high in the presence of a relatively weak central state. Meyer (1989) proposes that this fact can be explained by a particular faith in schooling and literacy (due to Protestantism) and its positive effects on nation building. These ideas originated in Britain and were spread to America. Once again, spiritual demand had been an important driver of education.

Demand for Education

The Pros and Cons of Mass Education

Although some early reformers had promoted mass education, formal schooling was not widespread before 1800. Instead, most education was provided in more informal settings. The diversity of opportunities to become literate is highlighted by Cipolla (1969), making reference to Rashin's (1958) study. Still, in 1883–1884, the study shows that in a province of Moscow, Russian Empire, the distribution of 7,123 factory workers with the capacity of reading according to the source of their education is that 38% became literate in their village, town, or district schools, 36% outside school, 10% in factory schools, 9% with clergy, and 7% during military service (Rashin 1958).

Mitch (1992b, referring to Briggs 1978; Mitch 1982) claims that fees for private schools at the elementary level were at levels that working-class families could afford in many European countries between the 16th and nineteenth century. He further argues that private instruction could then have achieved universal literacy without the state.

Yet different classes tended to have different demands for education. Whereas the middle class sought secondary or tertiary education for their children to move up the social ladder, workers might not seek education beyond the primary level because they failed to see the benefits of education. In this way educational differences among the classes, and perhaps also regions of countries, were exacerbated, and access to higher status jobs was often limited to particular groups of society, hindering mobility between social classes.

This point illustrates that the ruling elites in many countries were wary of the threat that more education posed to their political, economic, and social position. These elites feared that education would make the workers despise their lot in life, a

[14]More generally, the authors see the values in formal education close to those of Protestantism, as highlighted by Weber (1958): "formal education's emphasis on individual socialization and achievement parallels the Protestant emphasis on the individual's unmediated relation to God and individual salvation" (Soysal and Strang 1989, p. 279).

life that had been assigned to them "by nature." Therefore, education was perceived by some to lead them to seek more rights and make them more prone to resist domination (Graff 1991; Lindert 2004).

Despite these upper-class fears, there had been a growing demand for mass education for military and industrial reasons. Indeed, supporters of military conscription had particular interest in mass schooling. In fact, there had been an important demand for gunners in the sixteenth century (Cipolla 1969). Because gunners had to be literate to perform their profession, a number of European governments created specialized schools for their education. There they learned to read, write, and calculate, as well as some basics in ballistics. Sweden had promoted mass schooling for instilling national loyalty, teaching discipline, and lessening social tensions among the population, and their leaders found that literate soldiers were also considered much more effective in warfare (Malmström 1813). Similarly, Napoleon was aware of this fact and promoted the education of his recruits. Later on, his enemies used this insight to create an educated army (Vincent 2000). For instance, defeated Prussia successfully reformed its education system. This may have been one of the reasons Prussia (and the other German states) won the Franco-German War in 1870. "[I]n 1870 there were more than 20 per cent illiterates among the recruits of the French army while there were only about 3 per cent among those of the Prussian army [. . .]. As the French remarked at the time: Sedan est la victoire du maître d'école allemande" (Cipolla 1969, pp. 23–24, see also Fourrier 1965). Cipolla states that "[s]ocieties which produced an increasing number of literate soldiers had a decisive advantage over those that failed to do so" (Cipolla 1969, p. 23). Accordingly, Vincent notes that "[a]s Europe prepared for war, most of the potential combatants had ensured that their recruits would be able to read the instructions on their weapons and write back to their families" (Vincent 2000, p. 10). In other words, military objectives played a role in the provision of education by the state in the nineteenth century.

Books in the Industrial Revolution

The basic technology did not significantly change after Gutenberg until the nineteenth century (Chappell 1970). Afterward, the Industrial Revolution had a dramatic impact on the supply of printed materials. Steam power began to be used in the pressing process as early as 1810. At the same time, the first successful attempts were made with machine-produced (pulp fiber) paper, and machines for mass production entered the market in the 1840s. Further advances in paper production and in printing machines (in particular the linotype) immensely cut time and capital costs in printing in the second part of the century (Cook 2006). These changes had their effects on the number of published titles.

Weedon analyzes England over the period 1836–1916 in more detail. He concludes that the time between 1846 and 1916 "saw a fourfold increase in production and a halving of book prices" (Weedon 2003, p. 57, see also Fig. 4).

What were the determinants of these important changes in book and print production in the nineteenth century? Weedon suggests that technological transformations were essential in the process. She emphasizes that paper, an essential part of the production costs of a book, could now be produced by the use of machines. The cost of paper fell by about two-thirds between 1866 and 1896, making book production much cheaper.

Even more spectacular increases were recorded for the newspaper industry. Cipolla (1969) highlights that "in the United Kingdom in 1831, the average monthly issue of the newspaper press amounted to about 3,240,000 copies (equivalent to about 137 copies per 1,000 inhabitants). By 1882 it had gone up to 135 million (about 3,700 copies per 1,000 inhabitants) (Cipolla 1969, p. 107)."[15] To put UK newspaper circulation in perspective, Murch (1870) states that about 38,648,000 copies were in circulation in 1831, which increased to 546,059,000 copies in 1864 (i.e., an increase of 1,313%). Not only were more newspapers printed, but the cost of newspaper production decreased. For example, the price of newsprint fell by a factor of ten between the 1860s and 1890s (Cook 2006). In other words, steam power was revolutionizing all forms of print media.

Industrial Demands for Education

The industrial demand for literacy was slower to develop. In the first phase of the Industrial Revolution, industrial demand for skilled labor was rather limited (Galor 2011). Some factory workers needed to be literate and numerate, but most processes were still handled by illiterate workers (Landes 1969, in Galor 2011). Furthermore, the developing factory system increased the opportunity costs of education for young children because they could be employed in factories (Cipolla 1980; Venezky 1996), leading to stagnating educational levels. The evidence to date finds little importance for the role of formal education in the Industrial Revolution in Britain (Mitch 1999).

However, Becker et al. (2011) identify an important role for education in Prussia's industrialization catch-up during the second half of the twentieth century. Furthermore, industrial modernization led to the need for skilled labor and created demand in the second phase of the Industrial Revolution (Galor and Weil 2000; Galor 2011). Certainly by the second half of the nineteenth century, industrialists identified a growing need for workers to be educated in order to effectively use the machines.

Human capital became more important in the process of production because it was complementary to (physical) capital and technological progress. Labor productivity was also higher for skilled individuals (Galor et al. 2009; Galor 2011). In addition, educated individuals were considered to be more easily adaptable to

[15]Numbers from France emphasize the stark increase in newspaper production: "[i]n 1840 the monthly issue of all the Paris journals totalled less than three million copies. By 1882 it was up to 44 million copies (Cipolla 1969, p. 107)."

technological change and open to new arriving ideas (Cipolla 1969). Technological progress was advancing quickly during this period, and thus the corresponding demand for human capital took off. Therefore, Zeira argues that "[b]oth the mechanization of production and the increase in the scale of production changed dramatically the whole character of producing and marketing. It now required new skills, of reading, writing, and arithmetic. The need to handle and operate machines and to take care of them required some knowledge in science and engineering and at least some literacy skills, to read manuals, to correspond with producers on problems in machines, etc." (Zeira 2009, p. 602).

Mitch (1992a) estimates that a literate worker gained about 13% more than an illiterate one around 1870. The increasing opportunities for members of the lower classes to enter into competition for higher-ranked jobs and more frequent opportunities to use literacy skills in everyday life no doubt stimulated private efforts to become educated. More generally, the structural changes in the sectors of the working population had its own repercussions on the requirement of literacy with a long-run move out of agriculture, where literacy was less needed (Table 1). Finally, the demographic transition further increased the need to invest in each child due to the lower number of children.

According to Galor et al. (2009), capitalists had an interest in promoting public education because they needed skilled workers in their factories during the second part of the Industrial Revolution. In contrast, large landowners preferred to block educational expansion because industrialization threatened their social status. In addition, workers would be more prone to migrate to better-paying urban centers and would be less willing to accept existing working conditions in the fields. Finally, landowners were often those who had to pay the taxes for improving the educational system, and they did not want to take this financial burden (see also Baten and Hippe 2017).

Thus, against the interests of landowners, capitalists were active in lobbying for public education and tried to influence government actions toward intervention in the education field. In particular, technical education was seen to be an important way of providing skilled workers. The example of England may illustrate this. It was the industrial leader, but late in the provision of public education. According to Galor (2011), the government changed its prior laissez-faire politics when it became clear that other countries (such as Germany or France) were becoming more inventive and innovative in industry. According to a jury member of the Paris exhibition of 1867,

Table 1 Usefulness of literacy in English male population, 1841–1891

Category of occupational usefulness of literacy	1841	1851	1871	1891
Literacy required	4.9	5.6	7.9	11.1
Literacy likely to be useful	22.5	22.8	25.3	26.1
Literacy possibly useful	25.7	24.2	24.5	25.9
Literacy unlikely to be useful	46.9	47.0	42.3	37.0

Source: Mitch (1992a)
Note: "Usefulness" refers to the degree that literacy was useful in a work context

the exhibition showed that there was insufficient progress in the English industrial sector. The reason for this "upon which there was most unanimity conviction is that France, Prussia, Austria, Belgium and Switzerland possess good systems of industrial education and that England possesses none" (Green 1990, p. 296, in Galor 2011, p. 476). Later on, the vice president of the committee of the Council of Education believed that "[u]pon the speedy provision of elementary education depends our industrial prosperity . . . if we leave our work-folk any longer unskilled . . . they will become overmatched in the competition of the world" (Hurt 1971, pp. 223–224, in Galor 2011, p. 477). Thus, Galor (2011) argues that the English government finally gave in to the demand of capitalists and expanded public education.

Together with the process of industrialization, the public became increasingly aware that more modern skills and knowledge related to modern production processes had to be acquired. Therefore, in their analysis of the USA between 1890 and 1970, Rubinson and Ralph (1984) find a significant impact of technological change on the expansion of schooling (i.e., enrollment rates). However, they acknowledge the fact that schooling and literacy were already important in the USA before the Industrial Revolution. They were part of a process of status competition among different parts of the population (according to ethnic origins, occupational status, religious affiliation, etc.). Moreover, the importance of technological change weakened over time because the schooling system was expanded more by political considerations.

The Incentives for the Nation-State to Provide Mass Schooling

Alongside military and industrial reasons, illiteracy also started to be seen as a disgrace for a nation (Cipolla 1969), putting pressure on governments that did not want to fall behind the "civilized" leaders. Therefore, international reputation and competition can be seen as factors that influenced government policy in this context. In addition, there was a growing belief in the role of mass education in forming the beliefs of the population and in directing and possibly controlling their behavior.

Ramirez and Ventraensca (1992, p. 49) consider the later phases of the construction of mass schooling to be a project of the nation-state, emphasizing its overall transnational character. They argue that mass schooling evolved and standardized around similar ideological and organizational forms. It allowed the nation-state to connect with individuals. The rise of the nation-state was, therefore, intrinsically related to mass schooling. Public life became increasingly reordered around the nation-state, and former transnational populations were transformed into citizens of nation-states. Education as such became its own institution, with its own goals, interests, and stakeholders. "Almost all European governments took steps which homogenized their populations: the adoption of a state religion, expulsion of minorities, institution of a national language, eventually the organization of mass public instruction" (Tilly and Tilly 1973, p. 44, cited in Alesina and Spolaore 2005, p. 184). Later on, former colonies proceeded in a similar fashion to achieve the goal of nation building in the second part of the twentieth century.

In line with this idea, Alesina and Reich (2013) argue that the magnitude of homogenization aimed at by a government depended on its particular regime. In particular, they distinguish between democratic regimes, "safe" nondemocratic regimes, and "unsafe" nondemocratic regimes. For example, in France the Ancien Régime was powerful but not interested in pursuing homogenization. The French Revolution toppled the regime, but other elites soon came to power. Still, the threat of more democracy necessitated increased efforts of nation building and homogenization. One important way to homogenize the population was linguistic homogenization by enforcing French as the only language used at school. The cases of Italy and England similarly illustrate that mass schooling was perceived to be a requirement for nation building. Similarly, in many African states after decolonization, an important share of enrollments was only realized after the (re)organization of a more or less effective national state. Increasing enrollments had different advantages. It had a signaling effect, emphasizing that the nation-building process was underway. Moreover, it was aimed at showing that the ideals of progress were embraced by the elites, giving further legitimacy and credibility to their ruling. Thus, the erection of schools also had an important symbolic effect in several dimensions (Ramirez and Ventraensca 1992).

Government Intervention in Education

How can the European countries be classified according to the importance of government involvement in education? Mitch (1992b) distinguishes between two categories of European countries. First, the countries that were the leaders in industrialization (e.g., England, France, Germany) showed a gradual increase of literacy over several centuries prior to the eve of mass schooling (see Fig. 4 for England).[16] This gradual process may have involved local demand, which was met by local supply. Second, the late-industrializing nations (such as Russia, Italy, Spain) had much lower literacy levels, but then these countries increased them much more rapidly. In fact, the governments were aiming at closing the gap with the leading European countries. Thus, the role of the central state was much more important in these countries, and local elites had a less decisive influence in advancing educational levels.

Using a different approach, Soysal and Strang (1989) divide the European countries into three groups according to the conflict and competition within societies on the subject of education. The first category, the "statist construction of education," includes Prussia and Scandinavian countries (Denmark, Norway, Sweden). In these cases, the state was able to build upon the know-how and physical infrastructure of the national churches. In fact, instruction was mostly provided by churches before the nineteenth century (Vincent 2000). Thus, the alliance between the state and the church was crucial.

[16]The same can be said of the evolution in numeracy (A'Hearn et al. 2009).

Table 2 Introduction of compulsory education and primary enrollment ratios in 1870

Country	Introduction of compulsory education	Primary enrollment ratios in 1870
Prussia	1763	67
Denmark	1814	58
Greece	1834	20
Spain	1838	42
Sweden	1842	71
Portugal	1844	13
Norway	1848	61
Austria	1864	40
Switzerland	1874	74
Italy	1877	29
United Kingdom	1880	49
France	1882	75
Ireland	1892	38
Netherlands	1900	59
Luxembourg	1912	–
Belgium	1914	62
USA	–	72

Source: Soysal and Strang (1989)

Second, the "societal construction of education" took place in countries such as France, the Netherlands, Switzerland, Great Britain, and the USA. In these countries there were many conflicting interests articulated by important stakeholder groups. These were of a religious order (e.g., Great Britain and the Netherlands) or of local character (e.g., the USA). Therefore, a centralized, nationwide educational system was only constructed relatively late, although schooling had already been expanding quite significantly before the involvement of the state.

Finally, the "rhetorical construction of education" in countries such as Portugal, Italy, Greece, and Spain meant that compulsory education laws were passed relatively early, but actual action did not take place for many decades. The dates of the introduction of compulsory education and enrollment ratios in 1870 illustrate this discrepancy (Table 2). For example, although Greece introduced compulsory schooling quite early in 1834, its insufficient implementation led to primary enrollment ratios of only 20% in 1870. In this and the other countries that had a "rhetorical construction of education," there was no competition over education, and enrolment rates were also relatively low. In addition, the state was too weak and had too little power to enforce its own legislation in its territory.

Another question concerns the effectiveness of state policy in education. When was state action effective? Mitch (1992b) points out that a state's education policy might not have been very effective if demand was low. Even if a sufficient supply of education was guaranteed by the state, lacking interest or even resistance from parents meant that classrooms stayed half empty. Furthermore, state policy was

rather ineffective when there had already been a high level of literacy and a large supply of private schools. In this case, Mitch (1992b) argues that public policy initiatives were redundant, and in many cases already existing private schools were only turned into public ones. In contrast, a state's policy may have been quite effective when a clear shortage of supply did not meet popular demand. In addition, state policy could overcome local resistance where local elites blocked the provision of educational facilities against the will of the local population. Moreover, public policy could be effective when an initial level of instruction was present, but this level had to be advanced still further. Here, the state's policy (in particular in England, France, and Germany) was directed toward the construction of new schools, the standardization and improvement of teaching, the reduction or elimination of schooling fees, and the establishment of compulsory schooling laws. National literacy campaigns (e.g., in Scandinavia in the seventeenth century) may have also been an important tool to increase the effectiveness of education policy.

In consequence, Mitch emphasizes that an effective educational policy needed the concordance between the demand for instruction and the influence of policy measures. Both are dependent on the popular attitude toward literacy and the power relations within society.[17] Thus, education policy needed to count on popular and local elite support to be effective. Where one or both of those factors were lacking, national public policy was often without great effect. Therefore, local institutions and local elites played an important role. Accordingly, Lindert (2004) notes that many of the early high performers in mass education had decentralized education systems. For example, he argues that the Prussian education system was essentially based on bottom-up structures.[18] Other examples for decentralized education systems include the USA and Canada. In all of these countries, local governments were able to levy their own local taxes that financed the school system. Thus, local actors were able to choose whether they wanted to be committed to schooling or not and were freed of interests that may have been existing at the national level. Local leaders were thus influenced by local debates and local demand. Thus, local funding also meant that regional differences in education could become even more striking. In consequence, whereas in the first phase decentralization enabled educational levels to increase in the

[17]He distinguishes three cases: egalitarian, elite, and autocratic forms of power distribution. In the egalitarian case, the preferences of the majority were implemented by the elites. Thus, there was a high risk that policy measures were simply redundant or not meeting private demand. In these societies, the acquisition of education meant the prospect of moving up the social ladder, leading to popular demand. When the powers in society were more concentrated but upward mobility was still possible (e.g., in France and Germany), the effectiveness of public educational policy was probably higher. Finally, in the case of more autocratic forms of government, when power was extremely concentrated (such as in Spain and Portugal) and the masses lived at low standards of living, public action was ineffective. On the one hand, demand was low and actions by the state were perceived as intruding into family life. On the other hand, local elites blocked educational changes that may have been intended at the national level (Mitch 1992b).

[18]Overall, the Prussian kings were not fervent promoters of the spread of mass education. They "did as much, and said as much, to *block* schooling and free thought as to spread it" (Lindert 2004, p. 118).

most favorable regions, in the second phase, it tended to fail to increase educational standards in the laggard regions. In the latter phase, the government was needed to avoid such a market failure[19] and to centrally enforce higher educational spending.[20]

In what specific ways can the state act in order to avoid market failures in the education sector? First, the state can provide sufficient supply. In theory, the private sector could also be fully responsible for the supply side: according to Lindert (2014), investments in primary education have brought high private and social returns for at least the last 600 years. However, he points out that the private sector has never solved the "capital constraint" problem, i.e., that universal education investments are to be paid back by the increased future earnings of children. For this reason, the state has to step in to avoid underinvestment in education. Yet this supply of schools, teachers, and materials generates important costs for the government. This is one reason why the state was not able to finance public education throughout most of history. In fact, it did not have the necessary fiscal capacity due to inefficient state bureaucracies, etc. On the other hand, the willingness to invest in education, and thus to increase taxes, was dependent on the interests of the ruling classes. As long as these ruling classes, in particular landowners, were opposed to educational reforms, progress was limited or inexistent in this area (see also Baten and Hippe 2017). Therefore, Lindert (2014) argues that the provision of voting rights to larger segments of the population made a crucial change to the composition and attitude of governments toward educational investment. Thus, government spending on education has risen dramatically during the last two centuries. Higher spending does not necessarily translate into higher enrollments but is often well correlated (see Fuller 1983; Walters and O'Connell 1988).

Second, the state can create demand for education. Thus, investing in the quality and credibility of schools may have an important impact on demand and enrollment. The quality of schools depends on several factors, among other things on the quality of teachers. The supply of those teachers depends on their salaries and the social status of the teaching profession (Cipolla 1969).[21] Assuring the quality of teachers

[19]Here, it is assumed that all regions benefit socially from higher education levels.

[20]Some studies have suggested the existence of a human capital Kuznets curve, adapting the idea of Kuznets (1955) to education. In other words, increases in human capital inequality in earlier phases are followed by subsequent reductions in human capital inequality in later phases of economic development. For empirical contributions, see, e.g., De Gregorio and Lee (2002), Castello and Domenech (2002), Lim and Tang (2008), and Morrisson and Murtin (2013); for theoretical models see, e.g., Galor and Tsiddon (1996), Glomm and Ravikumar (1998), and Matsuo and Tomoda (2012).

[21]The level of salaries was quite diverse across Europe. However, for the most part, they seem to have been low. Thus, Cipolla (1969) suggests that the average salary of a teacher was comparable to that of a craftsman before the nineteenth century. In addition, the social status of teachers still varied importantly from one European country to the other in the nineteenth century. For example, teachers enjoyed high public respect in Germany (and Prussia in particular) (Cipolla 1969). Social prestige and income may thus be causes of high quality. Perhaps a long-standing tradition of the teaching profession was also an important factor. In contrast, schoolmasters and mistresses did not have a good reputation and did not have much prestige in England and southern Italy in the nineteenth century.

can thus be an important policy objective. In addition, the state can politically construct the organization of work. For example, it has the power to create minimum educational standards for the entry into jobs in the public and private sectors. If parents or their children want to enter this well-paid job market, they have to invest in education. In consequence, the state can create a public demand in a politically directed way. Likewise, the state can decrease the opportunity costs for sending children to school (e.g., by limiting child labor and giving schools a normative legitimization).

Finally, the ideological and symbolic compound associated with education (e.g., economic and social opportunity, but also Western ideals such as nation building, individual development, and progress) can be used by the state to advance mass schooling through corresponding signaling activities (such as literacy campaigns) (Fuller and Rubinson 1992).

However, it appears that the state's involvement to improve educational quality and quantity might have been too late in some countries. For example, Mitch (1986) indicates that this may have been the case in nineteenth-century England. In consequence, was the UK spending too little on education (and too late)? To have a better grasp of the amount of public spending in comparison with economic output, public investment in education can be directly measured as a share of GDP. Carpentier (2007) shows the relative position of the UK with other countries that were catching up. In particular, he considers the differences between the UK, France, and the USA (Fig. 6). The very low values of the UK until the end of the nineteenth century also appear to be quite low in comparison to these other industrializing countries. This illustrates the claim that the major deficiency of the UK in the second part of the Industrial Revolution was its lack of skilled workers due to a lack of education. The USA was by far the leading country among these three nations in 1870, spending almost 1% of its GDP on education. France follows with a large gap at about 0.5%, while the UK lags far behind. The UK became aware of this later on and the government increased its expenditures much more pronouncedly. France also increased its relative expenditure on education, for a decade becoming the leader of these three countries. Subsequently, the UK had the highest shares until First World War. However, the real takeoff in education spending shares occurred only between the 1950s and the 1970s. Since then, education spending has mostly been between 4% and 6% of GDP in all three countries.

These data illustrate the evolution of public expenditure on education. However, they do not include private investment in education. The overall investment in education might have been much higher than the public investment alone. Lindert (2004) provides some estimates on total (i.e., public and private) expenditure on all levels of education in a number of countries between 1850 and 1910 (see Fig. 7).

The data suggest that overall investment in education was highest in Germany throughout the second part of the nineteenth century, followed by the USA. France appears to have had much lower levels, and the gap to the leading countries

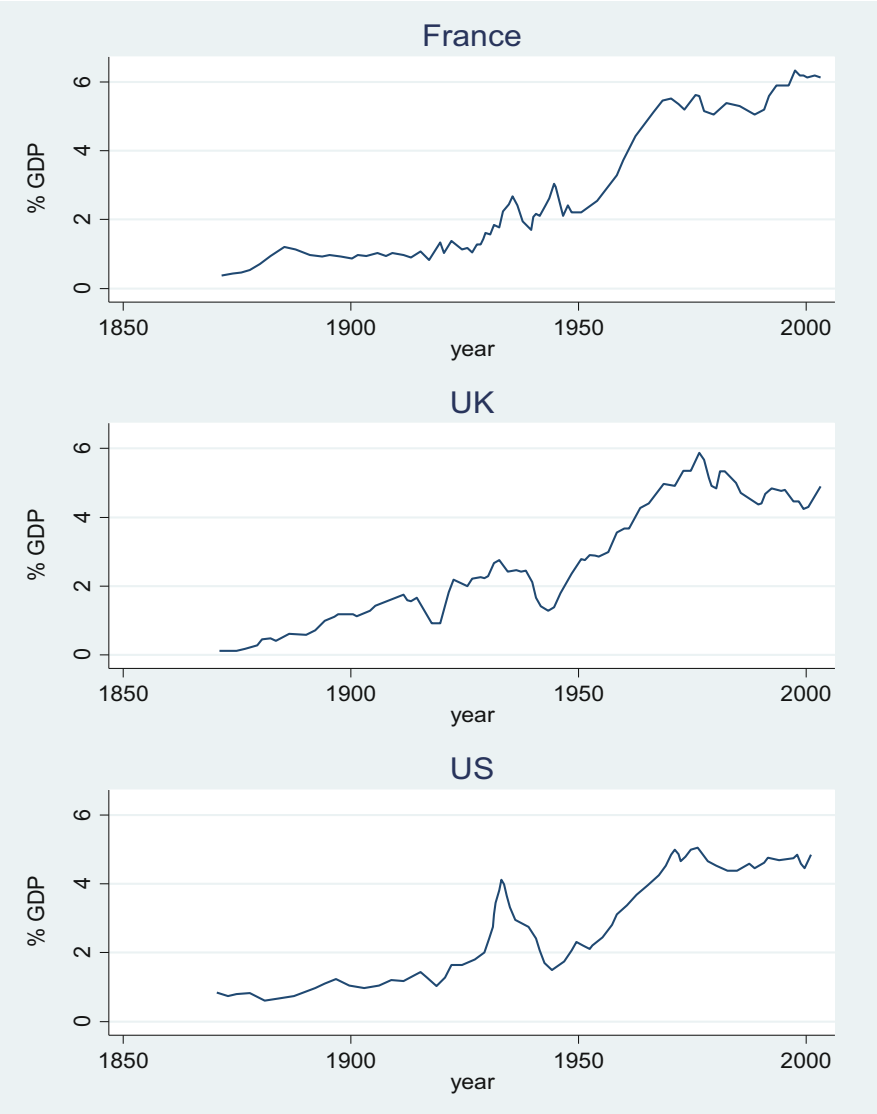


Fig. 6 Public expenditure on education as a share of GDP in the UK, France, and the USA, 1870–2003. (Source: Carpentier (2007); see also Diebolt (2000) and Diebolt and Fontvieille (2001))

widened at the end of the 19th and the beginning of the twentieth century. Data on Italy and the UK are even more tentative but could suggest a more intermediary position.

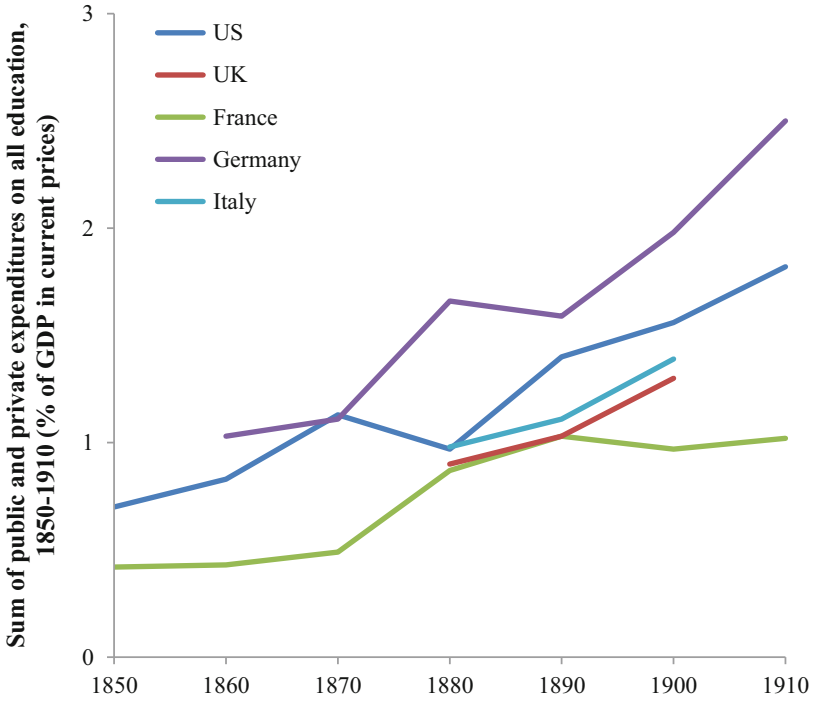


Fig. 7 Total of public and private expenditure as a share of GDP, 1850–1910. (Source: Derived from Lindert (2004))

The next logical step is to ask about the composition of this investment. Was private investment in education higher or lower than public expenditure? Thus, did public investment replace private investment over time? The next figure illustrates the relative importance of public and private investment (as a share of GDP) in these countries (see Lindert 2004 for details) (Fig. 8).

Three different patterns emerge. First, private and public expenditures were more or less at the same level in the USA in the 1850s. However, public expenditures as a share of GDP rise (almost) continuously, whereas private expenditures stagnate. The initially slightly higher levels for private expenditures might possibly suggest that public expenditures partly replace private ones. Second, public expenditures are much higher than private ones in both France and Germany throughout the period. The gap also widened, given the increased efforts of the government to provide schooling. Finally, private expenditures appear to have been higher in Italy and the UK at the end of the nineteenth century. Therefore, the tentative evidence would suggest that government spending became the major form of investment in education only in the twentieth century, much later than in other countries.

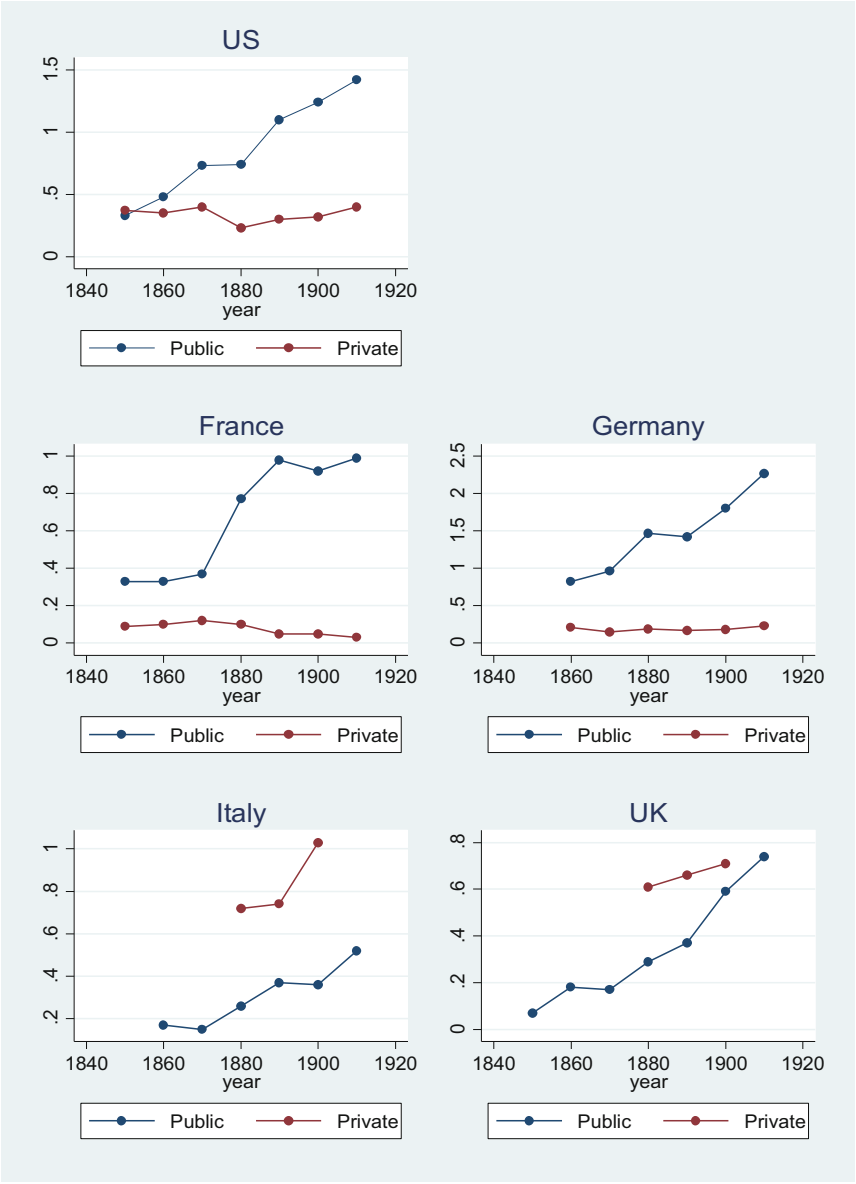


Fig. 8 Public and private expenditure as a share of GDP, 1850–1910. (Source: Derived from Lindert (2004). Note: Vertical axis refers to expenditures on all levels of education as a % share of GDP in current prices. The value for private expenditure for the UK in 1890 has been interpolated)

The Worldwide Human Capital Transition and the State

Trends in Worldwide Human Capital Levels During the Last Two Centuries

It appears important at this stage to provide a general overview of the long-run human capital transition in the world before moving on to the role of the state. We may begin with numeracy levels (Crayen and Baten 2010). Numeracy is measured here by the age heaping method, which measures the share of people who are able to count.[22] The industrial countries have the highest numeracy levels of all countries at the beginning of the nineteenth century. These countries are relatively closely followed by East Asia, which has similarly high numeracy levels as the industrial countries in the 1880s. Not far behind East Asia are the Eastern European/Central Asian countries, which achieve the maximum numeracy values by the 1900s. Thus, these three regions reach the upper limit of numeracy levels at the end of the nineteenth century. No other world regions were able to achieve such a level until the middle of the twentieth century. The most advanced area of these follower regions is Latin America/Caribbean, which has been inspired by the educational policies of the industrialized countries. The lagging world regions are Southeast Asia, Sub-Saharan Africa, and the Middle East/North Africa. South Asian countries had the lowest level of numeracy throughout the period.

Numeracy is just one indicator of human capital levels. Literacy is surely the most common variable for long-run education (see also Diebolt and Hippe 2017, 2018a, b; Hippe and Perrin 2017; Diebolt et al. 2018). The overall trends since 1870 (Fig. 9) are quite similar in many respects to the ones identified in numeracy. In particular, literacy levels are once again led by the West. The gap between the West and all other countries appears to be much more important than in numeracy. One important reason for this is that literacy levels are much lower than basic numeracy at the beginning of the period. This circumstance is related to the underlying concept of measurement: the numeracy indicator proxies very simple counting skills, which are even more basic than literacy skills. In any case, the West is again followed by East Asian countries (i.e., Korea-Japan and China) and Eastern Europe. Eastern European countries show a particular sharp increase in literacy between 1910 and 1950. In contrast to Korea-Japan, China does not substantially increase its literacy levels until the second half of the twentieth century and thus falls back. Latin America takes an overall similar evolution as China. Although its literacy levels speed up between 1910 and 1950, progress is more slowly achieved afterward. In China, the reverse is

[22]It uses a particular heaping phenomenon in the age distribution of censuses and other comparable data to calculate numeracy levels. More specifically, individuals over-reported ages ending in 0 and 5 to census takers, leading to significant spikes in the age distribution. The reason for this was that they were not able to count and to know their exact age, so that they rounded it. This is a well-known phenomenon that can be found in historical sources and in a range of developing countries today. For further information, see A'Hearn et al. (2009), Hippe and Baten (2012), Hippe (2012, 2013b, 2014).

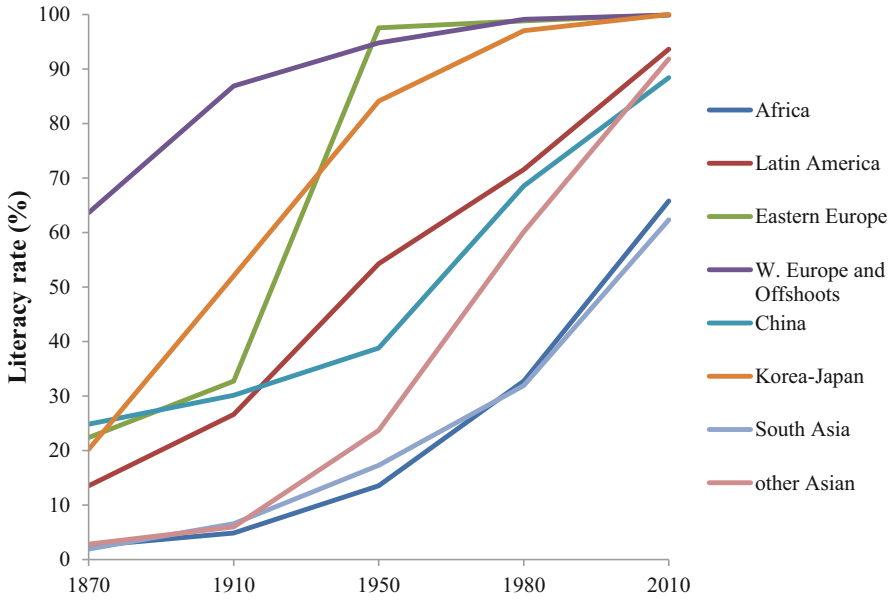


Fig. 9 Evolution of literacy in world regions, 1870–2010. (Source: Morrisson and Murtin (2013))

the case. Societies characterized by almost complete illiteracy in the 1870s come from South Asia, "other Asian" countries, and Africa. Whereas the "other Asian" countries are able to close the gap with the leading regions by 2010, South Asia and Africa show important, but not sufficient, progress. However, it seems that they may be able to join all other regions in the decades to come.

Taking another standard indicator, the years of primary education, the conclusions are quite similar (see Morrisson and Murtin 2013). Clearly, this variable is even more closely related to the state because most schooling has been provided by the state during the considered time span. The increases are not as impressive as in literacy in many cases, but this is partly due to the nature of the indicator. Still, the trends that were discernible in the literacy data can also be found here. Primary years of schooling amounted to about 3 years in the West in 1870, whereas it ranged from 0 to 1 years in all other regions. In 2010, three different regimes exist: about 6 years of primary education in the West, Korea-Japan, and Eastern Europe; about 5 years in Latin America, China, and other Asian countries; and, finally, about 3.5 years in South Asia and Africa. Except for the leading group, all other regions appear to be converging to the leaders in the longer run.

Primary education is a measure of basic human capital. Thus, it is similar to literacy and numeracy. If we want to enlarge the concept of human capital to include more advanced skills, we should also take into account the role of secondary education. Therefore, the total years of schooling indicator give a fuller comprehension of basic and more advanced human capital levels (Fig. 10). However, it is clear

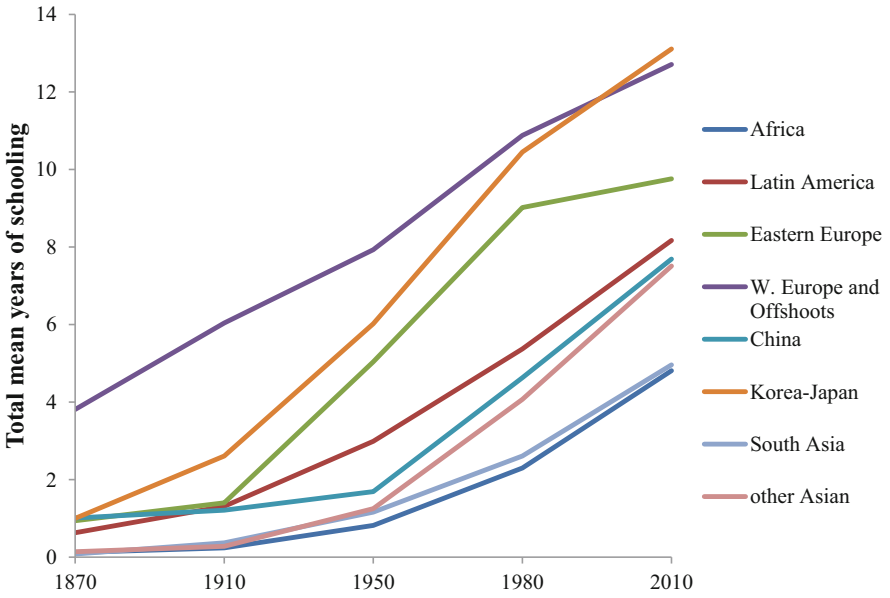


Fig. 10 Evolution of total years of schooling in world regions, 1870–2010. (Source: Morrisson and Murtin (2013))

that an important part of this variable is constituted by the levels of primary education, as shown in the previous figure. Most of the trends in the average total years of schooling are, therefore, in line with the results for primary education. The most evident change in the pattern is the evolution of Eastern Europe. In primary education, Eastern Europe was already at similar levels as the West and Korea-Japan by 1980. In primary and secondary education, there still remained an important gap of about 1.5–2 years. In addition, this gap was not closed, but actually widened until 2010. This change was probably motivated by the fall of Communism and the subsequent transition period in these countries.

One can also look at these data in another way. The notion of the technological frontier has been very important in the growth literature (e.g., Vandenbussche et al. 2006). The same idea can be applied to human capital. If we consider the leader of human capital (i.e., the West) as marking the frontier of human capital, we can calculate the simple distance of other regions to this frontier (using total years of schooling). Taking this perspective, it becomes clear that the gap between the West and "the rest" remained rather stagnant and very wide between 1870 and 1910. Only Korea and Japan exhibit significant progress. Afterward, all countries exhibit different rates of convergence to the human capital frontier, but (apart from Korea-Japan) all countries are still quite far away.

Thus, all the indicators employed in this analysis illustrate a move from relatively low human capital levels to much higher levels during the last 200 years. In schooling, this transition has taken place in all regions between

1870 and 2010. The West already had relatively high schooling levels in the second part of the nineteenth century (so that one needs to go back to the early nineteenth century to see the entire trend), but these levels are still very low compared to contemporary current levels. The human capital transition is a phenomenon of the last 200 years, and the state has been playing an important role in this transition.

Human Capital and the State

The involvement of the state in the transition from low mass levels of education to high levels becomes evident when considering different indicators. Passing compulsory schooling rules is one clear sign that a state intends – at least symbolically – to improve its educational levels. Figure 11 shows the evolution of the cumulative share of countries in a world region that have passed compulsory schooling rules. Western Europe appears to be the leader in state laws on education, achieving a rate

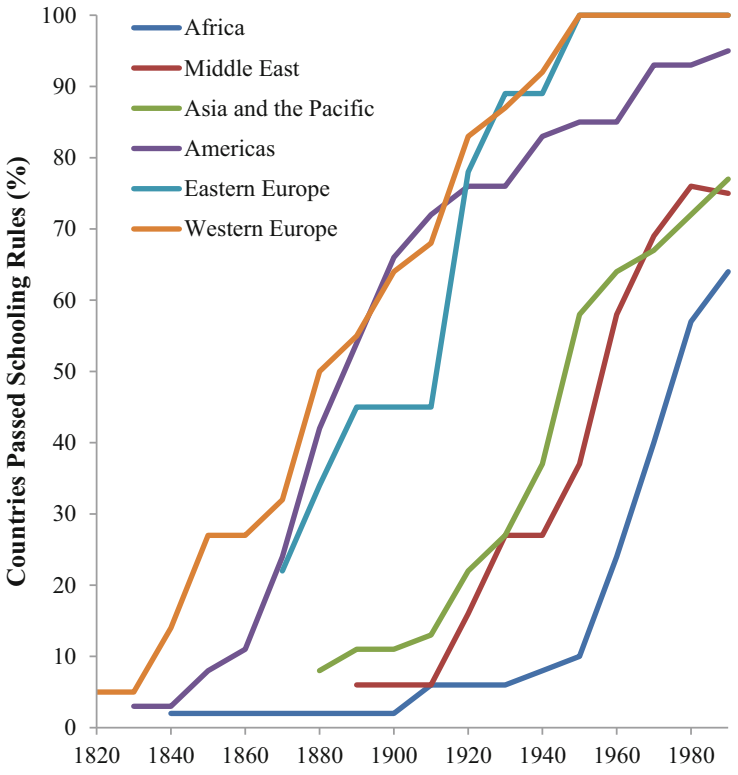


Fig. 11 Proportion of countries having passed a compulsory schooling rule, by region and decade, 1820–1990. (Source: Ramirez and Ventraensca (1992))

of around 50% by 1880. It is followed by the Americas and Eastern Europe, both culturally linked to the leaders. Asia and the Pacific were the next to follow slowly. The Middle East began its major involvement in state education even later, from the 1910s onward. Both regions show a parallel spectacular growth between the 1930s and the 1950s, with a lead of one decade for Asia and the Pacific region. Progress was less striking from then onward. In Africa, the turning point came in the 1950s and the period of decolonization. Afterward, similar levels as in the former two regions were almost reached. Accordingly, Ramirez and Boli-Bennett (1982) argue that there is an acceleration phenomenon, i.e., there is a correlation between the date of independence of a country and the time to pass compulsory schooling: the younger a country is, the faster it passes compulsory schooling laws (Adick 2003). Overall, the transition from almost non-existing official legal involvement in compulsory schooling to almost universal compulsory schooling in many parts of the world is very clear.

We can also consider another indicator related to state involvement. It is the cumulative share of countries within a world region that have created a central education authority (Ramirez and Ventraensca 1992). In some respects this indicator shows similar trends. Western Europe leads in this area until the 1870s. The share of countries with central education authorities then dramatically increases over the next century. Thus, all major world regions had almost universally central education authorities by the 1990s – except for the former leader Western Europe, where compulsory schooling laws and centralized state control have been contested for the last 200 years. For this reason, still in the 1990s, centralized state control was not implemented in a sizable share of Western European countries. In contrast, the idea of a central education authority has been deemed to be necessary in the other world regions that followed Western Europe. Less resistance and more conviction of the necessity of central state policy has characterized experiences in these continents. It highlights once again the connection between the (in part newly created) nation-states and mass schooling in these countries.

Conclusion

The expansion of the "knowledge economy" offers a potential way to place the global economy on a more sustainable trajectory. With this in mind, our aim in this chapter was to better understand the market and institutional forces that led to the human capital transition in present-day industrialized nations – that is, the development of a key pillar of the "knowledge economy."

To this end, we sought to answer the following questions: What factors have driven the demand for and supply of human capital over the last 1000 years in Europe? Has there been a market failure in the formation of human capital? What role has the state and other institutions played in the provision of schooling? Why

were government and the public willing to invest heavily in this public good? To better understand the forces that led to the human capital transition, we brought together a wealth of evidence on the history of education, and broader human capital formation, across Europe.

One of the central features of the history of the European human capital transition has been the decline in the price of written education, in the fifteenth century following the development of the Gutenberg Press and in the nineteenth century with its mechanization. This incentivized the development of written education and implied that traditional work skills were complemented by a broader education in literacy and numeracy and the ensuing improvements.

However, this narrative also seeks to emphasize, as Mitch (1992b) did, *the fundamental role demand played in driving the early human capital transition.* Early demands in transitions are often driven by "luxury" services, where some consumers have a relatively high willingness to pay, creating niche markets (Fouquet 2008). In the case of education, this was associated with the demand for spiritual guidance. The initial demand to read the Bible (either in Latin or, for most Europeans, in their mother tongue) was stimulated by Luther and other champions of the Reformation. In other words, capitalizing on a technological revolution, one persuasive voice (accompanied by a growing number of followers) can initiate a transformation in beliefs that stimulates an initial (immaterial) demand for one of the key sources of modern economic growth.

In time, the demand for written education stemmed from many additional sources, associated with commercial, military, industrial, parental, status, and "belief-formation" factors. So, the first phase of growth in human capital levels (particularly in literacy) in the leading countries can be explained by the interplay of supply and demand. This provided a basic level of education to upper and middle classes. Mirroring this experience, the declining costs of using ICT are enabling developing countries to accelerate investment in human capital and seek to catch up with the industrialized nations. However, if human capital transitions in developing economies unfold in a similar way to the transitions in Europe, then certain segments of the population will be left behind. That is, while probably upper and middle classes in developing economies become highly educated, a poorer segment will struggle to raise their levels of human capital. *Thus, without major efforts from government, educational and ultimately income inequality is likely to increase.*

Indeed, the second phase of the human capital transition (i.e., the spread of mass literacy and the construction of mass schooling beginning in the nineteenth century) was particularly driven by the state. In certain cases, *shocks (*e.g.*, wars, the loss of international competitiveness) played a key role in altering beliefs about the value of education and stimulating public investment in human capital.* For instance, Prussia's move to embrace education more forcefully was triggered by defeat in the Napoleonic wars. Similarly, industrialists in the UK at the end of the nineteenth century started to realize that their ability to keep up with foreign competitors, particularly Germany and the USA, depended on having a well-educated workforce from which to hire.

However, in general, the role of the state was more important in those countries that had to catch up with the leaders. Particularly in lagging economies in Europe, there was a market failure – in the sense that although the social and economic benefits of higher human capital levels would have been significant, education was not provided by the private sector. The construction of state fiscal capacities (including the building up of efficient bureaucracies) and falling resistance from social elites against public educational investments were often important prerequisites for state intervention. Once the state got involved in the education sector and created a mass public schooling system, this market failure was at least partly resolved over time. Naturally, recognition of the importance of the human capital investment needed to be followed up with effective implementation strategies and public financial support (often around 2–5% of GDP).

One reason the implementation and financial support for mass education often followed swiftly behind recognition was because the state, along with religious institutions and other nongovernmental organizations, placed a value on "belief-formation" and the control of the provision of education. The desire for nation building implied that instilling values and cultural attributes (e.g., language), which would encourage future generations to identify with the state, was an important driver of public investment in education (Pritchett 2003). In other words, this secondary benefit of education (to government) may well have been pivotal to the second phase of the human capital transition.

When considering potential future transitions, it is important to consider the benefits that government may gain from the transition. *Identifying the existence of a principal-agent problem and finding direct benefits for government might be a powerful means to incentivize government officials to develop effective implementation strategies and find public financial support.*

With this objective in mind, it is important to remember that major transitions are typically long-term processes. The human capital transition (here, defined as an increase from 10% to 90% literacy rate) in Europe took roughly 400 years, from Luther's printed translation of the Bible into a European vernacular language until the First World War. In Japan and Korea, it took a little over 100 years (from about 1850 to 1960). In South Asia and Africa, this is likely to take around 100 years. *Thus, even in countries implementing basic education programs today, there is a limit to how fast a transition can be achieved.* This is partly because new generations need to be educated, and this takes decades.

Yet even in the case of education, vested interests may inhibit a successful transition. *Therefore, these vested interests have to be identified and considered in any policy.* Indeed, the central state interacts with a range of stakeholders in educational policy. In particular, there are religious authorities (in Europe the churches), elites (landowners and capitalists), parents, ideological movements (including liberals and conservatives), and the local authorities. Any of these can alter the rate and nature of the transition. On the other hand, if government has direct benefits from a transition, then these vested interests are likely to be more swiftly overcome.

Cross-References

- ▶ Age-Heaping-Based Human Capital Estimates
- ▶ Cliometrics of Growth
- ▶ Economic-Demographic Interactions in the European Long Run Growth
- ▶ Education and Socioeconomic Development During the Industrialization
- ▶ GDP and Convergence in Modern Times
- ▶ Human Capital
- ▶ Innovation in Historical Perspective
- ▶ Institutions
- ▶ Preindustrial Economic Growth, ca. 1270–1820
- ▶ The Industrial Revolution: A Cliometric Perspective

Disclaimer The views expressed are purely those of the writers and may not in any circumstances be regarded as stating an official position of the European Commission.

References

A'Hearn B, Crayen D, Baten J (2009) Quantifying quantitative literacy: age heaping and the history of human capital. J Econ Hist, Cambridge University Press 68(3):783–808

Adick C (2003) Globale Trends weltweiter Schulentwicklung: Empirische Befunde und theoretische Erklärungen. Z Erzieh 6(2):173–187

Aghion P, Howitt P (1992) A model of growth through creative destruction. Econometrica 60:323–351

Aghion P, Howitt P (1998) Endogenous growth theory. MIT Press, Cambridge

Alesina A, Reich B (2013) Nation building, NBER working paper 18839, Feb 2013

Alesina A, Spolaore E (2005) The size of nations. MIT Press, Cambridge

Anderson CA (1965) Literacy and schooling on the development threshold: some historical cases. In: Anderson CA, Bowman MJ (eds) Education and economic development. Aldine, Chicago, pp 347–362

Ang JB, Madsen JB (2011) Can second-generation endogenous growth models explain the productivity trends and knowledge production in the asian miracle economies? Rev Econ Stat 93(4):1360–1373

Archer M (1979) Social origins of educational system. Sage, Beverly Hills

Barro RJ (1991) Economic growth in a cross section of countries. Q J Econ 106(2):407–443

Barro RJ, Lee JW (1993) International comparisons of educational attainment. J Monet Econ 32:363–394

Baten J, Hippe R (2017) Geography, land inequality and regional numeracy in Europe in historical perspective. J Econ Growth 23(1):79–109

Becker GS (1964) Human capital: a theoretical and empirical analysis with special reference to education, 1st edn. University of Chicago Press, Chicago

Becker GS (1981) A treatise on the family. Harvard University Press, Cambridge

Becker GS (1993) Human capital: a theoretical and empirical analysis with special reference to education, 3rd edn. University of Chicago Press, Chicago

Becker SO, Woessmann L (2008) Luther and the girls: religious denomination and the female education gap in nineteenth-century Prussia. Scand J Econ 110(4):777–805

Becker SO, Woessmann L (2009) Was weber wrong? A human capital theory of protestant economic history. Q J Econ 124(2):531–596

Becker SO, Hornung E, Woessmann L (2011) Education and catch-up in the industrial revolution. Am Econ J Macroecon 3(3):92–126
Benhabib J, Spiegel MM (1994) The role of human capital in economic development: evidence from aggregate cross-country and regional U.S. data. J Monet Econ 34(2):143–173
Biller P, Hudson A (eds) (1996) Heresy and literacy, 1000–1530. Cambridge University Press, Cambridge
Boli J (1992) Institutions, citizenship, and schooling in Sweden. In: Fuller B, Rubinson R (eds) The political construction of education. The State, State Expansion and Economic Change, Praeger, New York
Bolt J, van Zanden JL (2014) The maddison project: collaborative research on historical national accounts. Econ Hist Rev 67(3):627–651
Bowles S, Gintis H (1976) Schooling in capitalist America: educational reform and the contradictions of economic life. Basic Books, New York
Brewer DJ, Hentschke GC, Eide ER (2010) Theoretical concepts in the economics of education. In: Brewer DJ, McEwan PJ (eds) Economics of education. Elsevier, Oxford
Briggs J (1978) An italian passage. Yale University Press, New Haven
Buringh E, van Zanden JL (2009) Charting the "Rise of the west": manuscripts and printed books in Europe, a long-term perspective from the sixth through eighteenth centuries. J Econ Hist 69(2):409–445
Campbell R (1747) The London Tradesman, being a compendious view of all the trades, professions, arts, both liberal and mechanic, now practised in the cities of London and Westminster. T. Gardner, London
Carnoy M, Levin H (1985) Schooling and work in the democratic state. Stanford University Press, Stanford
Carpentier V (2007) Education Policymaking: economic and social progress. mimeo, London
Castello A, Domenech R (2002) Human capital inequality and economic growth: some new evidence. Econ J 112:187–200
Chappell W (1970) A short history of the printed word. Nonpareil Books, Boston
Chassant LA (1846) Dictionnaire des abréviations latines et françaises usitées dans les inscriptions lapidaires et métalliques, les manuscrits et les chartes du Moyen Âge. Cornemillot, Evreux
Chaves M, Gorski PS (2001) Religious pluralism and religious participation. Annu Rev Sociol 27:261–281
Ciccone A, Papaioannou E (2009) Human capital, the structure of production, and growth. Rev Econ Stat 91(1):66–82
Cipolla CM (1969) Literacy and development in the West. Penguin Books, Baltimore
Cipolla CM (1980) Before the industrial revolution: European society and economy, 1000–1700. Norton, New York
Clark G (2004) Lifestyles of the rich and famous: living costs of the rich versus the poor in England, 1209–1869. In: Paper presented in conference "Towards a global history of prices and wages". Available online at http://www.iisg.nl/hpw/papers/clark.pdf
Clark G (2007) A farewell to alms. Princeton University Press, Princeton
Clark G, Levin P (2001) How different was the industrial revolution? The revolution in printing, 1350–1869, Working paper. University of California, Davis
Cohen D, Soto M (2007) Growth and human capital: good data, good results. J Econ Growth 12:51–76
Cohn E, Geske TG (1990) The economics of education. Pergamon Press, New York
Collins A, Halverson R (2010) The second educational revolution: rethinking education in the age of technology. J Comput Assist Learn 26:18–27
Cook SDN (2006) Technological revolutions and the Gutenberg myth. In: Hassan R, Thomas J (eds) The new media theory reader. McGraw-Hill, Berkshire (originally published in Internet Dreams (1997), MIT Press, Cambridge, MA)
Corazzini AJ (1967) When should vocational education begin? J Hum Resour 2:41–50

Council of the European Union (2009) Council conclusions of 12 May 2009 on a strategic framework for European cooperation in education and training ('ET 2020'), C119/2, online. Last accessed on 22 Nov 2017. http://www.cedefop.europa.eu/EN/Files/ET_2020.pdf
Crayen D, Baten J (2010) Global trends in numeracy 1820–1949 and its implications for long-run growth. Explor Econ Hist 47:82–99
Cressy D (1980) Literacy and the social order: reading and writing in Tudor and Stuart England. Cambridge University Press, Cambridge, UK
Cuijpers PMH (1998) Teksten als koopwaar: vroege drukkers verkennen de markt: een kwantitatieve analyse van de productie van Nederlandstalige boeken (tot circa 1550) en de 'lezershulp' in de seculiere prozateksten. De Graaf, Nieuwkoop
De Gregorio J, Lee JW (2002) Education and income inequality: new evidence from cross-country data. Rev Income Wealth 51:395–416
De La Fuente A, Doménech R (2006) Human capital in growth regressions: how much difference does data quality make? J Eur Econ Assoc 4:1–36
Demeulemeester JL, Diebolt C (2011) Education and growth: what links for which policy? Hist Soc Res 36(4):323–346
Diebolt C (2000) Die Erfassung der Bildungsinvestitionen im 19. und 20. Jahrhundert. Z Erzieh 3(4):517–538
Diebolt C, Fontvieille L (2001) Dynamic forces in educational development: a long-run comparative view of France and Germany in the 19th and 20th centuries. Compare 31(3):295–309
Diebolt C, Hippe R (2017) Regional human capital inequality in Europe, 1850–2010. Région et Développement 45:5–30
Diebolt C, Hippe R (2018a) Remoteness equals backwardness? Human capital and market access in the European regions: insights from the long run. Educ Econ 26(3):285–304
Diebolt C, Hippe R (2018b) The long-run impact of human capital on innovation and economic development in the regions of Europe. Appl Econ. https://doi.org/10.1080/00036846.2018.1495820
Diebolt C, Hippe R, Jaoul-Grammare M (2018) Bildungsökonomie. Eine Einführung aus historischer Perspektive [Education Economics. An introduction in historical perspective; in German]. Springer, Wiesbaden
Dittmar J (2013) New media, firms, ideas, and growth: European cities after Gutenberg. National Bureau of Economic Research, Cambridge, MA
Ekelund RB, Hébert RF, Tollison RD (2002) An economic analysis of the protestant reformation. J Polit Econ 110(3):646–671
Ekelund RB, Hébert RF, Tollison RD (2004) The economics of the counter-reformation: incumbent-firm reaction to market entry. Econ Inq 42(4):690–705
Engelsing R (1973) Analphabetentum und Lektüre: Zur Sozialgeschichte des Lesens in Deutschland zwischen feudaler und industrieller Gesellschaft. Metzler, Stuttgart
Epstein SR (1998) Craft guilds, apprenticeship, and technological change in preindustrial Europe. J Econ Hist 58(3):684–713
Epstein SR (2004) Property rights to technical knowledge in premodern Europe, 1300–1800. Am Econ Rev 94(2):382–387
Fouquet R (2008) Heat, power and light: revolutions in energy services. Edward Elgar Publications, Cheltenham/Northampton
Fouquet R, Broadberry S (2015) Seven centuries of European economic growth and decline. J Econ Perspect 29(4):227–244
Fourrier C (1965) L'Enseignement Francais de 1789 à 1945. Institut Pédagogique National, Paris
Friedman M (1962) Capitalism and freedom. University of Chicago Press, Chicago
Fuller B (1983) Youth job structure and school enrollment, 1890–1920. Sociol Educ 56:145–156
Fuller B, Rubinson R (1992) Does the state expand schooling? Review of the evidence. In: Fuller B, Rubinson R (eds) The political construction of education. The State, State Expansion and Economic Change, Praeger, New York

Galor O (2005) From stagnation to growth: unified growth theory. In: Aghion P, Durlauf SN (eds) Handbook of economic growth, vol 1A. North Holland, Amsterdam, pp 171–293
Galor O (2011) Inequality, human capital formation, and the process of development. In: Hanushek EA, Machin S, Woessmann L (eds) Handbook of the economics of education, vol 4. Elsevier, Oxford, pp 441–493
Galor O, Moav O (2002) Natural selection and the origin of economic growth. Q J Econ 117:1133–1192
Galor O, Tsiddon D (1996) Income distribution and growth: the Kuznets hypothesis revisited. Economica 63:103–117
Galor O, Weil DN (2000) Population, technology and growth: from the malthusian regime to the demographic transition. Am Econ Rev 90(4):806–828
Galor O, Moav O, Vollrath D (2009) Inequality in landownership, the emergence of human-capital promoting institutions, and the great divergence. Rev Econ Stud 76:143–179
Gennaioli N, La Porta R, Lopez-de-Silanes F, Shleifer A (2013) Human capital and regional development. Q J Econ 128(1):105–164
Gilmont JF (1999) Protestant reformations and reading. In: Cavallo G, Chartier R (eds) A history of reading in the West. Polity Press, Oxford, pp 213–237
Glenn CL (2012) Educational freedom and protestant schools in Europe. In: Jeynes W, Robinson DW (eds) International handbook of protestant education. Springer Science+Business Media B.V., Dordrecht, pp 139–161
Glomm G, Ravikumar B (1998) Increasing returns, human capital, and the Kuznets curve. J Dev Econ 55:353–367
Gradstein M, Justman M, Meier V (2005) The political economy of education. Implications for growth and inequality. MIT Press, Cambridge
Graff HJ (1991) The literacy myth. Transaction Publishers, New Brunswick
Green H (1979) The education of women in the reformation. Hist Educ Q 19:93–116
Green A (1990) Education and state formation. Macmillan, Hampshire
Grossman GM, Helpman E (1991) Innovation and growth in the global economy. MIT Press, Cambridge
Guellec D (2004) Gutenberg revisité. Une analyse économique de l'invention de l'imprimerie. Rev Écon Polit 114(2):169–199
Hamilton K, Liu G (2014) Human capital, tangible wealth, and the intangible capital residual. Oxf Rev Econ Policy 30(1):70–91
Hanushek E, Kimko D (2000) Schooling, labor force quality, and the growth of nations. Am Econ Rev 90(5):1184–1208
Hanushek EA, Woessmann L (2008) The role of cognitive skills in economic development. J Econ Lit 46(3):607–668
Harris JR (1998) Industrial espionage and technology transfer. Britain and France in the Eighteenth Century. Ashgate, Aldershot
Hippe R (2012) How to measure human capital? The relationship between numeracy and literacy. Econ Soc 45(8):1527–1554
Hippe R (2013a) Are you NUTS? The factors of production and their long-run evolution in Europe from a regional perspective. Hist Soc Res 38(2):324–348
Hippe R (2013b) Spatial clustering of human capital in the European regions. Econ Soc 46(7): 1077–1104
Hippe R (2014) Human capital and economic growth: theory and quantification. Econ Soc 49(8): 1233–1267
Hippe R (2015) Why did the knowledge transition occur in the West and not in the East? ICT and the role of governments in Europe, East Asia and the Muslim world. Econ Bus Rev 1(1):9–33
Hippe R, Baten J (2012) Regional inequality in human capital formation in Europe, 1790–1880. Scand Econ Hist Rev 60(3):254–289
Hippe R, Fouquet R (2018) The knowledge economy in historical perspective. World Econ 18(1): 75–107

Hippe R, Perrin F (2017) Gender equality in human capital and fertility in the European regions in the past. Investigaciones de Historia Económica – Econ Hist Res 13(3):166–179
Höhener J, Schaltegger CA (2012) Religionsökonomie: eine Übersicht. Perspekt Wirtsch 13(4): 387–406
Humphries J (2006) English apprenticeship: a neglected factor in the industrial revolution. In: David PA, Thomas M (eds) The Economic future in historical perspective. Oxford University Press, Oxford, pp 73–102
Hurt J (1971) Education in evolution. Paladin, London
Iannaccone LR, Finke R, Stark R (1997) Deregulating religion: the economics of church and state. Econ Inq 35:350–364
Johnes G (1993) The economics of education. Macmillan Press, London
Jones CI (2002) Sources of U.S. economic growth in a world of ideas. Am Econ Rev 92:220–239
Krueger AB, Lindahl M (2001) Education for growth: why and for whom? J Econ Lit 39(4): 1101–1136
Kuznets S (1955) Economic growth and income inequality. Am Econ Rev 45:1–28
Landes DS (1969) The unbound prometheus: technological change and development in Western Europe from 1750 to the present. Cambridge University Press, Cambridge
Lange F, Topel R (2006) The social value of education and human capital. In: Hanushek EA, Welch F (eds) Handbook of the economics of education, Vol, vol 1, pp 459–509
Lauterbach U (1994) Apprenticeship, history and development of. In: Husén T, Postlethwaite TN (eds) The international encyclopedia of education, 2nd edn. Pergamon, Oxford, pp 310–318
Layard R (2005) Mental health: Britain's biggest social problem? LSE research online. http://cep.lse.ac.uk/textonly/_new/staff/layard/pdf/RL414_Mental_Health_Britains_Biggest_Social_Problem.pdf
Lim ASK, Tang HW (2008) Human capital inequality and the Kuznets curve. Dev Econ 46:26–51
Lindert PH (2004) Growing public. Social spending and economic growth since the eighteenth century, vol I. Cambridge University Press, Cambridge
Lindert PH (2014) Private welfare and the welfare state. In: Neal L, Williamson JG (eds) The Cambridge history of capitalism. Cambridge University Press, Cambridge, pp 464–500
Lochner L, Moretti E (2004) The effect of education on criminal activity: evidence from prison inmates, arrests and self-reports. Am Econ Rev 94(1):155–189
Lucas RE (1988) On the mechanics of economic development. J Monet Econ 22:3–42
Luther M (1909) Eine Predigt, daß man Kinder zur Schule halten solle (A Sermon on Keeping Children in School). In: Dr. Martin Luthers Werke: Kritische Gesamtausgabe, vol 30, Part 2. Verlag Hermann Böhlhaus Nachfolger, Weimar
Madsen JB (2010) The anatomy of growth in the OECD since 1870. J Monet Econ 57(6):753–767
Madsen JB, Saxena S, Ang JB (2010) The Indian growth miracle and endogenous growth. J Dev Econ 93(1):37–48
Malmström P (1813) Essai sur le système militaire de la Suède. Charles Delén, Stockholm
Mankiw NG, Romer D, Weil DN (1992) A Contribution to the empirics of growth. Q J Econ 107(2):408–437
Matsuo M, Tomoda Y (2012) Human capital Kuznets curve with subsistence consumption level. Econ Lett 116:392–395
McLaughlin E, Hanley N, Greasley D, Kunnas J, Oxley L, Warde P (2014) Historical wealth accounts for Britain: progress and puzzles in measuring the sustainability of economic growth. Oxf Rev Econ Policy 30(1):44–69
Meyer JW (1989) Conceptions of christendom: notes on the distinctiveness of the West. In: Kohn M (ed) Cross-national research in sociology. Sage, Newbury Park, pp 395–413
Mincer J (1958) Investment in human capital and personal income distribution. J Polit Econ 66:281–302
Mincer J (1974) Schooling, experience, and earnings. Columbia University Press, New York
Minns C, Wallis P (2013) The price of human capital in a pre-industrial economy: premiums and apprenticeship contracts in 18th century England. Explor Econ Hist 50(3):335–350

Mitch D (1982) The spread of literacy in 19th century England, PhD dissertation, University of Chicago
Mitch D (1986) The impact of subsidies to elementary schooling on enrolment rates in nineteenth-century England. Econ Hist Rev 39(3):371–391
Mitch D (1992a) The rise of popular literacy in Victorian England. The influence of private choice and public policy. University of Pennsylvania Press, Philadelphia
Mitch D (1992b) The rise of popular literacy in Europe. In: Fuller B, Rubinson R (eds) The political construction of education. The State, State Expansion and Economic Change, Praeger, New York
Mitch D (1999) The role of education and skill in the British industrial revolution. In: Mokyr J (ed) The British industrial revolution: an economic perspective, 2nd edn. Boulder, Westview, pp 241–279
Mokyr J (2009) Intellectual property rights, the industrial revolution, and the beginnings of modern economic growth. Am Econ Rev 99(2):349–355
Morgan NS (1997) Pen, print and Pentium. Technol Forecast Soc Chang 54:11–16
Morrisson C, Murtin F (2013) The Kuznets curve of human capital inequality: 1870–2010. J Econ Inequal 11(3):283–301
Murch J (1870) Five years' retrospect of literature, science and art. Bath Express and County Herald Office, William Lewis
Nelson RR, Phelps ES (1966) Investment in humans, technological diffusion, and economic growth. American Economic Association Papers and Proceedings 56(1–2):69–75
Peretto P (1998) Technological Change and Population Growth. J Econ Growth 3(4):283–311
Pritchett L (2001) Where has all the education gone? World Bank Econ Rev 15:367–391
Pritchett L (2003) "When will they ever learn?" Why all governments produce schooling, BREAD working paper no. 031
Psacharopoulos G, Patrinos HA (2004) Returns to investment in education: a further update. Educ Econ 12(2):111–134
Ramirez F, Boli-Bennett J (1982) Global patterns of educational institutionalization. In: Altbach P, Arnove R, Kelly G (eds) Comparative education. Macmillan, New York, pp 15–37
Ramirez F, Ventraensca MJ (1992) Building the institution of mass schooling: isomorphism in the modern world. In: Fuller B, Rubinson R (eds) The political construction of education. The State, State Expansion and Economic Change, Praeger, New York
Rappaport S (1989) Worlds within worlds: structures of life in sixteenth-century London. Cambridge University Press, Cambridge
Rashin AG (1958) Formirovanie rabochego Klassa Rossii. Sotsekgiz, Moscow
Romer PM (1986) Increasing returns and long-run growth. J Polit Econ 94(5):1002–1037
Romer PM (1990) Endogenous technological change. J Polit Econ 99(5):71–102
Rubinson R, Ralph J (1984) Technical change and the expansion of schooling in the United States, 1890–1970. Sociol Educ 57:134–151
Schultz TW (1961) Investment in human capital. Am Econ Rev 51:1–16
Schultz TW (1971) Investment in human capital. Free Press, New York
Schultz TW (1975) The value of the ability to deal with disequilibria. J Econ Lit 13(3):827–846
Schultz TW (1981) Investing in people: the economics of population quality. University of California Press, Los Angeles
Segerstrom PS, Anant ACT, Dinopoulos E (1990) A schumpeterian model of the product life cycle. Am Econ Rev 80(5):1077–1091
Sianesi B, van Reenen J (2003) The returns to education: macroeconomics. J Econ Surv 17(2): 157–200
Smits W, Stromback T (2001) The economics of the apprenticeship system. Edward Elgar, Cheltenham
Solow RM (1956) A contribution to the theory of economic growth. Q J Econ 70(1):69–94
Soysal YN, Strang D (1989) Construction of the first mass education systems in nineteenth-century Europe. Sociol Educ 62(4):277–288

Spence M (1973) Job market signaling. Q J Econ 87:355–379
Stöber R (2004) What media evolution is: a theoretical approach to the history of new media. Eur J Commun 19:483–505
Stone L (1969) Literacy and education in England, 1640–1900. Past Present 42:69–139
Sturm R (1993) How do education and training effect at country's economic performance? A literature review. RAND, Santa Monica
Swan TW (1956) Economic growth and capital accumulation. Econ Rec 32:334–361
Sweetland SR (1996) Human capital theory: foundations of a field of inquiry. Rev Educ Res 66(3):341–359
Thomas K (1986) The meaning of literacy in early modern England. In: Baumann G (ed) The written work: Literacy in transition. Claredon Press, Oxford, pp 97–131
Tilly C, Tilly L (1973) The rebellious century, 1830–1930. Harvard University Press, Cambridge
Uzawa H (1965) Optimum technical change in an aggregative model of economic growth. Int Econ Rev 6:18–31
Van Zanden JL (2009a) The long road to the industrial revolution: the European economy in a global perspective, 1000–1800. Koninklijke Brill NV, Leiden
Van Zanden JL (2009b) The skill premium and the 'great divergence'. Eur Rev Econ Hist 13(1):121–153
Vandenbussche J, Aghion P, Meghir C (2006) Growth, distance to frontier and composition of human capital. J Econ Growth 11:97–127
Venezky RL (1996) The development of literacy in the industrialized nations of the West. In: Barr R, Kamil ML, Mosenthal P, Pearson D (eds) Handbook of reading research, vol 2. Lawrence Erlbaum Associates, Mahwah
Vincent D (2000) The rise of mass literacy: reading and writing in modern Europe. Polity, Cambridge
Voth HJ (2001) The longest years: new estimates of labor input in England, 1760–1830. J Econ Hist 61(4):1065–1082
Wallis PH (2008) Apprenticeship and training in premodern England. J Econ Hist 68(3):832–861
Walters P, O'Connell PJ (1988) The family economy, work, and educational participation in the United States, 1890–1940. Am J Sociol 93:1116–1152
Weber M (1958) The protestant ethic and the spirit of capitalism. Charles Scribner's Son, New York
Weedon A (2003) Victorian publishing: the economics of book production for a mass market, 1836–1916. Ashgate Publishing, Aldershot
Young A (1998) Growth without scale effects. J Polit Econ 106(1):41–63
Zeira J (2009) Why and how education affects economic growth. Rev Int Econ 17:602–614

Education and Socioeconomic Development During the Industrialization

Sascha O. Becker and Ludger Woessmann

Contents

S. O. Becker
University of Warwick, Coventry, UK

CAGE, Coventry, UK

CEPR, London, UK

CESifo, Munich, Germany

IZA, Bonn, Germany

ifo, Munich, Germany

ROA, Maastricht, Netherlands
e-mail: s.o.becker@warwick.ac.uk

L. Woessmann (✉)
University of Munich and ifo Institute, Munich, Germany

CESifo, Munich, Germany

IZA, Bonn, Germany

CAGE, Coventry, UK

ROA, Maastricht, Netherlands
e-mail: woessmann@ifo.de

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_75

Abstract

This chapter discusses recent advances in our empirical knowledge of how education affected socioeconomic development during the industrialization. While early work attributed little role to formal education at the onset of the British Industrial Revolution, recent evidence is more positive. There is evidence, from Prussia and other European countries, that education played an important role in the first and second phases of industrialization in follower countries. While basic education seems to have been particularly relevant for the diffusion of the new industrial technologies, there is evidence that upper-tail human capital also played a role. In addition, the education of the population can account for major parts of the difference in Protestant and Catholic economic history. Beyond economic development, education also affected other societal developments during the industrialization. Educational expansion – in particular of advanced secondary schools – appears to have been an important force behind the decade-long process of secularization during the second phase of industrialization. In addition, the fertility decline during the demographic transition is closely related to increased education both in the generation of parents and of children, the latter indicating a significant trade-off between the quantity and quality of children.

Keywords

Education · Industrialization · Economic development · Economic history · Nineteenth century · Human capital · Schooling · Protestantism · Secularization · Demographic transition · Fertility

Introduction

For most of the history of mankind, the vast majority of human beings received hardly any formal education. For example, estimates put the share of the German population in early sixteenth century that was literate at 1% at most (Engelsing 1973). The virtual nonexistence of education for the masses coincides with their meagre and rather stagnant economic fate, certainly compared to the dynamics of economic advancement over the past century. In modern times, the expansion of mass education clearly contributed to economic development (Goldin 2016). Beyond educational attainment, it is particularly the actual basic skills of the population that appears to play a crucial role in global economic growth since

World War II (Hanushek and Woessmann 2008, 2012, 2016). What is less clear is the role of education in the period of transition between stagnation and modern growth, in particular during the onset and spread of industrialization.

This chapter assembles evidence on how education affected human development during the phase of industrialization. Traditionally, economic historians have tended to be quite skeptical about the role that the education of the population played during the industrialization. For example, despite his emphasis on the role of technological creativity for economic progress, Joel Mokyr (1990, p. 240) quite vividly emphasized that "If England led the rest of the world in the Industrial Revolution, it was despite, not because of, her formal education system." Over the past decade, however, contributions to cliometric research have advanced – and, in our view, altered – our understanding of the role of education during industrialization, in particular in the countries that followed England in industrializing. These recent advancements will be the focus of the first part of this chapter. In covering the recent evidence, it will distinguish between the first and second phase of industrialization as well as between different levels of education.

While understanding the sources of improvements in economic prosperity is clearly of vital importance, the role of education for human development during the industrialization certainly reaches beyond the economic sphere. Therefore, this chapter also aims to cover selected additional aspects of the effects of education on the socioeconomic development of the population. Addressing the differential development of regions along denominational lines, the second part will focus on the particular role of education in Protestant economic history. Staying with the religious realm, the third part will address the effect of education on religiosity itself, investigating its role in the process of secularization. The final part will cover evidence that education contributed to the demographic transition, addressing the role that the educational advancement of both children and mothers played in the fertility decline.

Education and Economic Development

We start with the role of education for economic development during the industrialization. The Industrial Revolution presumably marks the most fundamental technological shift in modern history. The period of industrialization is characterized by profound technological change sparked by such inventions as the steam engine and mechanical spinning, their diffusion, adaptation, and improvement, the rise of the factory system, and accompanying social changes in households and markets (cf. Mokyr 1999). We follow recent economic theory that tends to distinguish between a first phase of industrialization during which educational requirements were purportedly low and technological change was skill saving and a second phase during which skills became increasingly relevant for production as technological change increased the demand for human capital (cf. Galor 2005).

The Relevance of Education for Industrialization

To provide some conceptual background, we begin by discussing several aspects of the possible role of education for industrialization from a theoretical perspective. A first dimension in which the education of the population may have facilitated industrialization is the direct productive use of skills. If the tasks of a factory require a certain minimum level of skills, such as the ability to read basic instructions and perform basic calculations, establishing and operating a factory requires workers with basic literacy. Formal education may also impart behavioral traits and non-cognitive skills that are relevant for factory production, such as conscientiousness, discipline, orderliness, and perseverance (e.g., Field 1989). In addition, industrial production creates service jobs that require literacy and numeracy skills, such as accountancy, commercial transactions, banking, and lawyers (e.g., Laqueur 1974; Anderson and Bowman 1976).

A second dimension of education that may be relevant, particularly in the context of industrial catch-up, is its role in the adoption of new technologies. In motivating their catch-up model of technological diffusion, Nelson and Phelps (1966, p. 69) argue that "probably education is especially important to those functions requiring adaptation to change. Here it is necessary to learn to follow and to understand new technological developments." In a dynamic setting of changing technology, education plays a particular role by fostering the "ability to deal with disequilibria" (Schultz 1975), i.e., to perceive a given disequilibrium, to evaluate its attributes properly in determining whether it is worthwhile to act, and to undertake action to appropriately reallocate resources (see also Welch 1970). Such abilities are particularly relevant when technical change is disruptive rather than incremental, as is the case for most industries emerging during the Industrial Revolution, with the possible exception of the textile sector.

A third dimension in which education may facilitate industrial development is its role for entrepreneurship and innovation (e.g., Kocka 1977). Education may impart higher-level scientific skills and the ability to innovate that are necessary to advance technical knowledge. While this may seem foremost the task of higher education, a system of basic education that covers the broad masses may be a prerequisite to screen the highest-capable entrepreneurs and researchers. Thus, Landes (1980, p. 118) argues that "elementary schooling as such has been important . . . as a device for the recruitment of talent. . . . the bigger the pool one draws from, the better the chances of finding gifted and original scientists and technicians." This genuinely innovative dimension of the role of education may have been particularly relevant in sectors with fundamental technological breakthroughs during the second phase of industrialization, such as certain electrical and chemical industries.

The specific skills required to facilitate the adoption of the new technical and organizational modes of the industrialization are multifaceted and general, rather than applied to one particular craft, and may be best described as a general understanding of the functioning of the world. They start with the basic "three R's" of reading, writing, and arithmetic, required for commercial communication, accessing practical handbooks, decoding instructions, debugging new processes, and reading

books about foreign places – all relevant actions in the given historical setting (Anderson and Bowman 1976). They may also encompass socialization and the creation of an aspiring human personality with attitudes favorable to adopting new technology (Easterlin 1981). Relatedly, literacy may create awareness of non-conventional possibilities (Anderson and Bowman 1976).

Even though the industrialization may initially have created demand for uneducated labor – and often child labor – to perform routine tasks in some industries, the role of education in creating the ability to adjust to changing conditions and to develop and use the new industrial technologies may ultimately have been of great relevance during both phases of the industrialization.

Evidence from the First Phase of Industrialization

While these arguments suggest a potential role for education in the industrialization, initial evidence on the role of formal education during the first phase of the British Industrial Revolution does not really support this view. Quite to the contrary, Mitch (1993, p. 307) concludes his seminal review of the role of human capital in the first Industrial Revolution by stating that "education was not a major contributing factor to England's economic growth during the Industrial Revolution." These early studies seem to support the interpretation that the first phase of industrialization did not require substantial educational inputs.

It has to be acknowledged, though, that the early British evidence suffers from severe data constraints. In particular, studies had to rely mostly on proxying education by signatures in marriage registers for limited parish samples, observed concurrent to, but not before, the Industrial Revolution. Furthermore, the bulk of British evidence on the role of education is for the textile sector, a focus that may be rather specific as innovations were much more incremental and less disruptive in textiles than in other emerging industries (cf. Komlos 2000).

In any case, more recent evidence calls for a re-assessment of the role of education in the British Industrial Revolution. Based on an extensive new database, Madsen and Murtin (2017) conclude that education has been a crucial driver of British economic growth since 1270, with its contribution equally important before and after the first Industrial Revolution. Trends in new data on years of schooling in England from extensive source material suggest education facilitated preindustrial development, although this was not sustained after the initial stage of industrialization (de Pleijt 2018). Evidence on book production also suggests that Britain became one of the most literate countries in Europe during the centuries before 1750 (Baten and van Zanden 2008). Taking a perspective of more broadly defined human capital, Kelly et al. (2014) show that on the eve of the Industrial Revolution, British workers had higher levels of physical quality and mechanical skills than their continental counterparts.

Additional evidence from the first phase of industrialization comes from the follower countries that aimed to catch up to Britain as the technological leader during early industrialization. As indicated above, models of technological diffusion

in the spirit of Nelson and Phelps (1966) suggest that education may be a key ingredient to absorb new technologies and adapt to change. A number of descriptive studies have looked at the role of education in the industrialization of the textile sector in specific regions outside Britain, such as Catalan cotton factories from 1830 to 1861 (Rosés 1998), textile firms in Lowell, Massachusetts, around 1842 (Bessen 2003), and Southern Italian textile factories from 1861 to 1914 (A'Hearn 1998).

Using data on school enrollment and factory employment that link 334 counties from preindustrial 1816 to the first industrial phase in 1849, Becker et al. (2011) study the role of education for catch-up in the technological follower country Prussia during the first phase of industrialization until the mid-nineteenth century. Their evidence indicates that initially better-educated regions within Prussia responded more successfully to the opportunities created by the outside technological changes from Britain. Interestingly, formal education played a minor role at best for industrialization in the textile industry, where innovation was less disruptive and child labor more prevalent. By contrast, already in the first phase of the Industrial Revolution, basic school education is significantly related to industrial employment in the metal industry and in the other industries outside metals and textiles, such as rubber, paper, and food.

By themselves, such cross-sectional associations do not necessarily reflect a causal effect of education on industrialization. In fact, the process of industrialization may itself cause changes in the demand for education. On the one hand, factory production may increase the demand for low-skilled labor, drawing children out of school into factory work (e.g., Sanderson 1972). On the other hand, to the extent that industrialization increased living standards, education may have become more affordable for broader parts of the population.

To circumvent bias from such endogeneities, Becker et al. (2011) use education in 1816, before industrialization in Prussia, as an instrumental variable for education in 1849. This instrument is not affected by changes in the demand for education that emerged during industrialization, and thus isolates a part of the variation in education that is not determined simultaneously with industrialization. The instrumental variable model that restricts the analysis to that part of the educational variation in industrial times that is related to educational variation that pre-existed industrial times confirms a significant effect of education on industrialization during the first phase in Prussia. The validity of the instrumental variable specification is corroborated by the fact that results prove robust to the inclusion of an unusually rich set of control variables for the state of economic development before the onset of industrialization, indicating that preindustrial schooling is unlikely to capture the effects of other measures that are themselves related to subsequent industrialization.

Evidence from the Second Phase of Industrialization

In their analysis of the role of education for industrialization in Prussia, Becker et al. (2011) also investigate the second phase of industrialization, from 1849 to 1882. During the second phase of industrialization, the new industrial technologies may have increased the demand for human capital (e.g., Galor and Moav 2006), which

creates an additional source of potential endogeneity bias in cross-sectional analysis. Again using preindustrial education as an instrument for education levels during the second phase of industrialization, and controlling extensively for preindustrial development, the evidence indicates that basic education is significantly associated with nontextile industrial employment shares also during the second phase of industrialization in 1882.

To further address concerns that any preexisting omitted variables might drive the cross-sectional findings, Becker et al. (2011) also estimate panel data models that pool all three periods of observation – 1816, 1849, and 1882. Results confirm the effect of education on industrialization, and county fixed effects rule out that the findings simply capture unobserved heterogeneity across the counties.

The results suggest that it may not be arbitrary that Prussia, which was the educational world leader at the time (Lindert 2004), was particularly successful in catching up to – and ultimately passing in many sectors – the industrial leader Britain. Indeed, Prussian educational leadership seems to have translated into technological catch-up throughout the nineteenth century.

The particular mechanisms of educational expansion are studied further by Schüler (2016), who uses data on schooling inputs provided by the first two Prussian censuses on primary schools in 1886 and 1891. Her results indicate that teacher quality, as proxied by teacher unit costs, and to a lesser extent expenditures on educational infrastructure, had a positive effect on the change in income, proxied by per-capita income tax, in the industrializing Western part of Prussia. By contrast, there were no effects of differences in class size, despite the fact that class sizes were very large by modern standards. An exploration of the sectoral composition of the economy suggests that the rise in income may have stemmed from a shift to higher-skilled and better-paid occupations.

Beyond Britain and Prussia, Sandberg (1979) presents evidence consistent with an interpretation that education was a leading factor in the catch-up process of Sweden in late nineteenth century. In a cross-country analysis, O'Rourke and Williamson (1996) conclude that schooling had a modest effect on catch-up for a cross-section of 16 countries in the period 1870–1913. Using panel data for seven countries over the same period, Taylor (1999) confirms this result.

Different Levels of Education

As discussed above, different levels of education may affect different aspects of the industrialization process. In particular, basic education may prove important for general production and for the diffusion and application of technologies, whereas higher levels of education may be particularly relevant for innovation.

The results for Prussian industrialization in Becker et al. (2011) refer to measures of basic education, namely years of elementary and middle schooling (derived from enrollment rates) in 1849 and adult literacy rates in 1882. By contrast, enrollment rates in upper-secondary schools and the existence of universities are not significantly related to industrialization in their analysis. These results may partly

reflect the fact that education beyond the basic level was not very widespread at the time, as upper-secondary enrollment rates did not surpass 5% throughout the nineteenth century, and no more than eight universities existed in Prussia. In any case, the results suggest that basic follower mechanisms that stress the role of basic education for technology diffusion, rather than higher-skill or entrepreneurial channels, were most relevant for relative regional industrialization in Prussia in the nineteenth century.

In their analysis of Prussian data, Cinnirella and Streb (2017) find that various dimensions of human capital played an important role in innovation and economic development during the second phase of industrialization. In particular, they show that both the literacy rate of the population and the density of master craftsmen are related to higher patent activities and thus to technical change and higher incomes. While enrollment rates in secondary schools are not significantly related to patenting, they are in fact positively associated with income as proxied by income tax revenues.

As a proxy for advanced literacy skills, Baten and van Zanden (2008) build a dataset on the number of books per capita in preindustrial Europe. Their results indicate that countries with higher levels of per capita book production in the eighteenth century (1750–1800) had faster economic growth during the nineteenth century (1820–1913). They also show that countries with higher initial book production experienced faster wage growth in the preindustrial era.

Two papers document an important role for upper-tail human capital during industrialization. Meisenzahl and Mokyr (2012) highlight the role of technical competence, reflecting the upper tail of the human capital distribution, as a key factor in Britain's economic leadership. Focusing on the industrialization in France, Squicciarini and Voigtländer (2015) use Encyclopédie subscriptions in the mid-eighteenth century as a proxy for the presence of knowledge elites. They find that subscriber density as a measure of upper-tail knowledge predicts city growth after the onset of the French industrialization. A main mechanism appears to be increased productivity in innovative industrial technology. By contrast, while associated with development in the cross-section, literacy levels do not predict city growth in their analysis.

Education and Protestant Economic History

The economic role that education played during industrialization has an interesting repercussion for the relative economic performance of different Christian denominations: it provides an explanation for Protestant economic history.

A Human Capital Theory of Protestant Economic History

In his *The Protestant Ethic and the Spirit of Capitalism*, Max Weber (1904/05) famously put forward the thesis that a specific Protestant ethic promoted economic success, among others by making Protestants work harder and save more. As an

alternative hypothesis for Protestants' relative economic prosperity, Becker and Woessmann (2009) propose a simple human capital theory: a higher level of education among Protestants made them more productive and therefore increased their economic prosperity. The core argument is that Martin Luther had urged his followers to advance education so that they could read God's Word, the Bible.

As indicated, during Luther's lifetime about 99% of the population was unable to read and write (Engelsing 1973). Against this background, it is noteworthy that in 1524, Luther published his pamphlet "To the Councilmen of All Cities in Germany That They Establish and Maintain Christian Schools" (Luther 1524), in which he pressured the Protestant rulers to build and maintain schools. He thus tried to foster the supply side of education. He also addressed the parents, in a sermon published in 1530 under the title "A Sermon on Keeping Children in School" (Luther 1530), trying to encourage the demand side of education.

For Luther, the Bible was the Word of God and thus the direct connection between God and humankind. For this reason, he wanted all Christians to be able to read God's Word by themselves. But since being able to read the Word of God, in the first place, required being able to read, the emerging Protestant church had to make sure that the population became literate. To that extent, led by the "Educator of Germany" (Praeceptor Germaniae) Philip Melanchthon, the Protestant reformers regularly inspected the Protestant cities and congregations during church visitations to make sure that they had introduced a proper education system. As Dittmar and Meisenzahl (2018) show, many Protestant cities introduced city-level ordinances that regulated many aspects of life, and these "constitutions" were of crucial relevance for public education.

This Protestant push for education, coincidentally, may have had the side effect in the economic sphere of serving as human capital that brought economic development. Instruction in reading the Bible may thus have generated an educated workforce that made Protestants more productive, leading to economic prosperity, so the argument goes.

Using county-level data from the first Prussian population census, Becker and Woessmann (2010) show that in 1816 the school enrollment rate in Protestant-majority counties was at about two thirds. In Catholic-majority counties, it was at less than 50%. Protestantism already led to more schooling before the industrialization. Its beginning in Prussia is typically dated to around 1830. In 1871, when the Prussian Statistical Office first collected data about the literacy of the population, Becker and Woessmann (2009) find literacy rates to be at 90% in Protestant-majority counties, 8 percentage points higher than in Catholic-majority counties.

As this finding could in principle be the result of reverse causality – if more education-prone counties had been the ones to adopt Protestantism – Becker and Woessmann (2009) use an instrumental variables strategy that exploits the initial concentric dispersion of the Reformation to disentangle cause and effect. As an instrument, they use a county's distance to Wittenberg, the birthplace of the Reformation, to predict its share of Protestants in 1871. Using only that part of the variation in the share of Protestants predicted by the instrument, they find that higher shares of Protestants indeed caused higher literacy rates.

The positive effect of Protestantism on education has also been shown in other contexts. Using Swiss data, Boppart et al. (2013) confirm a persistent positive effect of Protestantism on several education indicators, in particular in areas with conservative milieus. In the Swiss data, the effect of Protestantism was particularly large for reading skills, but also existed in other subjects (Boppart et al. 2014). Goldin and Katz (2009) show that in the United States in 1910–1938, areas that led in secondary education had higher shares of Protestant population. Similarly, Go and Lindert (2010) report that in some specifications, Protestantism had a positive effect relative to Catholicism on several schooling outcomes across US counties in 1850. In Ireland in 1871, illiteracy among different Protestant factions was between 7% and 14%, while it was 40% among Catholics (Cipolla 1969). In Finland in 1880, only 1.3% of Lutherans were not able to read or write, against 54.4% of Catholics (Markussen 1990). Also across countries, Protestantism and literacy were very strongly correlated in 1900 (cf. Becker and Woessmann 2009). Evidence from many countries thus supports the positive association between Protestantism and education.

Education and Protestant Economic Development Over the Nineteenth Century

Becker and Woessmann (2009) go on to show that the Protestant quest for education also had economic consequences: a more highly educated workforce did make Protestants more productive. Indeed, they find that Protestant counties had higher salaries, higher income tax receipts, and a higher share of the workforce in manufacturing and services. They assess how much of these differences in economic outcomes can be accounted for by the education differences that can be traced back to the exogenous part of the spread of the Protestant Reformation. Their point estimates suggest that once income differences are adjusted for literacy differences, the remaining difference is no longer systematically related to the religious denomination. Thus, if education had the same effect on economic outcomes here as it has been shown to have in other settings, the results suggest that the higher literacy of Protestant regions can account for at least some of their economic advantage over Catholic regions. In fact, they are consistent with the hypothesis that literacy can account for most, if not all, of the Protestants' economic lead. Similarly, in a simple "horse race" between literacy and Protestantism in explaining economic outcomes, literacy has a large and significant effect and Protestantism loses all its association with economic outcomes, suggesting that the whole economic effect of Protestantism may run through increased human capital. These economic consequences of the Protestant Reformation were probably not directly intended, but rather a side effect of the Reformation – one that had very long-lasting consequences indeed.

Education may thus be an alternative to Max Weber's specific Protestant work ethic in explaining Protestant economic success. Once differences in literacy rates between Protestant and Catholic counties are controlled for, little difference, if any, remains in economic outcomes. This implies that there is little room for

alternative explanations that are more in line with the Weber thesis, such as differences in work effort and in the propensity to save, at least in the Prussian heartland of the Protestant Reformation.

Further analysis suggests that the effect of Protestantism on economic outcomes appears to have been restricted to areas outside the major cities. Cantoni (2015) does not find an effect of the Reformation on the growth of city sizes in Germany, consistent with an interpretation that literacy was relatively high in cities independent of religious denomination.

Gender-Specific Developments

Interestingly, the available evidence suggests that girls benefited from the Protestant education push even more than boys did. As early as 1520, 3 years after the Protestant Reformation started, Luther published one of his main pamphlets: "To the Christian Nobility of the German Nation Concerning the Improvement of the Christian Estate" (Luther 1520). He writes: "And would to God that every town also had a girls' school, in which the girls were taught the Gospel for an hour each day."

This call was followed up with local church ordinances. In particular, Johannes Bugenhagen, one of the leading Protestant reformers, was to set the standard for subsequent systems in his church ordinance for the city of Braunschweig in 1528, which provided for girls to learn reading. In this ordinance, Bugenhagen requested that the city should have both four boys' schools and four girls' schools. In his church ordinance for Wittenberg, he extended the request for girls' schooling to writing and calculating.

The church ordinances in turn seem to have led to effective changes in the provision of schooling, as documented by Green (1979) for the example of the visitations by church officials to the local parishes in the Electorate of Brandenburg. In this core state of what later became Prussia, the number of schools for girls increased more than tenfold between 1539 (when the Reformation was introduced) and 1600. The request to also foster education for girls is quite in contrast to Catholic thinking even a century later: for instance, in Bavaria, the biggest Catholic state in Germany at the time, there were still strong objections to schools in the countryside in general as late as 1614 (cf. Gawthrop and Strauss 1984).

As a consequence of the Protestant Reformers' attempt to educate girls, Protestantism indeed led to a decrease in the gender gap in basic education in Prussia in 1816 as well as in the gender gap in adult literacy in 1871 (Becker and Woessmann 2008). Interestingly, while compulsory schooling laws closed the gender gap in primary education for both denominations over the course of the nineteenth century, the pattern of effects of religious denomination on the gender gap then continues to show up in secondary and tertiary education in the twentieth century. For instance, in the first year when women were admitted to university in Prussia in 1908, there were more than eight times as many female students of Protestant denomination than of Catholic denomination.

The gender pattern is also evident in cross-country data. Across European countries in 1970, a higher share of Protestants in the population is associated with a higher educational gender parity index, measured as the ratio of years of education in the female and male population (Becker and Woessmann 2008).

Education and Secularization

The period of industrialization not only brought change on the economic front, but it also led to broader societal developments that we want to highlight in the next two sections. One relates to modernization and the role of religion. The other is about the role of the family and the number of children families had.

The Expansion of Advanced Schools and Religious Participation

Industrialization went hand in hand with new demands on education. Different advanced school types offered education beyond compulsory schooling, and new teaching materials might have challenged old truths. This also affected the role of the church in society. Karl Marx (1844) described religion as "opium of the people" that is required to alleviate the ailments of poor economic conditions. If he was right, improved material conditions may have reduced demand for religious consolation and hence reduced church attendance. This particular point was analyzed by Becker and Woessmann (2013) using Prussian census data from the late nineteenth and early twentieth century. They find no effect of increased income on church attendance in a sample of Prussian counties.

A similarly prominent hypothesis links education and religion. The fact that education increasingly moved to center stage also posed a challenge to the role of the church. In his book *The Future of an Illusion*, Sigmund Freud (1927) – like David Hume and others before him – took the view that progress in education and science would lead to a decline in religion (cf. McCleary and Barro 2006).

However, cross-sectional evidence generally contradicts this view and finds a positive association between education and church attendance (Iannaccone 1998). Many studies using modern cross-country and individual panel data come to mixed results. However, it is unclear how these results transmit to the context of actual decade-long societal developments during important historical phases of secularization, in particular during the times of the industrialization.

Only a few studies take a truly long-run view, tracing developments over several decades. Franck and Iannaccone (2014) construct a cross-country panel data set for 10 countries over the period 1925–1990 to identify determinants of church attendance. They exploit their rich panel data using panel fixed effects estimates. In most of their specifications they do not find a statistically significant effect of educational attainment on church attendance. But they find a link between government spending on education and church attendance in their models, which may indicate a particular role for how governments shape the content of schooling.

Becker et al. (2017) go back further in time, to the period of Germany's second phase of industrialization, at the turn from the nineteenth to the twentieth century. It is a period where the influence of the church in society, as measured by Protestant church attendance (participations in Holy Communion), fell significantly, by about 2 percentage points per decade. At the same time, secondary schooling developed and might indeed be related to the decline in church attendance. Toward the end of the nineteenth century, enrollment in compulsory schooling (ages 6–14) was basically universal, leaving little, if any, variation in enrollment in elementary and middle schools. The focus of the analysis is on advanced schools (*Höhere Unterrichts-Anstalten*), which taught students to the age of 18. Variation in enrollment in advanced schools seems particularly relevant to test the effect of education on secularization because advanced schools were most likely to convey the kind of scientific thinking stressed by secularization hypotheses. Advanced schools were restricted to male students, whereas the separate advanced school type for female students (*Höhere Mädchenschulen*) had a different focus.

Did the expansion of upper schooling that went along with the industrialization in Germany affect the attachment of Germans to the Church, as one pillar of society? The data used by Becker et al. (2017) cover 61 cities in Germany over the period 1890–1930. Cities are of particular interest because advanced secondary schools were generally found in cities, whereas in rural areas only the most basic type of education was provided. Several editions of the statistical yearbook of German cities provide data on advanced-school enrollment rates and other city-level characteristics. The church attendance data were assembled by Hölscher (2001). These unique data about participations in Holy Communion were originally collected as part of the statistics of the Protestant regional churches of Germany and refer to church districts. Those typically extend beyond the boundaries of cities, but can be linked to the city-level data.

Taking a cross-sectional view at the data, cities with a larger enrollment rate in advanced schools have higher church attendance. However, this could well be driven by unobserved time-invariant city characteristics that may be correlated with both education and church attendance. It seems more convincing to use city-fixed effects and identify from variation within cities over time, in order to relate changes in church attendance to changes in advanced school enrollment. In fact, once city-fixed effects are being used, a negative relationship appears: cities with a stronger increase in advanced school enrollment see a stronger decline in church attendance. This also holds when controlling for measures of economic development, e.g., municipal tax receipts per capita that reflect income in a city, as well as for total population to capture (speed of) urbanization.

Alternatively to panel-fixed effects estimates, first difference estimates give the same finding of a negative link between advanced school enrollment and church attendance. Results are robust to a series of alternative specifications, such as controls for changes in industry structure, welfare spending, political vote shares, and different income measures.

Dynamic panel-data methods suggest that changes in church attendance follow changes in advanced school enrollment. The effect of advanced-school enrollment is strongest after a lag of about 10 years, speaking against bias from reverse causation

and suggesting that the effect of advanced education on religious participation materializes over time. By contrast, lagged church attendance does not predict advanced-school enrollment. Results are also robust in models with lagged dependent variables to account for persistence in church attendance over time.

Using the supply of advanced schools as an instrument for advanced-school enrollment yields similar results. This specification alleviates remaining concerns about endogeneity from the demand side by exploiting variation in advanced-school enrollment that originates from the opening and closing of advanced schools in a city.

Different Levels of Education

A fascinating question is whether the link between schooling and secularization is driven by advanced secondary schools only and whether different types of advanced schools make a difference. Starting with the first question, one might wonder whether universities played a similar role as advanced secondary schools for the decline in church attendance. To check this, Becker et al. (2017) use an indicator for openings of new universities as well as for the share of university students in the city population. Neither enters the model significantly, and the coefficient on the enrollment rate in advanced schools remains qualitatively unaffected.

Another issue that is of interest is the role of church-run schools and secular schools. During the period of observation, some advanced schools changed from being religiously affiliated into nonreligious schools. This might have been associated with a change toward a less religious curriculum. If that was the case, the main effect of advanced school enrollment might capture a change in curriculum and educational content rather than an increase in education per se (cf. Franck and Iannaccone 2014). To address this concern, Becker et al. (2017) include the share of Protestant advanced schools in a city as an additional control. This variable does not enter the model significantly and does not qualitatively change the main result. The same is true for the share of Catholic or secular advanced schools.

One issue of interest when looking at advanced secondary school types is the role of critical thinking versus scientific knowledge, which are two candidate mechanisms to explain the negative link between advanced school enrollment and church attendance. The school type most directly oriented toward scientific knowledge was the newly emerging *Oberrealschule.* Their curriculum focused on the natural sciences and students learned scientific facts that may have reduced their belief in religious explanations of natural phenomena. The traditional *Gymnasium* had a humanistic curriculum focused on classical languages and literature. It may have instilled only limited knowledge in modern sciences, but instead have sparked critical thinking toward the church as an institution in general. The same may be true, to a lesser extent, for the *Realgymnasium* with its curricular focus on modern languages.

Regressions where enrollment in these different types of secondary schools are introduced separately indicate that the relationship between educational expansion

and decline in church attendance is as strong for classical humanist grammar schools as it is for the newly emerging upper secondary schools whose curriculum focusses strongly on natural sciences. This pattern is more consistent with a leading role for the conveyance of critical thinking in general – which may undermine belief in the institutionalized church – than with a particular role for learning specific scientific knowledge of facts of natural sciences.

Overall, the results suggest that the Protestant push for education that can be traced back for centuries may ultimately and accidentally have led to a decrease in Protestants' attachment to their church.

Fascinating new research by Cantoni et al. (2018) gives somewhat related evidence on the link between Protestantism and secularization, even though not directly on the effect of education per se and secularization. Their evidence shows that in the immediate aftermath of the Reformation, in the sixteenth century, secular authorities acquired enormous amounts of wealth from monasteries closed during the Reformation. Graduates of Protestant universities increasingly took secular, especially administrative, occupations. Protestant university students increasingly studied secular subjects, especially degrees that prepared students for public sector jobs, rather than church sector-specific theology.

Education and the Demographic Transition

Another seismic change coinciding with industrialization is the demographic transition, the decline in fertility and mortality. The demographic transition was often studied in isolation, but unified growth theory models the transition from Malthusian stagnation to sustained growth in a unified framework (cf. Galor 2005). A key feature of many of these models is that the new technologies that emerged during the process of industrialization increased the demand for education, which in turn triggered the demographic transition at the end of the nineteenth century. The trade-off between the quantity and quality of children is a central ingredient of unified growth theory. While the child quantity–quality trade-off has been documented extensively in modern data, only recently have attempts been made to trace it down in historic data for periods before or during the demographic transition.

The Trade-Off Between Children's Education and Fertility

Becker et al. (2010) present the first evidence on the existence of the child quantity–quality trade-off before the demographic transition. Using Prussian data, they give support to a central tenet of unified growth theory, namely that it is a perennial trade-off. Using data for the year 1849, well before the demographic transition in Prussia in the later decades of the nineteenth century, they show that the child quantity–quality trade-off was already in operation then. The basis for the analysis is a unique micro-regional dataset of more than 330 Prussian counties stemming from full population and education censuses conducted in 1849 by the

Prussian Statistical Office. Using ordinary least squares estimations, they regress the county-level child–woman ratio (as the measure of fertility) on school enrollment (as the measure of education), and vice versa. Regressions controlling for many confounding factors all show a negative conditional correlation between fertility and education.

In order to get at causality, and in order to closely link with the theoretical logic of the child quantity–quality trade-off, which requires exogenous variation in the price of education, in parents' preferences for education, and/or in the cost of raising children, instrumental variables estimation is employed. Specifically, land ownership inequality and distance to Wittenberg are used as instruments for primary education to identify the effect of education on fertility. The idea of using land ownership inequality builds on Galor et al. (2009) who present a theoretical model where inequality in the distribution of land ownership negatively affects the implementation of human-capital-promoting institutions (see Cinnirella and Hornung 2016 for Prussian evidence). Since large landowners would not benefit from the accumulation of human capital, given the low complementarity between land and human capital, they will hold back provision of education. The second instrument, distance to Wittenberg, follows the work of Becker and Woessmann (2009) discussed above showing that the higher share of Protestants associated with distance to Wittenberg also predicts stronger educational attainment. Using any or both of these instruments together, the negative effect of education on fertility is confirmed.

For the opposite direction of causality, the adult sex ratio serves as an instrument for fertility. The adult sex ratio constitutes a measure of marriage market tightness affecting marriage rates and fertility. Using this instrument, the negative effect of fertility on education is confirmed.

Having established mutual causation between education and fertility, Becker et al. (2010) conclude that the child quantity–quality trade-off was already in operation before the demographic transition, in the 1849 cross-section. Building on this insight, they go one step further and ask whether school enrollment in 1849 predicts the fertility transition in 1880–1905. It turns out that it does. Counties with a higher primary school enrollment rate in 1849 see a faster decline in fertility during the period of the demographic transition in Prussia.

Becker et al. (2012) go back even further in time, using data from 1816 Prussia, to show the existence of a child quantity–quality trade-off even in that earlier cross-section.

Similar results of a child quantity–quality trade-off in historic data have recently been found in other countries. Diebolt et al. (2017) use French département-level data from 1851. Their results show that a decline in fertility during the French demographic transition caused an increase in educational investments. The inverse relationship is not borne out in their data though, a result they interpret as implying that the education effect on fertility may take time before being effective.

Very recently, Fernihough (2017) and Klemp and Weisdorf (2018) made further headway in uncovering evidence of the child quantity–quality trade-off in historic data. Fernihough (2017) uses individual data from the 1911 Irish census and

documents that children who attended school were from smaller families, as predicted by a standard quantity–quality model. These results are also in line with the US findings by Bleakley and Lange (2009) who use data from 1910, which, however, is after the main phase of the fertility transition in the US context. Klemp and Weisdorf (2018) go furthest back in time and exploit a genealogy of English individuals living in the sixteenth to nineteenth centuries. Using exogenous variation caused by fecundity differences, they find that increases in the number of siblings caused reductions in adult literacy.

It is fair to say that there is now ample evidence of a link between education and fertility before and during the period of industrialization, across a wide range of countries.

Going beyond the child quantity–quality trade-off, higher levels of education before the onset of the demographic transition predict the speed of the fertility decline during the demographic transition (Becker et al. 2010). Murtin (2013) uses cross-country data over more than 100 years, covering the period 1870–2000, to study the long-term economic determinants of the demographic transition. As predicted by unified growth theory, he finds that primary schooling is the most robust determinant of the fertility transition.

Women's Education and Their Fertility

The child quantity–quality trade-off is the most obvious incarnation of a link between education and fertility. It follows directly from theory and considers the trade-off faced by parents optimizing over the number of children and their education levels, linking quantity and quality within the same generation.

It is, however, conceivable that intergenerational factors matter as well. Becker et al. (2013) combine Prussian county data from three censuses – 1816, 1849, and 1867 – to estimate the relationship between women's education and their fertility before the demographic transition. Despite controlling for several demand and supply factors, they find a negative residual effect of women's education on fertility. Instrumental-variable estimates using educational variation deriving from land ownership concentration, as well as panel estimates controlling for fixed effects of counties, suggest that the effect of women's education on fertility is causal.

Similar evidence has been found for France. Murphy (2015) uses French département-level data for the last quarter of the nineteenth century. He exploits regional variation to study the correlates of fertility, estimating various fixed-effects models. His findings confirm the important role played by education in general, and female education in particular. He uses literacy rates of adult males as well as a measure of the literacy gap between adult females and males as predictors of marital fertility. In this sense, his work is more closely related to the intergenerational effect of (parental) education on their fertility than to the child quantity–quality trade-off. To conclude, education played an important role for the demographic transition via several avenues.

Conclusion

Recent advancements in cliometric research have furthered our understanding of the link between education and socioeconomic development during industrialization. We can summarize the findings of recent research as follows.

First, while the role of education during the British Industrial Revolution has long been the subject of scholarly debate, evidence from follower countries such as Prussia and France suggests that education helped these countries to catch up with the industrial leader Britain. The results on the important role of education during the first phase of industrialization in follower countries might call for a revision of the common interpretation in the literature that the first phase of industrialization had purportedly low educational requirements.

Second, recent research also points to regional differences in education and economic development that follow denominational differences. Protestant areas of Prussia had higher literacy than Catholic areas, a result that has also been found for Switzerland. These differences in education translated into further economic development in Prussian counties, pointing to an alternative interpretation of denominational differences in economic development than Weber's Protestant ethic. Instead of different attitudes to work or different saving behavior, differences in educational attainment may be a credible alternative explanation.

Third, the expansion of education also affected secularization. Cities in Germany saw church attendance decline as a consequence of higher enrollment in advanced secondary schools. These schools may have conveyed general critical thinking that may have undermined belief in the church as an institution.

Fourth, education interacted with fertility and was thus a factor in the demographic transition during the nineteenth century. In the context of the child quantity–quality trade-off, parents traded off the number of children against investments in education per child. The increasing payoff to education during the nineteenth century meant that parents increasingly opted for more education of fewer offspring. At the same time, more educated parents had fewer children, whereby education had a double effect on the fertility decline.

While taking a broad perspective, we have given substantial prominence to findings from the Prussian industrialization period, not least because Prussian census data start earlier (in 1816) than in most other countries (cf. Becker et al. 2014). Prussian county-level data had already been analyzed by demographers 25 years ago (cf. Galloway et al. 1994). More recently, these findings have been complemented with evidence from other countries where large-scale digitization efforts started somewhat later.

Certainly, education had additional consequences on human development during industrialization beyond those covered here. For example, Cinnirella and Schüler (2018) show that the share of central spending in education positively affected votes for pronationalist parties in Prussia in 1886–1911, suggesting that public primary education may have played a role of indoctrination that helped the process of nation building in the context of Imperial Germany. Without a doubt, the role of education during industrialization leaves abundant open questions for cliometric research to address in the future.

References

A'Hearn B (1998) Institutions, externalities, and economic growth in Southern Italy: evidence from the cotton textile industry, 1861–1914. Econ Hist Rev 51(4):734–762

Anderson CA, Bowman MJ (1976) Education and economic modernization in historical perspective. In: Stone L (ed) Schooling and society: studies in the history of education. Johns Hopkins University Press, Baltimore, pp 3–19

Baten J, van Zanden JL (2008) Book production and the onset of modern economic growth. J Econ Growth 13(3):217–235

Becker SO, Woessmann L (2008) Luther and the girls: religious denomination and the female education gap in nineteenth-century Prussia. Scand J Econ 110(4):777–805

Becker SO, Woessmann L (2009) Was Weber wrong? A human capital theory of Protestant economic history. Q J Econ 124(2):531–596

Becker SO, Woessmann L (2010) The effect of Protestantism on education before the industrialization: evidence from 1816 Prussia. Econ Lett 107(2):224–228

Becker SO, Woessmann L (2013) Not the opium of the people: income and secularization in a panel of Prussian counties. Am Econ Rev 103(3):539–544

Becker SO, Cinnirella F, Woessmann L (2010) The trade-off between fertility and education: evidence from before the demographic transition. J Econ Growth 15(3):177–204

Becker SO, Hornung E, Woessmann L (2011) Education and catch-up in the Industrial Revolution. Am Econ J Macroecon 3(3):92–126

Becker SO, Cinnirella F, Woessmann L (2012) The effect of investment in children's education on fertility in 1816 Prussia. Cliometrica 6(1):29–44

Becker SO, Cinnirella F, Woessmann L (2013) Does women's education affect fertility? Evidence from pre-demographic transition Prussia. Eur Rev Econ Hist 17(1):24–44

Becker SO, Cinnirella F, Hornung E, Woessmann L (2014) iPEHD – the ifo Prussian Economic History Database. Hist Methods 47(2):57–66

Becker SO, Nagler M, Woessmann L (2017) Education and religious participation: city-level evidence from Germany's secularization period 1890–1930. J Econ Growth 22(3):273–311

Bessen J (2003) Technology and learning by factory workers: the stretch-out at Lowell, 1842. J Econ Hist 63(1):33–64

Bleakley H, Lange F (2009) Chronic disease burden and the interaction of education, fertility, and growth. Rev Econ Stat 91(1):52–65

Boppart T, Falkinger J, Grossmann V, Woitek U, Wüthrich G (2013) Under which conditions does religion affect educational outcomes? Explor Econ Hist 50(2):242–266

Boppart T, Falkinger J, Grossmann V (2014) Protestantism and education: reading (the bible) and other skills. Econ Inq 52(2):874–895

Cantoni D (2015) The economic effects of the Protestant Reformation: testing the Weber hypothesis in the German lands. J Eur Econ Assoc 13(4):561–598

Cantoni D, Dittmar J, Yuchtman N (2018) Reformation and reallocation: religious and secular economic activity in early modern Germany. Q J Econ 133(4):2037–2096

Cinnirella F, Hornung E (2016) Landownership concentration and the expansion of education. J Dev Econ 121:135–152

Cinnirella F, Schüler RM (2018) Nation building: the role of central spending in education. Explor Econ Hist 67:18–39

Cinnirella F, Streb J (2017) The role of human capital and innovation in economic development: evidence from post-Malthusian Prussia. J Econ Growth 22(2):193–227

Cipolla CM (1969) Literacy and development in the West. Penguin, Harmondsworth

de Pleijt AM (2018) Human capital formation in the long run: evidence from average years of schooling in England, 1300–1900. Cliometrica 12(1):99–126

Diebolt C, Menard A-R, Perrin F (2017) Behind the fertility–education nexus: what triggered the French development process? Eur Rev Econ Hist 21(4):357–392

Dittmar J, Meisenzahl RR (2018) Public goods institutions, human capital, and growth: evidence from German history. Rev Econ Stud (forthcoming)

Easterlin RA (1981) Why isn't the whole world developed? J Econ Hist 41(1):1–19
Engelsing R (1973) Analphabetentum und Lektüre: Zur Sozialgeschichte des Lesens in Deutschland zwischen feudaler und industrieller Gesellschaft. Metzler, Stuttgart
Fernihough A (2017) Human capital and the quantity-quality trade-off during the demographic transition. J Econ Growth 22(1):35–65
Field AJ (1989) Educational reform and manufacturing development in mid-nineteenth century Massachusetts. Garland, New York
Franck R, Iannaccone LR (2014) Religious decline in the 20th century West: testing alternative explanations. Public Choice 159(3–4):385–414
Freud S (1927 [1961]) The future of an illusion, Strachey J (ed). W. W. Norton, New York. [Original version (in German) published in: Psychoanalytischer Verlag, 1927]
Galloway PR, Hammel EA, Lee RD (1994) Fertility decline in Prussia, 1875–1910: a pooled cross-section time series analysis. Popul Stud 48(1):135–158
Galor O (2005) From stagnation to growth: unified growth theory. In: Aghion P, Durlauf SN (eds) Handbook of economic growth, vol 1A. North Holland, Amsterdam, pp 171–293
Galor O, Moav O (2006) Das Human-Kapital: a theory of the demise of the class structure. Rev Econ Stud 73(1):85–117
Galor O, Moav O, Vollrath D (2009) Inequality in land ownership, the emergence of human capital promoting institutions, and the great divergence. Rev Econ Stud 76(1):143–179
Gawthrop R, Strauss G (1984) Protestantism and literacy in early modern Germany. Past Present 104:31–55
Go S, Lindert PH (2010) The uneven rise of American public schools to 1850. J Econ Hist 70(1):1–26
Goldin C (2016) Human capital. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer, New York
Goldin C, Katz LF (2009) Why the United States led in education: lessons from secondary school expansion, 1910 to 1940. In: Eltis D, Lewis FD (eds) Human capital and institutions: a long-run view. Cambridge University Press, New York
Green L (1979) The education of women in the Reformation. Hist Educ Q 19(1):93–116
Hanushek EA, Woessmann L (2008) The role of cognitive skills in economic development. J Econ Lit 46(3):607–668
Hanushek EA, Woessmann L (2012) Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. J Econ Growth 17(4):267–321
Hanushek EA, Woessmann L (2016) Knowledge capital, growth, and the East Asian miracle. Science 351(6271):344–345
Hölscher L (2001) Datenatlas zur religiösen Geographie im protestantischen Deutschland: Von der Mitte des 19. Jahrhunderts bis zum Zweiten Weltkrieg, 4 vols. Walter de Gruyter, Berlin
Iannaccone LR (1998) Introduction to the economics of religion. J Econ Lit 36(3):1465–1495
Kelly M, Mokyr J, Gráda CÓ (2014) Precocious Albion: a new interpretation of the British Industrial Revolution. Annu Rev Econ 6(1):363–389
Klemp M, Weisdorf J (2018) Fecundity, fertility and the formation of human capital. Econ J (forthcoming)
Kocka J (1977) Entrepreneurship in a late-comer country: the German case. In: Nakagawa K (ed) Social order and entrepreneurship. University of Tokyo Press, Tokyo, pp 149–198
Komlos J (2000) The Industrial Revolution as the escape from the Malthusian trap. J Eur Econ Hist 29(2–3):307–331
Landes DS (1980) The creation of knowledge and technique: today's task and yesterday's experience. Daedalus 109(1):111–120
Laqueur TW (1974) Debate: literacy and social mobility in the Industrial Revolution in England. Past Present 64:96–107
Lindert PH (2004) Growing public: social spending and economic growth since the eighteenth century, 2 vols. Cambridge University Press, Cambridge
Luther M (1520) An den christlichen Adel deutscher Nation von des christlichen Standes Besserung (To the Christian nobility of the German nation concerning the reform of the Christian estate).

In: Dr. Martin Luthers Werke: Kritische Gesamtausgabe, vol 6. Verlag Hermann Böhlhaus Nachfolger, Weimar, 1888
Luther M (1524) An die Ratsherren aller Städte deutschen Landes, dass sie christliche Schulen aufrichten und halten sollen (To the councilmen of all cities in Germany that they establish and maintain Christian schools). In: Dr. Martin Luthers Werke: Kritische Gesamtausgabe, vol 15. Verlag Hermann Böhlhaus Nachfolger, Weimar, 1899
Luther M (1530) Eine Predigt, daß man Kinder zur Schule halten solle (A sermon on keeping children in school). In: Dr. Martin Luthers Werke: Kritische Gesamtausgabe, vol 30, Part 2. Verlag Hermann Böhlhaus Nachfolger, Weimar, 1909
Madsen JB, Murtin F (2017) British economic growth since 1270: the role of education. J Econ Growth 22(3):229–272
Markussen I (1990) The development of writing ability in the Nordic countries in the eighteenth and nineteenth centuries. Scand J Hist 15(1):37–63
Marx K (1844) Zur Kritik der Hegel'schen Rechtsphilosophie: Einleitung. In: Jahrbücher D-F (ed) Arnold Ruge, Karl Marx. Bureau der Jahrbücher, Paris, pp 71–85
McCleary RM, Barro RJ (2006) Religion and economy. J Econ Perspect 20(2):49–72
Meisenzahl RR, Mokyr J (2012) The rate and direction of invention in the British Industrial Revolution: incentives and institutions. In: Lerner J, Stern S (eds) The rate and direction of inventive activity revisited. University of Chicago Press, Chicago
Mitch D (1993) The role of human capital in the first Industrial Revolution. In: Mokyr J (ed) The British Industrial Revolution: an economic perspective. Westview, Boulder, pp 267–307
Mokyr J (1990) The lever of riches: technological creativity and economic progress. Oxford University Press, Oxford
Mokyr J (1999) The new economic history and the Industrial Revolution. In: Mokyr J (ed) The British Industrial Revolution: an economic perspective, 2nd edn. Westview, Boulder, pp 1–127
Murphy TE (2015) Old habits die hard (sometimes): can département heterogeneity tell us something about the French fertility decline? J Econ Growth 20(2):177–222
Murtin F (2013) Long-term determinants of the demographic transition, 1870–2000. Rev Econ Stat 95(2):617–631
Nelson RR, Phelps ES (1966) Investment in humans, technological diffusion, and economic growth. Am Econ Rev 56(2):69–75
O'Rourke KH, Williamson JG (1996) Education, globalization and catch-up: Scandinavia in the Swedish mirror. Scand Econ Hist Rev 43(3):287–309
Rosés JR (1998) Measuring the contribution of human capital to the development of the Catalan factory system (1830–61). Eur Rev Econ Hist 2(1):25–48
Sandberg LG (1979) The case of the impoverished sophisticate: human capital and Swedish economic growth before World War I. J Econ Hist 39(1):225–241
Sanderson M (1972) Literacy and social mobility in the Industrial Revolution in England. Past Present 56:75–104
Schüler RM (2016) Educational inputs and economic development in end-of-nineteenth-century Prussia. ifo working paper 227. ifo Institute, Munich
Schultz TW (1975) The value of the ability to deal with disequilibria. J Econ Lit 13(3):827–846
Squicciarini MP, Voigtländer N (2015) Human capital and industrialization: evidence from the age of enlightenment. Q J Econ 130(4):1825–1883
Taylor AM (1999) Sources of convergence in the late nineteenth century. Eur Econ Rev 43(9): 1621–1645
Weber M (1904/05) Die protestantische Ethik und der "Geist" des Kapitalismus. Arch Sozialwiss Sozialpolitik 20:1–54 and 21:1–110. Reprinted in: Gesammelte Aufsätze zur Religionssoziologie, 1920:17–206. [English translation: The Protestant ethic and the spirit of capitalism, translated by Talcott Parsons, 1930/2001, London: Routledge Classics.]
Welch F (1970) Education in production. J Polit Econ 78(1):35–59

Gender in Economic History

Joyce Burnette

Contents

Abstract

This chapter explores how cliometrics has helped us answer five questions about the role of women in economic history: How much did women participate in the economy? Why did women earn less than men? Why did men and women do different work? What determines gender roles? What role did women play in economic growth? The answers reveal that understanding the role of women is essential to understanding economic history.

Keywords

Women's work · Women's wages · Labor force participation · Gender gap · Occupational segregation · Gender roles · Gender norms · Girl Power

J. Burnette (✉)
Department of Economics, Wabash College, Crawfordsville, IN, USA
e-mail: burnettj@wabash.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_33

Introduction

What can the study of cliometrics offer to the study of gender? While one answer might be that it can quantify gender differences in wages and employment, in fact cliometrics has much more to offer. Cliometric provides a method for understanding the causes of the gender differences we observe. Its method is to use economic theories to form hypotheses about the past and then test these hypotheses with data. The cliometric method has helped us answer interpretive questions that do not seem to need quantitative analysis, such as why the gender gap exists or why gender roles vary across societies. It has helped us to answer questions about why we observe gender differences in behavior or how norms change over time. Before cliometrics, the relationship between culture and women's work was usually seen as one-directional: gender ideology explains what women did and what they were paid. Cliometrics has given us a window into how gender ideology is determined and has turned this causation around, demonstrating how economic realities determine the gender roles of different societies. To demonstrate how cliometrics has shaped our understanding of gender in history, I will explore the answers to five questions about gender in economic history.

How Much Did Women Participate in the Economy?

In economics the history of women's work is often written as one of increasing participation. In the USA, female labor force participation grew from 18% in 1890, and less than 5% for married women, to approximately 60% in the early twenty-first century (Costa 2000). Similarly, in England, married women's labor force participation grew from 10% in 1900 to 74% in 1998. While this growth is striking, focusing on the twentieth century is misleading because it leaves the impression that women's participation before the twentieth century was low. Historians have corrected this error by pointing out that in fact women's participation in the past was quite high, and the low participation rates of the early twentieth century were not the historical norm but a brief deviation from the long-established pattern of women's economic activity. Goldin (1995) noted a U-shaped relationship between female labor force participation and GDP per capita across a wide range of countries, suggesting that women's participation declines during industrialization and then rises. The historical evidence for the USA and UK fits that hypothesis; in both countries married women's labor force participation declined during the nineteenth century before rising during the twentieth century.

Any attempt to determine women's participation in the past must confront the fact that our definition of labor force participation is both unfair and anachronistic. An individual is counted as part of the labor force if he or she worked for pay. The self-employed and those who work from home are included, but today these categories contain only a small percentage of workers. Home production of goods consumed by the household is not included, even though these activities are economically productive. While economists know that home production should be counted as part of a

country's economic output, they have failed to adjust measures of GDP, arguing that home production is too difficult to value. In acknowledgement of this failure, labor force participation is often described as "work outside the home." Such a definition, however, is problematic for the gender historian.

The modern definition of labor force participation is unfair to women because it defines work usually done by women as not economically productive. A child-care worker employed by a day care counts as participating in the economy, while a mother taking care of her own children does not. If counted, household production for family use would increase GDP between 25% and 40%. By failing to count housework, we ignore a substantial portion of productive activity that has historically been women's work. In 2003, women did nearly twice as much housework as men, down from 1965 when women did three and a half times as much housework as men. Labor force participation does a particularly bad job of describing women's work in the past. A seventeenth-century woman who spent her day milking the cow, spinning yarn, helping her husband in his craft, and preparing the family meals did not have a job, but she was working nonetheless.

Failing to count household production as work is also anachronistic. At the beginning of the nineteenth century, housewives were considered productive workers, but during that century, a combination of economists who did not value women's contributions and male unions who wanted to argue that men's wages should be high enough to support a family led to re-definition of housewives as dependents rather than productive workers (Folbre 1991). In Britain, the 1861 census included wives and widows without a stated occupation in the category "Persons Engaged in the Domestic Offices." By 1881, however, wives were included in the category "Persons without Specified Occupations." Our implicit assumption that household production is not work is itself the result of a historical process that defined housewives as unproductive.

The concept of labor force participation is also anachronistic because in the past most people, men as well as women, produced goods that they used themselves. Households grew their own food and manufactured their own clothing. Households also made their own soap, brooms, and furniture. Relatively few people worked for wages, and most of those did not live entirely on their wages. In the sixteenth century, two-thirds of English rural residents had access to enough land to produce their own food and did not have to work for wages. Among those who did work for wages, the vast majority kept animals and thus produced at least some of their own food. Over time, households specialized more and purchased a greater percentage of their consumption goods from the market, so that eventually self-provisioning became a marginal activity rather than the norm. By ignoring household production, labor force participation ignores the majority of what men as well as women did in the past.

In addition to being a flawed concept, female labor force participation before the twentieth century is poorly measured. Most measures are based on censuses or tax records, which systematically ignore women's work. Because women did many different kinds of part-time work, or simply because they were women, they were not recorded as having occupations. Sometimes the work of women was not

recorded because a husband legally represented his wife; medieval English court records often list men as paying brewing fines even though their wives did the brewing (Bennett 1996). Amsterdam's 1742 tax register lists various ship captains as the owners of shops; these shops must have been run by their wives when they were at sea, but the wife is not mentioned (de Vries and van der Woude 1997). Numerous historical studies have identified workers who were paid by an employer but were listed with no occupation in the census. If these errors are corrected, female market participation was not as low in 1900 as it first appeared, but there is still a U-shaped pattern, with participation declining in the nineteenth century and increasing in the twentieth.

While reported occupations are often more a statement of gender ideology than of actual work, we can often find information about what women did in records not intended to record occupation, such as statements of witnesses in court cases. In fact, these records probably recorded what women did more accurately than official occupational lists. Examples of studies that have used court records to examine women's work are Earle (1989) for London, Ogilvie (2003) for Germany, and Agren (2017) for Sweden. All of these studies reveal wide participation of women in productive work.

If we include all productive activity, women's participation in work was nearly universal in the pre-industrial world. Even noble women participated in textile production. If we limit ourselves to work for the market, female participation was still widespread. In the early eighteenth century, witnesses in London courts were asked how they made their living; 54% of female witnesses claimed to be wholly maintained by their own employment, 18% partly maintained by employment, and only 28% claimed to have no paid employment (Earle 1989). Paid employment was likely less common in the countryside, but still much higher than the 10% observed around 1900. Horrell and Humphries (1995) examine the budgets of poor English families from 1787 to 1865 and find that around half of the wives had cash earnings to contribute to the family income. In seventeenth-century Netherlands, women outnumbered men in the fish market and the eel market, and Montague noted that "more women are found in the shops and business in general than men" (van den Heuvel 2007, pp. 41, 98). Hufton (1975, p. 10) concludes that in eighteenth-century France, "No girl expected to renounce work on marriage." Women's occupations were less likely than men's occupations to appear in official records because the man was seen as the head of the household, because many wives worked as assistants to their husbands, and because women often combined many different part-time activities. Nevertheless, women certainly worked.

If female participation in the market economy was so high before 1800, why did it decline during the nineteenth century? A number of factors combined to produce this decline. One reason was simply that male earnings rose. While there has been a great deal of debate about whether real wages in Britain rose during the period 1760 to 1830, it is generally agreed that wages rose after 1850. In the Netherlands as well, male earnings rose and measured female labor force participation fell during the second half of the nineteenth century (Botar 2017). Labor supply usually falls in response to higher non-labor income (leisure is a normal good), so we would expect

female labor supply to fall in response to rising male earnings. Another reason for declining market participation was that it became less socially acceptable for women to work outside the home. Trade unionists argued for higher wages by claiming that men should earn a family wage that would allow them to support their dependents. In doing so they created a situation where a man lost status if his wife was seen to work outside the home, because it implied he could not provide for his family. Mokyr (2000) suggests that increased knowledge of diseases and the discovery of germs led to an increased demand for cleanliness. Households purchased cleanliness by shifting the wife's labor from the market to housework. A less optimistic interpretation of the decline in female participation is that the demand for female labor was declining, so that women were unable to find employment (Humphries and Sarasua 2012). If we ignore household production, the market participation of married women reached a historic low around 1900 when the labor force participation rates of married women were about 10% in Britain and 6% in the USA (Costa 2000, p. 106).

The rise in female market participation over the twentieth century has been more extensively studied, but most studies describe the changes rather than identifying their cause. It is difficult to determine the cause of rising participation because most of the important variables were both a cause and effect of the changes in participation. Lower fertility, increased education, and changing cultural expectations contributed to rising female market participation, but were also a result of that rise. Analysis is made more difficult by the fact that women make important lifetime decisions about education and fertility when they are young, based on what they expect their opportunities to be in the future, so that at a point in time women of different cohorts behave differently.

While increasing male incomes tended to discourage female participation, women's labor supply response to their own wages has been positive during the twentieth century. This means that the substitution effect, which implies that women work more because work is more rewarding, dominates the income effect, which implies that women with higher incomes would spend some of that income on leisure and work less. However, women's wages were also rising during the nineteenth century, when their labor force participation was falling (Goldin 1990). Either other factors were important or the relationship between wage and labor supply changed over time. Galor (2005) suggests that there was a threshold, after which the effect of women's wages on their labor supply turned positive.

Since female wages rose during the twentieth century at the same time as participation increased, there must have been an increased demand for female labor. The sectoral shift from farming and manufacturing to services and the feminization of clerical jobs increased the demand for female labor since women have a comparative advantage in services, and women preferred white-collar work to blue-collar work. Women responded to the increase in demand, and the higher wages they could earn in the market, by working more. Later in the century, women entered a wide range of professional occupations. Entry into these occupations was, of course, made possible by women's increased education, which may itself have been the result of women's expectations of increased participation. Improved

contraception, especially the legalization of the pill, made women more willing to make human capital investments in their twenties, and thus increased the number of women in professional schools (Goldin and Katz 2002). Cheaper child care also contributed to increasing female participation. In Europe, public policy in the form of paid maternity leave and subsidized child care has encouraged female participation.

While we might expect that household technologies such as plumbing, washing machines, refrigerators, and vacuum cleaners reduced the demand for household production, the evidence suggests otherwise. These innovations did not reduce the time spent by a wife in housework because households either shifted work from paid domestic servants to the housewife or increased standards of cleanliness. However, time-saving appliances may have had effects on the next generation of women. Lewis (2015) finds that electrification had no immediate effect on female employment, but that it was associated with increased school attendance by girls and subsequently higher participation and earnings for women in the next generation.

Attitudes towards women's work outside the home changed considerably over the twentieth century, a change that was to some extent a result of increasing participation. The fraction of Americans who say they approve of "a married woman earning money in business or industry if she has a husband capable of supporting her" increased from around 20% in the 1940s to around 80% in the 1980s (Fernandez 2013, p. 473). While in 1968 only about a third of teenage girls expected to be working outside the home at age 35, by the 1980s over 80% expected to be working at age 35 (Goldin et al. 2006). Given their changed expectations, young women invested more in education and delayed childbearing. Women whose mothers were in the labor force are more likely to be in the labor force themselves. Gay (2017) shows that regions of France that had higher World War I mortality also had higher female participation, and that daughters of these women also had higher participation rates. Men's preferences also mattered; men who had working mothers are more likely to have working wives, and World War II mobilization had a temporary impact on the generation who were mothers during the war, but a permanent effect on the next generation (Fernandez et al. 2004).

While rising female participation is the story of the twentieth century, it is not the story of earlier periods. The low levels of market participation around 1900 were a historical aberration. If we also include production of goods and services for the use of the household, then even the supposedly low participation rates of the late nineteenth and early twentieth centuries would be much higher. Women's work is not a new invention, but a long-established habit. For women as for men, work has been the normal state of existence throughout history.

Why Did Women Earn Less than Men?

With a few rare exceptions, women have consistently earned less than men. For some, this consistency is proof that wages were socially determined. Women's lower pay has been attributed to the patriarchal structures of society, custom, or institutions. An alternative hypothesis is that women's wages were lower than men's

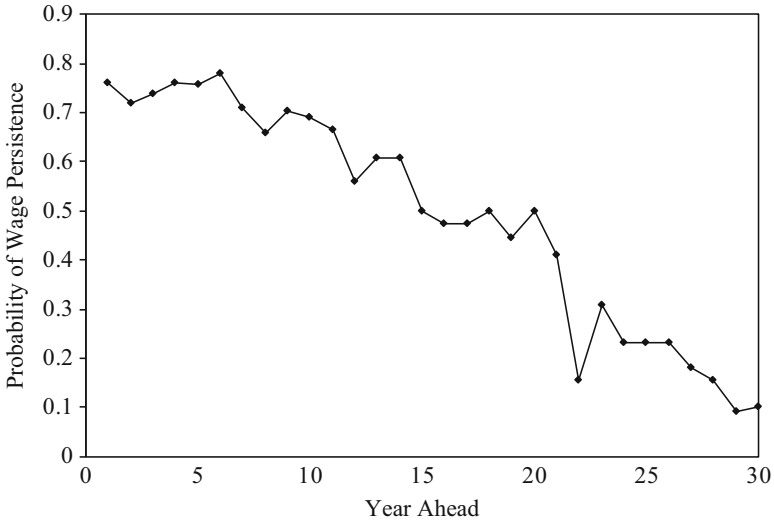


Fig. 1 The Persistence of Female Agricultural Wages over Time. (Source: Burnette 2008. Note: Based on a sample of 50 English farms. For each farm I record the modal summer wage for female day-laborers. The graph shows the probability that two wages at the same farm X years apart were the same.)

because their productivity was lower. Economic theory suggests that the wage is equal to the marginal product of labor, that is, the increase in the value of the firm's output that occurs when the worker is hired. If women had lower marginal productivity than men, a competitive market would pay them lower wages. Women could have had lower productivity than men even if they did the same work and put forth as much effort as men. Women might have had lower marginal productivity because they had fewer skills (which itself is socially determined) or because they are physically different from men. Women have less physical strength than men and, unlike for other gender differences, there is relatively little overlap in the male and female distributions; the average woman's maximum lift capacity is 2.5 standard deviations below the average man's (Burnette 2008). Throughout much of history, many production tasks depended on strength, and thus women's lesser strength would have given them lower productivity.

Since there are multiple hypotheses that predict that women will earn less than men, we need to examine other evidence to determine which hypothesis is correct. There is, of course, no reason to expect that the same hypothesis is correct for all times and all places. Future cliometricians will surely add refinements to the picture presented here. Currently the data suggest that, for most occupations before 1900, the gender gap was determined by market forces rather than custom.

If wages were set by custom, we would expect them to change relatively slowly, and they should not respond to changes in demand. The data allow us to reject these implications. Contrary to what has sometimes been suggested, women's wages were not fixed at the same level for decades, but did in fact change frequently. Figure 1

shows the probability that the female summer wage at a particular English farm remained the same a certain number of years in the future. The probability that a wage remained the same one year in the future was 76%, indicating that in 24% of the cases the wage changed from one year to the next. After 15 years, the probability of no change was only 50%. This flexibility suggests that women's wages were not determined by custom. There is also evidence that women's wages responded to changing economic circumstances. Women's as well as men's wages changed rapidly in response to the Black Death. For unskilled construction workers in Navarre, Spain, men's daily wages jumped from 9 pence in 1348 to 17.8 in 1349 and then fell a bit to 15 pence in 1353. Women's wages rose a bit more slowly but tripled over this period, rising from 4 pence in 1348 to 7 in 1349 and 12 in 1353 (Hamilton 1936). We also observe regional changes in English agricultural wages in response to the Industrial Revolution. Around 1750 women's wages were lower in the north of England than in the south. Over the next hundred years, however, women's wages rose in the north, where the factory production mushroomed, and stagnated in the south, where women lost traditional work such as hand spinning, so that by the early nineteenth century, women's farm wages were higher in northwestern England than in southern England (Burnette 2004).

It is also possible that wages did not match productivity because they were set by government regulation. There are numerous examples of governments setting wages, and some of these regulations were more effective than others. After the Black Death, the British government set maximum wages in an attempt to keep wages from rising. However, given that wages rose substantially in this period, the attempt seems to have been largely unsuccessful. Early modern Germany was more successful in enforcing maximum wages for spinners and seamstresses. The guilds and the state combined to enforce the maximum piece-rates, confiscating the wool and yarn of merchants promising to pay higher wages to spinners (Ogilvie 2003). The Germans had more success controlling wages because they generally had more control over activities in their community, including who was allowed to live in the community and who was allowed to marry. Thus market wages seem to have prevailed where markets were competitive, and non-market wages could be enforced only where there was a great deal of monopoly power.

Some have suggested that physical strength cannot explain the gender wage gap because this gap changed dramatically over time (Bardsley 1999; van Zanden 2011). In England, the female-male wage ratio for casual daily labor ranged from a high of 0.90 in the 1390s to a low of 0.32 in the 1840s (Humphries and Weisdorf 2015). However, the observed changes in the wage ratio are compatible with strength being an important determinant of the wage ratio. There is no reason to expect that the market price for strength would stay the same over centuries, any more than the market price of land or grain would remain fixed. The extent to which strength matters varies by task. It matters a lot for plowing, a little for weaving, and not at all for hand spinning. If we line up all the tasks in the economy by this ratio, an efficient economy would assign women tasks requiring less strength and men tasks requiring more strength. Employers will observe the market

wage ratio and hire women if the productivity ratio in their task is higher than the market wage ratio, and hire men if the opposite is true. The employer will hire both men and women only if the wage ratio is equal to the productivity ratio, so it is not surprising that men and women most often did different tasks, and only occasionally are found working together. The market wage ratio is then determined by the demand for different tasks and the number of male and female workers available. When a limited number of men were available relative to the demand for strength-intensive tasks, strength earned a premium. If a society had a high demand for strength-intensive tasks and the marginal task was plowing, the gender wage gap would be large, whereas if the society had a low demand for strength-intensive tasks and the marginal task was weaving, the gender wage gap would be smaller. Since labor markets were local, we would expect to find different wage ratios and different divisions of labor in each town. Even if strength did determine the female-male wage ratio, we would not expect the ratio to be constant over time or across locations.

While the female-male wage ratio in England fell from 80% in 1300 to around 40% in 1800, that fall was not continuous. The gender wage gap for day laborers increased between 1400 and 1600, fell until 1750, and then rose again (Humphries and Weisdorf 2015). During the fifteenth century, the expansion of the wool industry drove up wool prices and made sheep farming more profitable than arable agriculture. Evidence from later periods suggests that increased sheep-farming was associated with a decline in the demand for female labor in agriculture, reducing the relative female wage (Burnette 1999). If sheep-farming had the same effect on the demand for female labor in the fifteenth century, then we would expect to see an increase in the gender wage gap during this time. The fall in the English gender gap after 1600 corresponds to the increased demand for spinning due to the shift from woolen cloth to lighter worsted textiles, known as the "new draperies," which increased the relative demand for spinners. Muldrew (2012) estimates that for the older woolen cloths 59% of the workers were spinners, while for the "new draperies," 83% were spinners. This increase in the demand for spinning should have decreased the strength premium paid by the market. After the invention of the spinning jenny in 1764, machinery rapidly put hand spinners out of work, and the resulting decline in demand for female workers explains declining relative female wages over the period 1750–1850. Thus changes over time in the gender wage ratio are consistent with a strength-related gap, since the market price of strength can change.

When examining whether workers were paid for their productivity we must remember to compare outputs, not inputs. If a man and a woman did the same task for an hour, they provide the same input (one hour of labor) but do not necessarily produce the same output. If a woman reaped less grain in an hour than a man, and if she was paid the same rate per acre reaped, then she would earn less per hour. Evidence suggests that, when working at the same piece-rate, women did earn less per hour because they produced fewer units per hour. For nineteenth-century US manufacturing, Goldin (1990) finds that women produced 80% as much as men when working for identical piece-rates.

Table 1 Testing for wage discrimination

Date	County	Industry	Wage ratio	Productivity ratio	Source
1832	USA	Manufacturing	0.46	0.42	a
1839–1845	France	Cotton spinning	0.54	0.63	b
1839–1845	France	Wool spinning	0.49	0.43	b
1839–1845	France	Cotton and wool weaving	0.51	0.53	b
1850	USA	Manufacturing	0.49	0.41	a
1860	USA	Manufacturing	0.55	0.53	a
1870	Canada	Manufacturing	0.38	0.49*	c
1900	USA	Mfrg., blue collar workers	0.55	0.45	a
1900	USA	Mfrg., white collar workers	0.48	1.17*	a
1966	USA	Manufacturing	0.53	0.75*	d
1977	USA	Manufacturing	0.54	1.01*	d
1986-1993	Norway	Manufacturing	0.82	0.83	e
1989	Israel	Manufacturing	0.77	0.82	f
1990	USA	Manufacturing	0.55	0.84*	g
2002	USA	Manufacturing	0.72	0.84*	a

*Productivity ratio is significantly above the wage ratio, indicating wage discrimination
Sources:
a. Burnette (2015)
b. Cox and Nye (1989)
c. McDevitt et al. (2009)
d. Leonard (1984)
e. Haegeland and Klette (1999)
f. Hellerstein and Neumark (1999)
g. Hellerstein et al. (1999)

One way to test the productivity hypothesis is to use firm-level data to estimate the productivity ratio and then directly test whether the wage ratio equaled the productivity ratio. A small number of studies have done this for historical and modern labor markets. Table 1 summarizes the results of these studies. Most studies find that female workers had a lower marginal product than male workers, a difference that reflects their relative age and work experience as well as strength, though two studies find that women were at least as productive as men. Generally the productivity ratio rises over time; twentieth-century ratios are higher than nineteenth-century ratios. There is evidence of wage discrimination in the twentieth-century USA and in Canada in 1870, in the sense that the female-male wage ratio was lower than the productivity ratio. In other countries, and in the USA during the nineteenth century, there is no evidence of wage discrimination. In 1900, US manufacturing exhibits wage discrimination among white-collar workers but not among blue-collar workers.

The evidence in Table 1 supports Goldin's (1990) claim that wage discrimination appeared in the USA during the early twentieth century among clerical workers. The gender wage gap was negligible when workers were first hired, but grew with

experience as males gained more from their years with the firm than females. Explanations of wage discrimination involve firms offering incentives to men but not women. Goldin (1986) suggests that firms chose different monitoring systems for men and women, paying women, who were expected to quit, piece-rates, and offering men, who were expected to stay at the firm longer, wages that increased with tenure to discourage quits and encourage hard work. Owen (2001) suggests that men's lower turnover rates were the result of internal labor market policies. She finds that men's quit rates were initially higher than women's, but fell more rapidly than women's in the 1920s, the time during which the firms she studied centralized employment decisions and began to reward tenure. Owen suggests that women did not benefit as much as men from incentives such as pensions and company housing, and thus when firms implemented internal labor markets, male turnover fell more than female turnover and became less sensitive to labor demand conditions. Thus, wage discrimination seems to result from the fact that men, but not women, benefited from internal labor market policies designed to reward persistence at the firm. Estimated productivity ratios from US manufacturing in 2002 suggest that among workers under 35, the wage ratio is close to the productivity ratio, but that as workers age these ratios diverge. While men and women have similar increases in productivity as they age, male wages increase substantially more than female wages, so that among older workers there is substantial wage discrimination (Burnette 2015).

Wage discrimination is not the only possible form of discrimination women could face. If women were confined to low-wage occupations due to discrimination, they would have earned less due to the fact that crowding lowered the marginal product in those occupations and we would not find evidence of wage discrimination. Since occupational segregation was a different type of discrimination, we need to ask why women typically worked in lower-paid occupations than men.

Why Did Men and Women Do Different Work?

Throughout history, women and men have usually done different work. In societies that used the plow, men dominated agricultural work. Where land was cultivated with a hoe, women did more of the field work. Women produced most textiles, but sometimes men did certain tasks in textile production. Hand spinning was women's work, while finishing tasks such as fulling and shearing were usually men's work, and handloom weaving could be either. Routine household tasks such as cooking, cleaning, fetching water, and child care were generally women's work. Women also prepared most food and drink, though men also participated, particularly where production was on a larger scale. For example, men took over brewing when the move from ale to beer introduced large-scale production in the brewing industry (Bennett 1996).

While there are clearly gendered patterns of work, we should not exaggerate the extent of the differences. Agren (2017) suggests that work done by men and women in Sweden was not as different as occupational titles would suggest. While men were labeled as doctors and women were not, both men and women provided care for the

sick. While the frequency of tasks differed across the genders, there were few tasks that were never done by one of the genders. Men occasionally provided childcare, while women sometimes drove horses, shoveled soil, or cut down trees.

Even when women produced for the market rather than for domestic consumption, they generally specialized in food, drink, and clothing. Earle (1989) uses the statements of witnesses in London courts c. 1700 to examine the work of men and women. Most of the women claimed they were at least partly maintained by their own employment, and most of the work they did involved cleaning or preparing food or clothing. Work relating to domestic service, textiles and clothing, food provision, and nursing accounted for four-fifths of all women's work in this sample. Such observations have led some historians to conclude that women were confined to occupations that were associated with household production. Perhaps women were allowed to participate in the market, but only in areas associated with the domestic sphere, and thus seen as appropriate for women. This claim, however, ignores the extent to which men were also involved in producing food and clothing. For English towns in the late fourteenth century, Goldberg (1992) finds that around three-fourths of female household heads were employed in victualing or in textile and clothing trades, but she also reports that 45% of all household heads (who were primarily male) were in these trades. As late as 1841 nearly half of British men were engaged in trades relating to food and drink (including agriculture), textiles and clothing, or domestic service (compared to 89% for women). There were differences of course. Women were more concentrated in domestic work than men, and women were much less likely to work in the most high-paid and high-status occupations, but the difference is one of degree.

The same is true of the claim that women worked in the home and men worked outside the home. Gender differences in the location of work were not as great as ideology would suggest. Ogilvie (2003) finds that 45% of the time that women are mentioned in German church-court minutes they were in the home, while men were in their homes 39% of the time. For medieval England, Hanawalt (1986) finds that 21% of women's accidental deaths, and 8% of men's accidental deaths, occurred at home. While it is true that women were more likely to be at home than men, we should not conclude that women were confined to the home.

One hypothesis for why women did different work than men is that entry into certain occupations was constrained, so that men monopolized the most lucrative occupations and women were confined to lower-pay and lower-status occupations. In the pre-industrial period, guilds were the most likely source of occupational constraints. When guilds were strong, they could decide who was allowed to enter their trade. With industrialization unions and professional organizations replaced guilds as sources of occupational constraints. Women may also have found it hard to enter capital-intensive occupations as a result of law and inheritance practices that left them with less control over their wealth and less access to credit than men. To the extent that such constraints kept women out of better-paying occupations, women earned less than they otherwise would have.

An alternative explanation for the sorting we observe is that women choose different work than men because they maximized their incomes by doing

so. Women would not choose occupations where output was most sensitive to strength, such as plowing, and would choose occupations were strength mattered little or not at all, such as spinning. According to this explanation, occupational sorting does not reduce women's incomes, because if a woman did choose to plow she would not be very productive and would earn less from doing so than from textile production or food preparation. According to this view, differences in men's and women's work are simply responses to the different comparative advantages of men and women.

There is evidence in favor of both of these hypotheses, depending on the location and type of work. In general, comparative advantage was more important in low-skilled work, and in more competitive societies such as England and the Low Countries, while constraints were more important for skilled work, and for more highly regulated societies such as Germany.

Men's comparative advantage in tasks requiring strength is able to explain some of the broader historical patterns of work. For example, women rarely plowed, mowed with a scythe, or hewed coal, because these jobs required a great deal of upper body strength. The broad association of men with plow-based agricultural and women with textiles, which is found in many different societies, is consistent with comparative advantage. In one of the earliest Western texts, Homer's *Odyssey*, we observe women engaged in textile production (spinning and weaving) and men engaged in agriculture and war. The association of women with textiles and men with agriculture is also observed in China. Within textile production, hand spinning was consistently women's work. Handloom weaving, which requires some strength, was sometimes women's work, and sometimes men's work, depending on the demand for strength in the economy. Around the year 1000 in Europe looms became heavier and men took over weaving. During the Industrial Revolution, women and children continued to spin with the spinning jenny and the throstle, but with the invention of the spinning mule, adult males took over spinning because the mule required a great deal of strength, at least in its early form. The mechanization of spinning greatly reduced the number of spinners needed, and many women entered handloom weaving.

In a few cases, certain domestic tasks were assigned to women in spite of the fact that they required strength. Before wind- and water-powered mills were built, early medieval women used hand mills to grind grain into flour. Laundry has also been women's work, in spite of the strength required. Laundry likely required more strength than weaving, suggesting room for improved efficiency in nineteenth-century Europe.

Constraints on women's employment were most effective when enforced by specific institutions and least effective when they were simply cultural expectations. While many people expressed gender role preferences, such expectations do not seem to have been strong enough to prevent employers from hiring women when it was profitable to do so. There are many examples of employers acting in ways that are inconsistent with their own ideologies. In 1833, a farmer from Cornwall noted that "Farmers in this neighborhood are obliged to employ men for what women and children should do" ("Rural Queries" Newlyn East, Cornwall). In 1876, the owner of

a lace warehouse told a parliamentary committee that "we have as a rule an objection to employing married women, because we think that every man ought to maintain his wife without the necessity of her going to work." But he also admitted to employing 49 married women and added that "we wish that the present state of things as regards married women should not be disturbed" (quoted in Rose 1992, p. 32). While both men clearly had ideas about proper gender roles, they were quite willing to act in ways that conflicted with their ideology. The empirical evidence also suggests that employers were willing to change the gender composition of their workforce in response to changes in wages. In the 1770s, farmers hired more men where female wages were high, and more women where male wages were high.

Institutional barriers were sometimes more effective in limiting female employment. When guilds or unions were powerful enough to enforce their rules, they were able to prevent women from entering an occupation. German guilds, with the support of the state, excluded all women except wives from their trades. British mule-spinners, using violence, were able to prevent women from working as mule-spinners. Requiring a university education, which was not open to women, was an effective method of excluding women from medicine and law. Women in business could also be limited by their legal status. Generally, such barriers were weakest in Britain and the Low Countries, and strongest in Germany.

In England, occupational constraints were relatively weak. Many guilds had female members, and some guild regulations were written in inclusionary language (Power 1975). Girls could be apprenticed, and women could take apprentices, though few actually did. In the eighteenth century, only 4% of apprentices (not including parish apprentices) were girls, and only about 3% of those taking apprentices were mistresses (Simonton 1991). Not all guilds were open to women, though, and some specifically forbade the employment of women not related to the master. Even when they did try to limit women's work, however, English guilds had little power. In the 1690s only 38% of London merchants were members of a guild. Large numbers of weavers practiced the trade without an apprenticeship. By the early nineteenth century, only 5–10% of weavers had served an apprenticeship. In Leeds, cloth makers who were not guild members had their own hall for selling cloth (Burnette 2008). As a result, women can be observed doing a wide range of work. Women made nails and screws, and some worked as auctioneers or merchants. In the 1841 census, only 23% of occupations were exclusively male (Burnette 2008).

Women in the Low Countries were also relatively free to engage in the market. With many Dutch men either at sea or dead due to the high mortality rates there, women outnumbered men in Dutch towns and women conducted much of the business of the towns. In some Dutch markets, more than half of the sellers were women (van den Heuvel 2007). Guilds were not particularly strong in the Low Countries, perhaps because there were so many cities competing for trade and industry. During the fourteenth and fifteenth centuries, trade shifted from Bruges, to Antwerp, and then to Amsterdam, each time moving to a town where guilds had less power. Later, in the sixteenth through eighteenth centuries, Dutch guilds increased in number. However, the power of these guilds was still limited. In the sixteenth century, cities attempted to prohibit new industry in the countryside, and

even obtained a prohibition on new rural industries from the central government in 1531, but this prohibition proved unenforceable (de Vries and van der Woude 1997). While some Dutch guilds were closed to women, many guilds had women as members, and a few guilds were exclusively female. The hacklers' guild in Gouda and the guild for pulling boards through the canals of Utrecht were all-female (Schmidt 2009; de Vries and van der Woude 1997). Attempts to exclude women often proved unsuccessful. While the tailors' guild attempted to exclude women in the late seventeenth century, they were not successful and eventually admitted women to the guild. In Gouda in 1788–1789, one-third of all guild members were female, mainly because the largest guild, the tailors' guild, included seamstresses and was 84% female (Schmidt 2009).

However, even in these relatively open economies, not all work was available to women, and it was the most skilled occupations that effectively excluded women. By the nineteenth century, unions and professional organizations were more important than guilds in limiting female employment. While many male unions attempted to prevent female employment, only the most effective were able to do so. Scottish coal miners attempted to ban women workers in 1836, but failed. Generally, unions of low-skilled workers, such as handloom weavers, were unsuccessful. In skilled occupations, unions were more successful in their demands that employers not hire women. Scottish calico printers and English woolcombers excluded women. The mule-spinner's union successfully excluded women even after the self-actor reduced the strength requirements of the machine. The English compositors' union was relatively strong, so England had fewer female compositors than Edinburgh or Sweden. Tailors excluded women until 1834, when the union lost a strike and women entered the trade in large numbers.

Professions also succeeded in excluding women, often by controlling licensing or requiring formal education which as not available to women. The medical profession had historically included women. Early modern women acted as healers and sometimes joined the barber-surgeons guild, and before the eighteenth century all midwives were women. In sixteenth-century Norwich, 10 of 73 medical practitioners were women (Burnette 2008). In the eighteenth and nineteenth centuries however, women were forced out of medicine. The Royal College of Physicians admitted only university-educated men and gradually eliminated their competition by convincing the public that others were not qualified. Beginning in the seventeenth century, and culminating in male dominance of the profession in the nineteenth century, physicians somehow convinced people that women were not qualified to act as midwives. In the clergy as well, professionalization edged out women. While the Church of England did not allow women to be priests until 1994, Quakers and Methodists both had women preachers during the early years of their movements. However, as those movements became established and began to keep official lists of clergy, women preachers were no longer welcome. Thus, even in a relatively open society such as Britain, women did not have access to the most skilled occupations.

In Germany, occupational constraints were more severe. While fourteenth-century guilds admitted women, and Cologne had some all-female guilds, over time guilds placed restrictions on women's work, arguing that women's proper

place was in the home. By the eighteenth century, German guilds were strong and kept tight control over employment. All occupations except farming, spinning, housework, and manual labor were guilded, and only males were admitted to guild apprenticeships (Ogilvie 2003). This meant that women's work options were narrow. As a result, women sometimes ended up doing strength-intensive tasks such as plowing, which were not their comparative advantage. The system benefited men, who had less competition for their guilded work and thus higher earnings. In addition to limiting the range of occupations in which women could work, guilds also used wage ordinances to keep the wages of spinners and seamstresses low. The piece-rates set for spinners in the seventeenth and eighteenth centuries were below market rates and thus required active enforcement.

French guilds were less successful than German guilds in excluding women, and French women were not confined to the lowest-paid occupations. Wives worked with their husbands when married and had the right to continue in the trade when widowed. Many of the women guild members were not widows of masters, but were wives or widows of men in other trades. Some guilds were exclusively female. In thirteenth-century Paris, at least 5 of 100 guilds had only female members (Power 1975). In eighteenth-century Paris, the linen draper and seamstress guilds were all-female. Other guilds included large numbers of women members. In Saint-Malo about a quarter of women were drapers in the early seventeenth century, and about half were women at the beginning of the eighteenth century (Collins 1989). In Rouen, 10% of guild masters were female (Schmidt 2009). Where men did exclude women from their guilds, they faced serious competition from women producing outside of the guild system. In Paris, the tailors' guild excluded women from membership (while using the labor of their own wives and daughters) but faced serious competition from the female seamstresses' guild, which claimed the right to produce clothing for women and children, and from unguilded workers. During the eighteenth century, the reluctance of local government to enforce guild rules contributed to the weakening of the guilds. Fourteenth-century Parisian doctors attempted to exclude women. In 1322, Jacqueline Felicie de Alminia was prosecuted for practicing medicine without a license. However, exclusion does not seem to have been very successful; this was not the first time Jacqueline was prosecuted for practicing medicine, and there were several other women practicing in Paris at the time (Power 1975).

The extent to which legal status constrained women also varied greatly by country. Married women were usually considered to be *feme couvert* who had no legal status apart from their husbands, so they owned no property and could not make contracts. Such laws would seem to limit their ability to engage in business, but in many cases they were not as binding as they would appear. Married women could escape these constraints by operating as a *feme sole*. The concept of *feme sole* was widespread throughout Western Europe, but in practice countries differed in how much freedom women had. At the one extreme was the Netherlands, where *feme sole* status applied to any married woman in business. In Britain *feme sole* status was available but declining over time, though women could still maintain control over their property within marriage using contracts enforced in equity courts.

Phillips (2006) argues that couverture was not a significant impediment, and was sometimes an advantage, for women in business. While in Britain and the Low Countries unmarried women and widows could conduct their own legal business, Germany limited the rights of even these women. German law required all women to be represented in court by a *Kriegsvogt*, a man who was supposed to be her legal representative. In practice, the *Kriegsvogt* often acted in the interests of the men of the community, and the system failed to protect wives from abuse or widows from having their land taken (Ogilvie 2003).

While some of the gender division of labor was due to sorting by comparative advantage, some was due to institutional constraints on what work women could do. The extent of such constraints varied over time and place.

What Determines Gender Roles?

Gender norms clearly exist. The relevant questions are where they come from and to what extent they influence behavior. Cliometrics, by demonstrating the economic origins of gender roles and examining their persistence, has contributed a great deal to our understanding of gender roles. Gender roles do affect behavior, but they are also affected by behavior and are not independent of economic incentives.

In spite of the rapid changes that have occurred over the past few decades, traditional gender roles still matter. Fisman et al. (2006) demonstrate that women care more about intelligence, and men care more about appearance when choosing a partner, and that men avoid women who are more ambitious than themselves. Violating the norm that husbands should earn more than their wives carries a price. Women who earn more than their husbands are less likely to report a happy marriage, are more likely to divorce, and spend relatively more time on housework. Women with potential earnings higher than their husbands are more likely than other women to drop out of the labor force or reduce their hours of work (Bertrand et al. 2015). Cultural determinants of behavior are probably deeper than we realize. Men generally enjoy competition more than women, but this difference is cultural; in a patriarchal society, men are more likely than women to choose competition, while in a matriarchal society, women are more likely to compete (Gneezy et al. 2009).

While gender norms clearly matter, we should not exaggerate their importance in determining behavior. People often say one thing and do another, and gender ideology was often ignored in practice (Vickery 1993). As noted above, we observe employers hiring women even when they express ideological objections to doing so. Sometimes women bridged the gap between ideology and reality by actively hiding their economic role to conform to social expectations. Sophie Henschel, who inherited her husband's locomotive company when she was widowed, was in fact the decision-maker in the company, but public statements presented her son as the head of the business and emphasized Sophie's charitable work (Beachy 2006). The popularity of work such as taking in washing or lodgers may have been partially due to the fact that, since the work was done in the home, women could earn money without publically violating norms.

Since gender roles appear in ideological statements, it has generally been assumed that their origin was ideological. Bennett and Karras (2013, p. 5) claim that medieval gender roles were "rooted in the three religions of the time, Christianity, Judaism, and Islam, in scientific teachings, and in political traditions." They do not mention economic incentives as a contributor. Recent developments in cliometrics, however, have demonstrated the extent to which gender roles are themselves the result of occupational sorting driven by economic incentives.

Some studies demonstrate that women are more highly valued in regions where they have greater economic value. Two papers have shown that areas where women have lower economic value also have more boys than girls. Sex ratios favoring boys, which may result from either selective abortions or differential care leading to more deaths among girls, are generally seen as a sign of a social preference for males. An advantage of this measure is that we don't have to wonder whether people are hiding their true opinions when we ask them what they think about gender roles. Qian (2008) exploits the fact that tea-growing uses more female labor and fruit-growing uses more male labor. She shows that when reforms increased the price of cash crops, the survival of girls increased relative to the survival of boys in tea-growing areas, but not in fruit-growing areas. Carranza (2014) shows that Indian areas with more loamy soil, and thus more time spent plowing and less time spent weeding, have lower female participation in agriculture and fewer girls per boy under age six.

Economic incentives affect current gender norms, and these norms have effects that persist over time. Cliometric studies have demonstrated that current outcomes are affected by economic circumstances in the distant past. Since plowing requires a great deal of strength, in regions using the plow, men specialized in agriculture and women in more domestic tasks. Regions using the hoe for cultivation, however, saw greater female participation in agriculture. Alesina et al. (2013) show that regions of the world that used the plough in the pre-industrial period have lower female labor force participation and express less equal gender attitudes today. Hansen et al. (2015) show that countries with a longer history of agriculture have lower female labor force participation today. They claim that use of the plough was less important than whether cereals or root crops were grown, as cereals required more processing by the women. In China weaving of cotton cloth has historically been women's work, and technological improvements in the fourteenth century increased women's earnings from weaving and thus their social status. These positive attitudes towards women have persisted over centuries. Xue (2016) shows that areas in China that are climatically better suited to cotton weaving have relatively more equal sex ratios today (relatively more girls) and their residents are less likely to express opinions of male superiority. Cultural norms must have some persistence if agricultural practices thousands of years ago can explain current attitudes and outcomes. Thus cultural norms do matter for behavior, though these norms are not exogenous but were themselves formed in response to the economic demand for female labor.

Other studies demonstrate persistence over shorter time horizons. Fernandez and Fogli (2009) demonstrate that in 1970, the labor force participation of a second-generation US woman was affected by her father's country of origin. The lower the female labor force participation in the father's country of origin, the lower the

US-born daughter's participation. Women are also affected by the ethnic origin of their husbands. The lower the female labor force participation in the country of origin of the husband's father, the less likely the wife was to work. Gay (2017) finds that French women who grew up in regions with higher World War I mortality were more likely to have had working mothers when young and were more likely to be in the labor force as adults. Grosjean and Khattar (2015) demonstrate that in Australian regions that had more men relative to women in the nineteenth century, due to the transportation of convicts, women were more likely to marry, less likely to work outside the home, and less likely to work in a high-earning occupation. These effects were persistent. In regions with higher sex ratios and lower female participation in the past, people today are more likely to agree that a woman's place is in the home, and women are less likely to work in high-ranking occupations. These studies suggest that gender roles, though persistent, are determined by economic incentives more often than we realize. Thus, they acknowledge that cultural norms have some power over our behavior, but suggest that these norms have economic origins.

Received gender norms can, of course, be changed. Xue (2016) shows that, even with inherited Confucian values, areas of China with cotton production and thus higher economic productivity for women came to value women more. Sometimes gender norms change quite quickly. The fraction of teenage girls expecting to work outside the home at age 35 doubled during the 1970s. Exposure to women leaders can reduce discrimination in a relatively short time. Beaman et al. (2009) show that only 10 years after West Bengal reserved one-third of local council seats for women, men in villages that were required to have a female leader evaluated female leaders more positively.

Thus, there is evidence both that gender norms are malleable and that they are persistent. Gender roles change in response to new technology or new ideas, and in some cases can change quite quickly. Yet in spite of this malleability, economic activities millennia ago have a measurable impact on behavior today. How are these findings compatible? One possible solution is to posit that gender roles are difficult to change but that when they do change they change quickly. This would happen if there are multiple equilibria and it takes a fairly large shock to move out of one equilibrium. Xue (2016) suggests that gender roles change only in response to relatively large changes in circumstances, not to small changes. Alternatively, it could be that gender roles didn't change for centuries because the underlying economics that caused them didn't change, but then changed rapidly when circumstances changed. However, it seems implausible that persistence could be explained by the absence of large shocks over periods of thousands of years.

Some economists have sought to explain why some societies experience more change than others. Giuliano and Nunn (2017) show that changes in ideology are faster for people from areas of the world that have historically experienced less stable weather. This makes sense because the tradition is less valuable in societies where circumstances vary more. Exposure to foreigners also reduces the persistence of prejudice. Voigtländer and Voth (2012) show that anti-Semitism is persistent over the period 1350–1930, but that such views were less persistent in cities that were

members of the Hanseatic League, suggesting that involvement in trade with foreigners decreased prejudice.

One way to reconcile evidence of persistence with evidence of rapid change is to note that studies of persistence are picking up small effects. Typically, the portion of the variance explained by the distant past is quite small. Historical use of the plough explains only 6% of variation in female labor force participation and 11% of variation in female business ownership (Alesina et al. 2013). Even with geographic control variables, Grosjean and Khattar (2015) are able to explain only 2% of variation in attitudes towards women's work and 3% of the variation in female labor force participation in 2011. Xue (2016) finds that premodern textile production explains 0.15% of the variation in the sex ratio in 2000, and only 3–5% of variation in stated gender values. While economic incentives in the past can be linked to small changes in the probability of behaviors today, these influences are clearly not the most important determinants of women's work today.

What Role Did Women Play in Economic Growth?

Perhaps the most surprising finding of the cliometrics literature is that women's economic empowerment, or "Girl Power," was an essential foundation of modern economic growth. It was not a nice byproduct, but part of the process. Women's high earnings in Northwestern Europe resulted in later marriage, lower fertility, and increased human capital, which led to rising incomes. In unified growth theory, technological change during the nineteenth century generated increased demand for human capital and for female labor (Galor 2005). In the Girl Power story, the demographic shock of the Black Death and the associated change in the relative prices of land and labor generated the higher female wages, which in turn increased human capital. The story has been told in different ways by different authors; what follows is an attempt to synthesize the literature into a coherent story.

The story starts in Northwestern Europe after the Black Death, as pastoral agriculture expanded. The plague, by making labor scarce and land relatively abundant, made animal husbandry more profitable. Various explanations have been given for why the shift towards pastoral agriculture was concentrated in Northwestern Europe. Voigtländer and Voth (2013) suggest that Southern Europe was geographically not well suited to cattle, and that Eastern Europe did not shift to animal husbandry because they had higher grain productivity and grain remained profitable. De Pleijt and Baten (2017) suggest that the shift to pastoral agriculture was concentrated in regions where the population was more lactose tolerant.

Given their physical endowments, men have a comparative advantage in grain production, where tasks such as plowing and mowing require upper body strength, and women have a comparative advantage in dairy production. Thus, areas that increased dairy production also saw an increase in the demand for female servants. The female-male wage ratio rose after the Black Death, and these higher wages gave women the option of living on their own wages. Also contributing to female independence were marriage rules established by the Catholic Church requiring

consent of both partners, giving women as well as men a say in who they married (de Moor and Van Zanden 2010). The resulting increase in female autonomy has been called Girl Power.

Increases in female agency led to increased economic growth for two reasons. First, female independence led to later marriages and thus lower fertility, which allowed for increases in income per capita. Second, Girl Power also led to higher levels of investment in human capital for women and for their children. These mechanisms were mutually reinforcing but could operate independently.

Higher female wages were associated with higher ages of marriage, a greater portion of women remaining single, and married couples living separately from their parents, all of which were features of the European Marriage Pattern. With opportunities for market work at good wages, women could afford to support themselves while waiting for a suitable spouse. Additionally, jobs for women as dairy maids typically took the form of year-round live-in service that was more suited for single workers, encouraging later marriage. Voigtländer and Voth (2013) estimate that the average English woman married 4 years later as a result of the shift to pastoral agriculture. Fertility then declined for two reasons. First, lower fertility was a mechanical result of higher age at marriage. Since fertility was low outside of marriage and relatively unconstrained within marriage, fewer years spent in marriage would reduce fertility. Second, female empowerment also created incentives to lower fertility. Women's higher opportunity costs of time increased the cost of having a child and encouraged a shift from child quantity to child quality. Lower fertility eased population pressures and increased human capital as parents with fewer children invested more in each child.

Female empowerment also increased human capital directly. Human capital in this context is not necessarily formal schooling, which was relatively unimportant for most work. Human capital in this context should be seen as the acquisition of "useful knowledge" (Mokyr 2009). Higher wages and labor force participation for women encouraged investments in human capital for girls. More skilled mothers, in turn, could transmit more human capital to their children. We would expect women's greater bargaining power to increase education for all children, and to increase relative expenditure on girls. Modern studies have shown that mothers are more likely than fathers to spend their income on children's education (Qian 2008) and that grandmothers are more likely to spend their money on granddaughters (Duflo 2003). Countries with lower fertility and higher education entered a virtuous circle: higher investments in child quality increased incomes, which in turn encouraged investment in child quality.

Cliometricians have demonstrated that Girl Power lead to higher human capital by using age-heaping to measure numeracy. Numerate individuals are more likely to report their actual age, rather than rounding to the nearest multiple of five, so countries where more individuals report age 40 than age 39 or 41 were less numerate than countries where reported ages are evenly spread over those years. Baten et al. (2017) demonstrate that, within Eastern Europe, numeracy was higher in areas where fewer women in their twenties were married and in regions where the soil was less suitable for grain production. Comparing eighteen countries in Europe, de Pleijt and

Baten (2017) show that regions with more lactose tolerant populations had higher ages at marriage and, as a result, higher numeracy. (The connection between female power and human capital has been shown in other contexts as well; Qian (2008) shows that increases in female income increased educational attainment for both boys and girls in twentieth century China.)

High human capital, in turn, was important for economic growth. While Mokyr (2009) does not emphasize the importance of formal schooling in the British Industrial Revolution, he does emphasize the mechanical skills developed in trades such as clockmaking, skills obtained through apprenticeship, and scientific knowledge. Statistical evidence of the link between human capital and growth comes from de Pleijt and van Zanden (2016), who demonstrate that book consumption, their measure of human capital, was the most important determinant of GDP in Europe between 1300 and 1800.

This process is said to explain the "Little Divergence" in which England and the Low Countries pulled ahead of the rest of Europe in the seventeenth century. Many of the changes described above were reversed during the "Malthusian intermezzo" of the eighteenth century (van Zanden 2011). Women's relative wages fell, their age of marriage fell, and fertility rose. Somehow, in spite of the reversals, England was the first country to experience the "Great Divergence" of the Industrial Revolution, many centuries removed from the rise of Girl Power. Any connection between the two requires long lags. Fortunately for our story, the origins of the Industrial Revolution have been traced to processes that began much earlier. Mokyr (2009) has argued that it was the scientific revolution, beginning with Francis Bacon, that set the stage for the dramatic gains of the Industrial Revolution. Van Zanden (2009) emphasizes the increase in book production in Europe, beginning even before the printing press, both as evidence of and contributor to increasing human capital. In northwest Europe, literacy rose continuously between 1500 and 1800. Numeracy, as measured by age heaping, rose throughout Europe between 1350 and 1800, though data is scarce in the earlier centuries (A'Hearn et al. 2009). Though there were several centuries between the Black Death and the onset of modern economic growth, the story still makes sense if you believe that the Industrial Revolution was made possible by the gradual accumulation of human capital over many centuries.

Dennison and Ogilvie (2014) challenge the Girl Power story, noting that countries with the highest age at marriage were not necessarily those that developed first. Denmark and Iceland had higher age at marriage than England, but England reached modern economic growth first. However, while age at marriage is commonly used as a measure of Girl Power, it is not the same thing. The crucial change was female independence: women gained the ability to support themselves by wage labor and the ability to decide for themselves whether and whom to marry. This power led to a higher age at marriage, but a high age at marriage that appeared for other reasons would not have the same effects. Dennison and Ogilvie note that women's economic power was limited in some countries that exhibited characteristics of the European Marriage Pattern, leading them to doubt its causal power. In Germany, the age of marriage was high, but women were prevented from entering high-paid work and

were not allowed free choice in where they lived or whether they married. The solution to this dilemma is to see women's economic power as the underlying cause and the European Marriage Pattern as a byproduct. While a higher age at marriage was important for a reduction in fertility, the human capital mechanism does not require it. Increased female wages and participation in the labor market could increase human capital formation even if the age at marriage did not increase. It is also true that high earnings and a high age at marriage were not sufficient, if not accompanied by independence. Chinese women had high earnings in textile production, but did not get the control over their earnings, marriages, or education that European women got (Xue 2016). The Girl Power story then is truly one of the benefits of female empowerment.

Conclusion

In economic history, women have gone from an after-thought to a central character in the stories that we tell. We cannot understand economic growth without examining what happened to women. In general, we cannot understand the past unless we take women's contributions seriously. The statistical methods of cliometrics have been central to uncovering and proving the importance of women to history. Statistical methods have not only measured the extent of women's work and wages but have also identified the sources of gender norms and have demonstrated the importance and persistence of those norms.

References

A'Hearn B, Baten J, Crayen D (2009) Quantifying quantitative literacy: age heaping and the history of human capital. J Econ Hist 69(3):783–808

Agren M (2017) Making a living, making a difference: gender and work in early modern European society. Oxford University Press, Oxford

Alesina A, Giuliano P, Nunn N (2013) On the origins of gender roles: women and the plough. Q J Econ 128(2):469–530

Bardsley S (1999) Women's work reconsidered: gender and wage differentiation in late medieval England. Past Present 165(1):3–29

Baten J, Szoltysek M, Campestrini M (2017) 'Girl Power' in Eastern Europe? The human capital development of central and eastern Europe in the seventeenth to nineteenth centuries and its determinants. Eur Rev Econ Hist 21(1):29–63

Beachy R (2006) Profit and propriety: Sophie Henschel and gender management in the German locomotive industry. In: Beachy R, Craig B, Owens A (eds) Women, business, and finance in nineteenth-century Europe. Berg, Oxford

Beaman L, Chattopadhyay R, Duflo E, Pande R, Topalova P (2009) Powerful women: does exposure reduce prejudice? Q J Econ 124(4):1497–1540

Bennett J (1996) Ale, beer and brewsters in England: women's work in a changing world, 1300–1600. Oxford University Press, Oxford

Bennett J, Karras RM (2013) Women, gender, and medieval historians. In: Bennett J, Karras RM (eds) The Oxford handbook of women and gender in medieval Europe. Oxford University Press, Oxford

Bertrand M, Kamenica E, Pan J (2015) Gender identity and relative income within households. Q J Econ 130(2):571–614
Botar C (2017) Dutch divergence? Women's work, structural change, and household living standards in the Netherlands, 1830–1914. PhD dissertation, Wageningen University
Burnette J (1999) Labourers at the Oakes: changes in the demand for female day-laborers at a farm near Sheffield during the agricultural revolution. J Econ Hist 59(1):41–67
Burnette J (2004) The wages and employment of female day-laborers in English agriculture, 1740–1850. Econ Hist Rev LVII(4):664–690
Burnette J (2008) Gender, work, and wages in industrial revolution Britain. Cambridge University Press, Cambridge
Burnette J (2015) The paradox of progress: the emergence of wage discrimination in US manufacturing. Eur Rev Econ Hist 19(2):128–148
Carranza E (2014) Soil endowments, female labor force participation, and the demographic deficit of women in India. Am Econ J Appl Econ 6(4):197–225
Collins J (1989) The economic role of women in seventeenth-century France. Fr Hist Stud 16(2):436–470
Costa D (2000) From the mill town to the board room: the rise of women's paid labor. J Econ Perspect 14(4):101–122
Cox D, Nye JV (1989) Male-female wage discrimination in nineteenth-century France. J Econ Hist 49(4):903–920
de Moor T, van Zanden JL (2010) Girl power: the European marriage pattern and labour markets in the North Sea region in the late medieval and early modern period. Econ Hist Rev 63(1):1–33
de Pleijt A, Baten J (2017) Girl power generates superstars in long-term development: female autonomy and human capital formation in early modern Europe. EEHS conference
de Pleijt A, van Zanden JL (2016) Accounting for the 'Little divergence': what drove economic growth in pre-industrial Europe, 1399–1800? Eur Rev Econ Hist 20(4):387–409
de Vries J, van der Woude A (1997) The first modern economy: success, failure, and perseverance of the Dutch economy 1500–1815. Cambridge University Press, Cambridge
Dennison T, Ogilvie S (2014) Does European marriage pattern explain economic growth? J Econ Hist 74(3):651–693
Duflo E (2003) Grandmothers and granddaughters: old-age pensions and intrahousehold allocation in South Africa. World Bank Econ Rev 17(1):1–25
Earle P (1989) The female labour market in London in the late seventeenth and early eighteenth centuries. Econ Hist Rev 42(3):328–353
Fernandez R (2013) Cultural change as learning: the evolution of female labor force participation over a century. Am Econ Rev 103(1):472–500
Fernandez R, Fogli A (2009) Culture: an empirical investigation of beliefs, work, and fertility. Am Econ J Macroecon 1(1):146–177
Fernandez R, Fogli A, Olivetti C (2004) Mothers and sons: preference formation and female labour force dynamics. Q J Econ 119(4):1249–1299
Fisman R, Iyengar S, Kamenica E, Simonson I (2006) Gender differences in mate selection: evidence from a speed dating experiment. Q J Econ 121(2):673–697
Folbre N (1991) The unproductive housewife: her evolution in nineteenth-century economic thought. Signs 16(3):463–484
Galor O (2005) From stagnation to growth: unified growth theory. In: Aghion P, Durlauf S (eds) Handbook of economic growth, vol 1A. Elsevier, North Holland
Gay V (2017) The legacy of the missing men: the long-run impact of World War I on female labor force participation. Unpublished
Giuliano P, Nunn N (2017) Understanding cultural persistence and change. NBER working paper 23617
Gneezy U, Leonard KL, List J (2009) Gender differences in competition: evidence from a matrilineal and a patriarchal society. Econometrica 77(5):1637–1664

Goldberg PJP (1992) Women, work and life-cycle in a medieval economy: women in York and Yorkshire c. 1300–1520. Clarendon Press, Oxford
Goldin C (1986) Monitoring costs and occupational segregation by sex: a historical analysis. J Labor Econ 4(1):1–27
Goldin C (1990) Understanding the gender gap. Oxford University Press, New York
Goldin C (1995) The U-shaped female labor force function in economic development and economic history. In: Schultz TP (ed) Investment in women's human capital and economic development. University of Chicago Press, Chicago
Goldin C, Katz L (2002) The power of the pill: oral contraceptives and women's career and marriage decisions. J Polit Econ 110(4):730–770
Goldin C, Katz L, Kuziemko I (2006) The homecoming of American college women: the reversal of the college gender gap. J Econ Perspect 20(4):133–156
Grosjean P, Khattar R (2015) It's raining men! Hallelujah? Unpublished
Haegeland T, Klette TJ (1999) Do higher wages reflect higher productivity? Education, gender and experience premiums in a matched plant-worker data set. In: Haltwanger J, Lane J, Spletzer J, Theeuwes J, Troske K (eds) The creation and analysis of employer-employee matched data. Elsevier, Amsterdam
Hamilton EJ (1936) Money, prices, and wages in Valencia, Aragon, and Navarre, 1351–1500. Harvard University Press, Boston
Hanawalt B (1986) Peasant women's contribution to the home economy in late medieval England. In: Hanawalt B (ed) Women and work in preindustrial Europe. Indiana University Press, Bloomington
Hansen CW, Jensen P, Skovsgaard C (2015) Modern gender roles and agricultural history: the Neolithic inheritance. J Econ Growth 20(4):365–404
Hellerstein J, Neumark D (1999) Sex, wages, and productivity: an empirical analysis of Israeli firm-level data. Int Econ Rev 40(1):95–123
Hellerstein J, Neumark D, Troske K (1999) Wages, productivity, and worker characteristics: evidence from plant-level production functions and wage equations. J Labor Econ 17(3): 409–446
Horrell S, Humphries J (1995) Women's labour force participation and the transition to the male-breadwinner family, 1790–1865. Econ Hist Rev XLVIII(1):89–117
Hufton O (1975) Women and the family economy in eighteenth-century France. Fr Hist Stud 9(1):1–22
Humphries J, Sarasua C (2012) Off the record: reconstructing women's labor force participation in the European past. Fem Econ 18(4):39–67
Humphries J, Weisdorf J (2015) The wages of women in England, 1260–1850. J Econ Hist 75(2):405–447
Leonard J (1984) Antidiscrimination or reverse discrimination: the impact of changing demographics, title VII, and affirmative action on productivity. J Hum Resour 19(2):145–174
Lewis J (2015) Short-run and long-run effects of household electrification. Unpublished
McDevitt C, Irwin J, Inwood K (2009) Gender pay gap, productivity gap and discrimination in Canadian clothing manufacturing in 1870. East Econ J 35(1):24–36
Mokyr J (2000) Why 'More work for mother?' Knowledge and household behavior, 1870–1945. J Econ Hist 60(1):1–41
Mokyr J (2009) The enlightened economy: an economic history of Britain 1700–1850. Yale University Press, New Haven
Muldrew C (2012) 'Th'ancient Distaff' and 'Whirling Spindle': measuring the contribution of spinning to household earnings and the national economy in England, 1550–1770. Econ Hist Rev 65(2):498–526
Ogilvie S (2003) A bitter living: women, markets, and social capital in early modern Germany. Oxford University Press, Oxford
Owen L (2001) Gender differences in labor turnover and the development of internal labor markets in the United States during the 1920s. Enterp Soc 2(1):41–71

Phillips N (2006) Women in business, 1700–1850. Boydell Press, Woodbridge
Power E (1975) Medieval women. Cambridge University Press, Cambridge
Qian N (2008) Missing women and the price of tea in China: the effect of sex-specific earnings on sex imbalance. Q J Econ 123(3):1251–1285
Rose S (1992) Limited livelihoods: gender and class in nineteenth-century England. University of California Press, Berkeley
"Rural Queries", Report of His Majesty's Commissioners for Inquiry into the Administration and Practical Operation of the Poor Law, Appendix B, British Parliamentary Papers, 1834 (44) XXX
Schmidt A (2009) Women and guilds: corporations and female labour market participation in early modern Holland. Gend Hist 21(1):170–189
Simonton D (1991) Apprenticeship: training and gender in eighteenth-century England. In: Berg M (ed) Markets and manufactures in early industrial Europe. Routledge, London
van den Heuvel D (2007) Women and entrepreneurship: female traders in the northern Netherlands, c. 1580–1815. Aksant, Amsterdam
van Zanden JL (2009) The long road to the Industrial Revolution: the European economy in a global perspective, 1000–1800. Brill, Leiden/Boston
van Zanden JL (2011) The Malthusian Intermezzo: women's wages and human capital formation between the late Middle Ages and the demographic transition of the 19th century. Hist Fam 16(4):331–342
Vickery A (1993) Golden age to separate spheres? A review of the categories and chronology of English women's history. Hist J 36(2):383–414
Voigtländer N, Voth HJ (2012) Persecution perpetuated: the medieval origins of anti-Semitic violence in Nazi Germany. Q J Econ 127(3):1339–1392
Voigtländer N, Voth HJ (2013) How the west invented fertility restriction. Am Econ Rev 103(6):2227–2264
Xue M (2016) High-value work and the rise of women: the cotton revolution and gender equality in China. Unpublished

International Migration in the Atlantic Economy 1850–1940

Timothy J. Hatton and Zachary Ward

Contents

Abstract

This chapter focuses on the economic analysis of what has been called the age of mass migration, 1850–1913, and its aftermath up to 1940. This has captured the interest of generations of economic historians and is still a highly active area of research. Here we concentrate on migration from Europe to the New World as this is where the bulk of the literature lies. We provide an overview of this literature focusing on key topics: the determinants of migration, the development of immigration policy, immigrant selection and assimilation, and the economic effects of mass migration as well as its legacy through to the present day. We explain how what were once orthodoxies have been revisited and revised and how changes in our understanding have been influenced by advances in methodology,

T. J. Hatton (✉)
Department of Economics, University of Essex, Colchester, UK

Research School of Economics, Australian National University, Canberra, Australia
e-mail: hatton@essex.ac.uk

Z. Ward
Department of Economics, Hankamer School of Business, Baylor University, Waco, TX, USA
e-mail: zach.a.ward@gmail.com

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_41

which in turn have been made possible by the availability of new and more comprehensive data. Despite these advances, some issues remain contested or unresolved, and, true to cliometric tradition, we conclude by calling for more research.

Keywords
Mass migration · The Atlantic economy · Immigrants and emigrants

Introduction

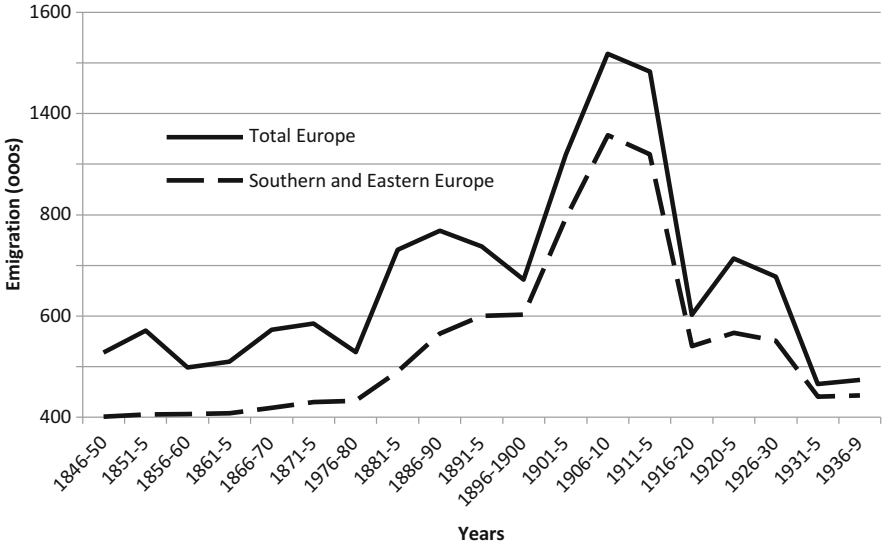
This chapter provides a survey of cliometric research on international migration. We focus on studies of migration from Europe during the age of mass migration from the middle of the nineteenth century up to 1913 and during the interwar de-globalization that followed. While there were other important migrations within Europe and Asia during this period, we focus on what might be called the "greater Atlantic economy" and especially on migration to the United States. This transformed the world economy, and it is where the vast bulk of the cliometric literature has focused attention. The literature has evolved over recent years with old interpretations being challenged as new insights have emerged, although some issues are still contested. The application of new approaches has strong parallels with the research of economists on more recent periods, and this has been underpinned by the emergence of new and more comprehensive data.

We start with an overview of the determinants of aggregate migration streams, a well- established literature that has been less active in recent years. This includes time series analysis of the push-pull variety as well as studies that focus on cross-sectional differences in emigration intensity between origin countries and regions. Migration flows were shaped by immigration policies that became increasingly restrictive in the early twentieth century. We review the literature on the formation of immigration policy, focusing on the political economy determinants. We then turn to the question of whether international migrants were positively or negatively selected from source-country populations. This micro-level analysis, based on unit record data, has been the focus of much of the recent literature and has challenged the conventional wisdom that migrants were positively selected. The self-selection of migrants has implications for the old debate about immigrant assimilation, and we review recent research that marks a return to the more pessimistic view of the older literature. We then turn to the economic effects of immigration in New World destinations as well as the effects of emigration on the Old World. The traditional concerns about the effects on wages and living standards have been examined using a variety of methods that yield a range of different results. Finally we briefly review a recent literature that seeks to explain economic outcomes in the present as the very long-run consequences of the age of mass migration.

Determinants of International Migration

From 1850 to 1940, an estimated 50 million migrants traveled from Europe to the New World (Fig. 1), with about three fifths going to the United States and smaller streams heading for South America and the British Dominions. There is a long-standing interest in what determined the year-to-year movements in migration, a literature that goes back to work by Jerome (1926) and Thomas (1941). This literature sought to explain cycles in transatlantic migration using time series models that focused on variations in economic conditions in source and destination countries. The "push" versus "pull" literature that flourished in the 1960s and 1970s was critically reviewed by Gould (1979). As he pointed out, when cyclical indicators such as industrial production are included for both source and destination countries, the latter typically dominate in terms of the size and significance of the coefficients. On the whole, these studies suggested that short-term ups and downs in aggregate migration were dominated by pull rather than by push factors. And in their presence, other variables representing the incentive to migrate, such as income or wages at home and abroad, often failed to achieve significance. Gould (1979, p. 668) concluded that "By and large, reaction to this literature is one of some disappointment, for not only has it failed to generate important new insights . . . it has had only limited success in confirming or denying old interpretations."

It is difficult to believe that emigration decisions were determined exclusively by conditions in the destination or that they did not depend on the prospects of higher



Source: Kirk (1946), p. 297.

Fig. 1 Emigration from Europe 1846–1939 (five-year averages). (Source: Kirk (1946), p. 297)

wages or incomes. Emigration must have been determined by some assessment of the expected gain and hence by some comparison, however approximate, between conditions at home and abroad. But the first generation of studies lacked a coherent underlying model of the emigration decision, which made the coefficients difficult to interpret. Using a basic economic framework, Hatton (1995) developed a model in which representative potential migrants base their decisions on the comparison of future expected incomes at home and abroad. In the manner of Todaro (1969), expected income in a location depends on both the wage rate and the probability of employment. The latter accounts for the effects of the business cycle and, if migrants are risk-averse and there is greater uncertainty about conditions abroad than at home, that would help to explain the dominance of cyclical conditions, especially at the destination. In addition, the time series dynamics that were often included to soak up serial correlation can be accounted for by adaptive expectation formation.

A number of studies have used this framework or can be interpreted within it. Estimates for emigration from the United Kingdom from 1870 to 1913 show that the wage gap and employment rates at home and abroad all matter in a manner consistent with the model (Hatton 1995). While short-run fluctuations are largely accounted for by the business cycle, longer-term trends can be accounted for by the slowly changing income gaps between origin and destination countries. For example, the 17% fall in the ratio of US to Irish wage rates between 1876–1880 and 1909–1913 accounted for a long-run decline in emigration from Ireland of 4 per thousand of the population (Hatton and Williamson 1998, p. 83). Another important variable that features in these models is the stock of previous migrants at the destination. This strong pull effect of emigrant networks or diasporas is often interpreted as reducing the costs and risks for new migrants. In the Irish case, the declining ratio of emigrant stock to home population reduced emigration by another 4 per thousand. New migrants were often assisted through remittances or prepaid tickets sent by previous emigrants. Consistent with this, Magee and Thompson (2006) find that the flow of money orders to the United Kingdom from 1880 to 1913 depended on both the migrant stock and the average income of the English-speaking destination countries.

The direct costs of passage should also matter, but these have been absent from most econometric studies for lack of suitable data. The cost of sea voyage fell after 1850 with the shift from sail to steam and the effective price fell even more steeply with shorter passage times (Keeling 1999; Cohn 2005; Sánchez-Alonso 2007), not to mention improvements in overland travel to and from seaports. On the North Atlantic routes, sharp swings in ticket prices were associated with changes in the effectiveness of shipping cartels. Using quarterly data for 1899–1913, Deltas et al. (2008) find that passenger numbers on routes to the United States and Canada were 22% lower at times when cartels were keeping prices up. Other elements of costs also mattered. Assisted passages to Australia substantially reduced the costs for UK emigrants in 1911–1913, leading to a surge of emigration (Pope 1981). For Spanish emigrants, the depreciation of the peseta between 1882 and 1905 increased the travel costs of emigration from Spain, which had the opposite effect. Sánchez-Alonso (2000a) finds that this reduced emigration during those years by as much as 30%.

Time series models account fairly well for the ups and downs of well-established migration streams, and perhaps this is one reason that they have not been the focus of much of the recent literature. A greater challenge is to account for the initial rise of emigration that got these streams started and the differences in emigration intensity between origin countries and localities. Among Western European countries, annual average gross emigration rates varied widely from 12 per thousand for Ireland to less than 1 per thousand for France. The trends also differed. The Irish emigration rate declined from the 1860s, and those of Germany, Norway, and Sweden fell from the 1880s. From that time, emigration from Italy and Spain and from some countries in Eastern Europe increased steeply up to the outbreak of war in 1914. One stylized fact is that, during the onset of modern economic growth, emigration rates often increased from low levels, reaching a peak after some decades and then gently declining. This long swing in migration, though often not complete, has been identified for at least some European countries (Akerman 1976; Hatton and Williamson 1998, Chap. 3).

Several studies suggest that demographic trends were important. The present value of migration would be greatest for those with the longest time horizons and the lowest costs: the young and single. Ellis Island records show that the average age of immigrant arrivals to the United States in 1900–1910 was 26, two thirds were male, and the majority were single (Bandiera et al. 2013). Using decade-average emigration rates for 12 European countries, Hatton and Williamson (1998, Chap. 3) found that the birth rate lagged two decades had a large positive effect. Annual time series for Scandinavian countries also supports this view (Quigley 1972; Larsen 1982; Hatton and Williamson 1998, Chap. 4). Interestingly, Greenwood (2007) finds that a higher current birth rate tended to reduce the emigration rate for those of parenting age, probably reflecting the higher costs of family migration. However, demographic effects were weaker for countries such as Italy and Spain (Sánchez-Alonso 2000b; Hatton and Williamson 1998, p. 113) because other constraints mattered more.

Emigration often gathered momentum just when the pace of development was quickening at home. In the early stages, most migrants were too poor to migrate, even though the incentive was large, but economic development served to ease the "poverty constraint." Thus on the upside of the long swing, emigration grew as wages at home increased, while on the downside, further increases in the home wage reduced the incentive to emigrate. Using annual time series from the 1870s to 1913, Faini and Venturini (1994) find for Italy, and Sánchez-Alonso (2000b) finds for Spain, that rising income per capita at home positively influenced emigration. The evidence also suggests that poverty constraints became less binding as the stock of previous emigrants, who could provide assistance, increased (Hatton and Williamson 2005, p. 65). This interaction helps to explain why emigration from Ireland (with its large post-famine emigrant stock) declined while emigration from equally poor Italy (with a low initial stock) increased as development gathered pace. Bohlin and Eurenius (2010) find a positive effect of the interaction between poverty and the emigrant stock in a panel of emigration from Swedish counties 1881–1910. Poverty constraints are also observable at the micro-level. In the Hesse-Cassel region of Germany from 1832 to 1857, Wegge (1998) found that networked emigrants

carried less cash with them. By the 1920s, these effects were weaker, but Armstrong and Lewis (2017) provide evidence that the need to save served to delay emigration to Canada.

There were many political and economic events that led immigrants to seek refuge during the age of mass migration, including the Great Irish Famine in the late 1840s, failed political revolutions in 1848, the Swedish and Finnish famine of 1866 to 1868, the Mexican Revolution of the 1910s, and the religious persecution of Jews in the 1880s, the early 1900s, and the interwar period. Mokyr and Ó Gráda (1982) and Cohn (1995) examine immigrant flows during the Great Famine, and Boustan (2007) and Spitzer (2014) study Jewish emigration. While short-term shocks boosted emigration in the short-term and then subsided, their effects often persisted through the powerful effect of the stock of previous migrants on subsequent flows.

Less attention has been paid to how potential migrants chose among alternative destinations in the New World. Differences between destinations in the pre-existing migrant stock, colonial ties, common language, and cultural affinity meant that they were often poor substitutes (Taylor 1994). But there was greater scope for choice among English-speaking countries, as between the United States and Canada (Green et al. 2002) and also within South America. Balderas and Greenwood (2010) examine the effect of emigration to one destination on emigration to other destinations using annual time series for emigration from 12 source countries to 3 destinations. Using instrumental variables, they find evidence of substitution between Argentina and Brazil, but not between either of these and the United States. The political climate in host countries could also matter. Bertocchi and Strozzi (2008) examine the effects of political institutions on migration between 11 European countries and 3 New World countries (Argentina, Canada, and the United States) on decade-average data for 1870–1910. In addition to economic and demographic variables, they find that immigration was positively related to political participation (democracy and suffrage) as well as to rights for immigrants (access to citizenship, land, and education).

Many migrants retuned, often after just a few years. The conventional estimates indicate that the ratio of out-migrants to in-migrants was about 40% (Gould 1980; Kuznets and Rubin 1954). However, recent estimates by Bandiera et al. (2013) put the figure much higher. Using the Ellis Island records, they calculate that gross immigration from 1900 to 1910 was about 20% higher than the official figures report. Comparing this inflow with the change in the stock of immigrants from the census implies that outflows were 60% of inflows, as compared with the previously accepted figure of 40%. The same calculation for the following decade, when the immigrant stock hardly changed, yields an outflow-inflow ratio of 75–80%. Some of these migrants may have been short-term visitors, but, even so, high return rates call into question the treatment of migration as a once-and-for-all decision, suggesting instead a significant share of circular migration. Indeed, Ward (2016) shows that between 1897 and 1914 repeat entrants (foreign-born who had previously been in the United States) accounted for 10–20% of all foreign-born arrivals. These were not mainly the *golondrinas* or seasonal workers from Southern Europe so often referred

to in the literature: repeat entrants were instead often from Northern and Western Europe and were higher skilled, suggesting that the circular flow was not dominated by low-skilled laborers.

Immigration Policy

The era before the World War I is seen as one of free migration in the greater Atlantic economy – but only for some migrants. The United States severely restricted immigration from China in 1882 and from Japan in 1907. Immigration from the poorest parts of the world was unrestricted only because there was no pressing "threat" of immigration. In 1885 Canada introduced an entry tax on Chinese immigrants, which was raised in subsequent steps. Similar policies were adopted in the Australian colonies and in 1901 newly federated Australia followed the 1897 policy of Cape Colony and Natal by introducing a dictation test designed to keep out non-Europeans. This was adopted by New Zealand in 1907 and by Canada in 1910. Other more incremental changes included entry taxes and restrictions on criminals, lunatics, and those "liable to become a public charge."[1] And the positive inducements to immigration offered by some countries such as Argentina, Australia, and Brazil were eventually reduced or withdrawn.

In the face of continuing anti-immigration pressure, more radical restrictions were introduced from World War I onward. In 1917, the United States introduced a literacy test for all immigrants and a ban on immigration from the "Asiatic Barred Zone." This was shortly followed by the first numerical quota in 1921, which became tighter in the Act of 1924 and its amendment in 1929. As illustrated in Fig. 1, international migration was dramatically lower in the 1920s as compared with the decades before 1914, and much of this was due to policy. The decline in immigration was most dramatic in the United States; for the source countries covered by the 1921 quota, immigration fell by two thirds in the following year. By linking them to the source composition of the foreign-born in the past, the quotas bore heavily on the so-called "new immigrant" countries, as illustrated in Table 1. Although the sharp shift toward restriction in the United States has received most of the attention, other countries followed. In 1923, Canada introduced a formal distinction between immigrants from preferred and non-preferred countries. Escalating restrictions were adopted in South Africa and Brazil, culminating in quota systems in 1930 and 1934, respectively. Even the British Dominions adopted severe restrictions limiting immigration from Britain: Australia in 1930, New Zealand in 1931, and Canada in 1932.

The retreat from free migration is sometimes seen as a backlash against mounting numbers (Williamson 1998). But what were the political economy mechanisms? As Foreman-Peck (1992, p. 360) put it: "The two key questions of any political

[1]For example, the US Immigration Act of 1882 introduced a head tax (extended in 1891); the Act of 1903 imposed restrictions on anarchists, beggars, those with epilepsy, and importers of prostitutes.

Table 1 Prewar immigration and postwar quotas

	Annual immigration	Quota allocation		
	1908–1914	1921	1924	1929
Belgium	5186	1563	512	517
Denmark	6117	5694	2789	1181
France	8209	5729	3954	3086
Germany	31,292	68,059	51,227	25,957
Ireland	27,571	–	28,567	17,853
Italy	202,222	42,057	3845	5802
Netherlands	6625	3607	1648	3153
Norway	11,874	12,202	6453	2377
Portugal	9166	2520	503	440
Spain	5021	912	131	252
Sweden	16,642	20,042	9561	3314
UK (includes Ireland in 1921)	42,658	77,342	34,007	65,721

Source: Greenwood and Ward (2015), p. 79

economy of international migration are: (1) who gains and who loses from migration? And (2) who is in a position to do something about it?" The answer to the first question depends on who was most likely to face direct competition from immigrants in the labor market. As immigration from Southern and Eastern Europe expanded, the immigrants were increasingly low-skilled workers. In the 1860s, the average immigrant to the United States came from a country with GDP per capita 95% of the United States. By the 1900s, the ratio had fallen to 50%, and similar declines occurred for immigration to Canada and Argentina (Hatton and Williamson 2007, p. 223). On the other hand, the owners of capital and land, and perhaps the more skilled and educated workers, were less likely to favor restriction.

In order to test for these effects, Timmer and Williamson (1998) developed an annual index of immigration policies for five New World countries for 1870–1930. They found that for Argentina, Brazil, Canada, and the United States, more restrictive policies were associated with declining relative unskilled wages. In Canada and the United States, where immigrant origins were more diverse, it was the rise of immigration from low-wage countries, differing in ethnicity and religion from earlier waves, which helped to close the door. For the United States, this is consistent with the imposition of quotas that radically reduced immigration from Southern and Eastern Europe. In Argentina and Brazil, where immigrant origins were less diverse, the rising share of all foreign-born also led to restriction.

While immigration policies typically get tighter in recessions (Shughart et al. 1986), recessions seem to have more decisive effects when they are preceded by a gradual accumulation of forces that shift opinion against immigration. This may help to explain the imposition of the emergency quota in the United States, as the postwar resurgence of immigration coincided with an increase in the unemployment rate from 5.2% in 1920 to 11.7 in 1921. While the United States led in the early 1920s, the door was closed in other New World countries as unemployment increased

during the Great Depression. But another factor may also have been important. The interwar period saw a dramatic decline in international capital mobility. So the effect of capital inflows in muting the wage effects of immigration would have been smaller than before 1914 (see further below). For this reason, workers could have been more opposed to immigration than in the era when capital chased labor (Hatton and Williamson 2007).

At first sight, the answer to Foreman-Peck's second question depends on who had the right to vote. As suffrage widened, it typically percolated down the hierarchy of class and income, diluting the political weight of landowners and capitalists and giving a stronger voice to urban blue-collar workers. At the turn of the century, voting rates were about one third of the adult population in North America (much higher for males) but less than 10% in Latin America (Engerman and Sokoloff 2005). Severe immigration restrictions came earlier in the United States than in the (comparably democratic) British Dominions, which chose to remain open, at least to immigrants from the United Kingdom. On the other hand, in independent Latin America, where the franchise was much narrower and excluded immigrants, and where the Latifundia retained their grip on power, restriction came earlier than might have been expected. Sánchez-Alonso (2013) analyzed a newly created index of immigration policy for Argentina from 1870 to 1930. She found that some of the shift toward restriction was accounted for by rising immigration, declining migrant education relative to natives, and growing inequality. But restrictive immigration policies were instigated not through the ballot box but as a result of strikes and protracted labor unrest.

The underlying political economy mechanisms have been explored in the greatest detail for the United States. Turning to the attitudes of workers, Richardson (2005) analyzed the opinions of 2,000 blue-collar workers surveyed by the Kansas State Labor Bureau in 1895–1897. He found that the vast majority wanted immigration to be restricted or to be completely suppressed. Opposition to immigration was stronger among union members than among non-unionists; interestingly those with incomes in the middle of the range were more opposed than those with the lowest wages. Not surprisingly, the first- or second-generation foreign-born were the least likely to favor restriction. There was also some evidence that opposition to immigration was greater in counties with strong growth in the immigrant population.

How did this anti-immigrant sentiment get translated into restrictive immigration policy? In an influential paper, Goldin (1994) examined voting patterns in the US Congress. Beginning in 1897, a series of bills incorporating a literacy test failed to pass into law until 1917 when the House of Representatives and the Senate both overrode President Wilson's veto. By 1906, representatives from the South had joined those from other rural areas in supporting restrictions, and the key battleground was the cities of the East and the Midwest. Goldin found that a representative was more likely to vote for an override in 1915 the more rapid the growth in the foreign-born population and the slower the growth of wage rates in their city in the preceding decade. But the higher the immigrant share in the district (the level, not the change), the less likely was a vote for restriction and especially when the share reached 30%. Looking across cities from 1910 to 1930, Tabellini (2017) finds that

greater immigrant inflows, especially from "new" source countries, reduced electoral support for the Democrats, led to the election of more conservative congressmen and increased the likelihood that they would vote for the quota act of 1924.

Biavaschi and Facchini (2017) analyze the House votes on all 14 immigration bills from 1897 to 1924, estimating with fixed effects by session and by congressional district. They find that a representative was less likely to vote for restriction the greater the share of employment in manufacturing and the lower the share urban in the district. The characteristics of the Representative mattered too, with northern democrats and those with non-Ivy League backgrounds more likely to vote for restriction. Interestingly they also find that, while higher shares of foreign-born made a vote for restriction less likely, this was tempered where state residency laws made it more difficult for immigrants to naturalize and to gain the vote. This is consistent with the findings of Shertzer (2016) that the rate of naturalization was greater where immigrant interests aligned with other sizeable minorities to form a winning coalition.

Contemporary controversies over immigration policy have stimulated renewed interest in the shaping of immigration policies in the past. Recent cliometric studies have uncovered a range of influences that are broadly consistent with the political economy of immigration policy mapped out in more qualitative accounts. Working through interest group politics, immigration itself has mixed effects. While a sizeable and well-established stock of past immigrants tends to favor more liberal immigration policy, the recent inflow has the opposite effect. It also seems likely that under certain conditions, shocks such as wars and recessions trigger an immigration backlash. As most of the evidence is for the United States, it is less clear how these influences played out in countries with different political and institutional settings.

Immigrant Selection

Who migrates and how they compare with those left behind is a long-standing question in the literature. The selection of immigrants is related to the fortunes of immigrants in the host-country labor market, immigration's impact on the wages and incomes of the native-born, and to the longer-run effects of immigration – topics explored further in later sections. The selection of immigrants can be measured in a variety of ways: one dimension is the origin-country composition, which, as we have seen, shifted toward the poorer parts of Europe, raising the concerns of policy makers. Another is the demographic profile; immigration to the United States increasingly comprised young unattached males – those with low dependency burdens and high labor force participation rates. But the focus of much of the recent cliometric literature is on selection by skill or ability.

The basic framework is the Roy model as adapted to migration by Borjas (1987). In this model, the incentive to migrate depends on the individual's skill level and the returns to skills in the home country compared with the destination country. If the return to skills is greater at the destination, then the skilled will be more likely to

emigrate; if the return to skills is higher at the source, then the unskilled will be more likely to emigrate. But the costs of migration must also be considered; higher costs of migration will reduce the net present value of migrating more for poorer, less skilled workers than for those with higher skills and incomes. This framework is particularly applicable to the age of mass migration, an era before the advent of restrictive immigration policies, and a time when most migrants were moving for economic reasons rather than fleeing famines, wars, or political persecution. If the postbellum United States became more unequal than European sources, then that would favor more positive selection from any given source country. On the other hand, declining transport costs and, above all, assistance from the growing stock of previous emigrants would work in the opposite direction.

Testing different theories on the selection of immigrants first requires measuring selection, which can be difficult for numerous reasons. First, measures of immigrant quality in historical data are often crude and rarely include key metrics of interest such as ability, entrepreneurial spirit, or attitude toward risk. One way to get a summary measure of a migrant's productivity is by using his wage, but since the researcher does not observe the immigrant's wages in the source country if he had stayed, we need to either measure selection on a premigration variable or estimate the counterfactual wage. Despite these methodological problems, there have been significant advances in the literature due to the recent digitization of census and immigrant arrival records. Arrival records document immigrants prior to any assimilation forces causing a change in occupation or other observables, which allows for a (mostly) straightforward comparison of immigrants relative to stayers as observed in source-country census data. However, the quality of information in ship records may be low if ship captains were careless filling them out, and variables in the arrival records may not always match with variables in census data. Instead, some researchers have linked censuses across countries, allowing us to observe the immigrant's characteristics before and after migration.

The predominant finding of studies using arrival records or census data is that European immigrants to the United States during the age of mass migration were intermediately or negatively selected. Using individual census data on Norwegian immigrants to the United States in the late nineteenth century, Abramitzky et al. (2013) find that they were negatively selected as the fathers of movers had less wealth than the fathers of stayers. Negative selection has also been found for Irish emigrants prior to 1850, based on the age-heaping reported in arrival records and in the Irish Census (Mokyr and Ó Gráda 1982); Irish immigrants to the United States were also negatively selected in the early twentieth century based on linked census records (Connor 2016). Italian emigrants in the early twentieth century were shorter than the average Italian left behind, suggesting that they had a worse health environment than stayers (Spitzer and Zimran 2017).

Most studies measure migrant selection on observable characteristics, but the selection of immigrants on unobservable characteristics is more difficult to gauge. Furthermore, it is unclear whether selection on observable and unobservable characteristics is correlated; it could be that a low-skilled immigrant from an observably poor family was also highly ambitious and entrepreneurial. Abramitzky et al. (2012)

use an innovative approach to measure selection on unobservables: they compare the return to immigration measured with and without brothers' fixed effects. The idea is that a naïve estimate of the return to immigration by comparing migrants' occupations to stayers' occupations is biased if immigration is correlated with a positive or negative omitted variable; however, one can gauge the direction of this bias by estimating the return to immigration within brothers to control for unobservables across households. Abramitzky et al. (2012) find that the naïve return to immigration is more negative than the estimate within brothers and interpret the negative bias on the naïve estimate as reflective of negative selection into immigration. The negative selection on unobservable characteristics is consistent with other findings of negative selection on observable characteristics such as father's wealth, perhaps suggesting that the observable measures of selection are a good proxy for unobservable selection.

Can these findings of mostly intermediate or negative selection be rationalized in the context of the Roy model? Stolz and Baten (2012) test the Roy model on data from the age of mass migration by comparing measures of inequality with measures of selection. They use measures of inequality and selection that are well known within the cliometric literature: for inequality, the amount of variation in height data and, for selection, the amount of age-heaping on zeroes or fives for immigrants relative to those remaining in the source (Steckel 1995, 2008; Tollnek and Baten 2016). The coefficient on relative inequality across sources offers strong support for the Roy model: migrants to a destination from more equal origins were more negatively selected on numeracy. They also find that sharing a common language is associated with positive selection while wars and upheavals at the origin lead to negative selectivity. This latter effect may also apply to famines, which would help explain the negative selection of emigrants from mid-century Ireland. However, neither the predicted positive effect of origin-country poverty nor the negative friends and relatives effect could be consistently identified.

The effect of liquidity or wealth constraints on migrant selectivity is something of a puzzle, but it seems likely to have become less important over the nineteenth century, if for no other reason than assistance from the growing stock of previous migrants. For example, in 1908, 92% of immigrant arrivals to the United States were joining a relative or friend at arrival and about one third had their tickets paid for by someone else. Therefore, while on the margin poverty constraints did limit some migrants, they were not binding for many. However, wealth constraints were likely stronger during the earlier stages of the mass migration, when networks were less developed and prior to cost reductions due to the diffusion of steam technology following the Civil War (Cohn 2009). As noted earlier, Wegge (1998) shows that networked migrants from the Hesse-Cassel region in mid-century Germany required less cash to move, implying that they depended on help from previous migrants. And even though Irish emigrants were negatively selected overall, during the Great Famine the very poorest were less able to emigrate and were more likely to have starved (Mokyr and Ó Gráda 1982; Cohn 1995). There is other evidence that poverty constraints limited migration from the lower end of the skill distribution. While Italian immigrants were negatively selected overall, they were positively selected

from the poorest provinces, suggesting that costs restricted migration of the poorest (Spitzer and Zimran 2017).

Focusing on the drivers of immigrant skill levels, Covarrubias et al. (2015) use US inflow data from 1899 to 1932 to determine the relative importance of source-country income and costs of immigration on the skill composition of inflows. They find that higher origin-country income levels increased the total number of migrants but slightly decreased their quality, suggesting that a poverty constraint was alleviated. On the other hand, transport costs had little effect on the skill level of the inflow, perhaps because costs are represented by freight rates rather than by passenger fares. The severe restriction on immigration imposed from the early 1920s caused the selection of immigrants to be more positive, at least in terms of occupation, as the implied policy costs fell more heavily on the lower skilled. Therefore, not only did the immigration quotas have a between-country effect as the quotas imposed different degrees of restriction across origin countries, but there was also a within-country effect as lower-skilled individuals were more likely to be restricted (Covarrubias et al. 2015; Massey 2016).

Perhaps what policy makers and economists are more interested in is the quality of the net flow rather than just the inflow. As we have seen, since millions of migrants returned to Europe, the quality of the net flow depends on the selectivity of return migration. The evidence suggests that the average return migrant was lower-skilled than the average permanent immigrant, a pattern that holds both within and between source countries (Abramitzky et al. 2014; Ward 2017a).[2] The negative selection of return migrants presents something of a puzzle as the Roy model predicts that the selectivity of the return flow should be opposite to the selectivity of the inflow (Borjas and Bratsberg 1996), which was also negatively selected. This may be because those who returned home were more likely to receive a negative shock in the US labor market, perhaps due to occupational downgrading after arrival.

The recent literature has challenged the view often presented in qualitative accounts that immigrants into the United States were, on the whole, positively selected. Instead the evidence suggests that immigrants overall were somewhat negatively selected but with variations by country and region of origin. So far, there is little comparable evidence for other New World destinations during the age of mass migration, but there is reason to think that the results could be different. For example, Italian migrants to migrants to Argentina had relatively high literacy rates, especially those from Spain and Portugal (Sánchez-Alonso 2007). And the occupational composition of emigrants to Australia 1877–1913 was more skilled than that of Canada and the United States (Pope and Withers 1994).

[2]Abramitzky et al. (2016b) provide new microdata on returners to Norway and show that despite return migrants being negatively selected relative to both movers and non-movers, they achieved higher outcomes upon return to Norway, reflecting a positive return to temporary migration.

Immigrant Assimilation

These new findings on immigrant selectivity raise questions about other widely debated issues. Following (Borjas 1987), the selectivity of immigrants has been intimately connected with their assimilation into the US labor market. The assimilation of immigrants tackles a similar question to the selection literature but from a different reference point. Instead of comparing immigrants to nonmigrants in the source country, the assimilation literature compares immigrants to the native-born in the destination country. This comparison is most often for wages or occupational status but has also been made along different dimensions such as geographical location, fertility, or marriage rates. In this section, we will primarily focus on standard economic outcomes such as occupation, wages, and employment.

The classic finding in the literature is that for the post-1950 decades, immigrants initially received lower wages than natives and then converged to native wages in the decades after arrival (Chiswick 1978). This finding is often interpreted in a human capital framework: at arrival, immigrants lose some value of their human capital acquired at home since it does not transfer across borders perfectly, but then they gain host-country-specific human capital, such as language fluency, after arrival. Another reason for convergence between immigrants and natives (which links to the selection issue) is that immigrants are thought to have different abilities or ambition and drive than the native-born, which implies that they would upgrade income throughout the life cycle faster than the native-born. However this optimistic view of immigrant assimilation has been downgraded in the subsequent literature that takes account of cohort effects and selective return migration (Borjas 1985; Lubotsky 2007).

Immigrant assimilation in the late-nineteenth-century US labor market has been explored using data on the earnings of immigrants and native-born from a range of reports from state labor bureaus. The initial results seemed to support the revisionist view that the age-earning profiles of immigrants were no steeper than those of the native-born and that, over the life cycle, immigrants may even have fallen further behind their native-born peers (see for example Hanes 1996). By contrast, Hatton (1997) showed that the results turn partly on the shape of the age-earning profile combined with the fact that the immigrants in these samples are typically older than the native-born. Taking this into account, immigrants that arrived before the age of sixteen had age-earning profiles very similar to the native-born. Those that arrived as adults had earnings that were initially lower and that increased with age, but did not overtake the native-born. However these results stem from cross-sectional analysis of subsets of blue-collar workers, and they do not account for cohort effects or for selective return migration.[3]

[3]However, analysis of grouped data by origin country from the reports of the US Immigration Commission and of cohort occupational progression in the 1900 and 1910 censuses reaches similar conclusions (Hatton 2000; Minns 2000). For a more detailed review of this generation of studies, see Hatton (2011).

This view has now been subjected to further revision. Abramitzky et al. (2014) find that the average immigrant arrived similarly skilled and upgraded at the same rate as natives. The difference in results is due to data structure: both cross sections and repeated cross sections, as were used in the earlier literature, overstate immigrants' assimilation rate because of declining cohort quality and negative selection of return migrants. Abramitzky et al. (2014) improve on repeated cross sections with new panel data by linking immigrants across the 1900–1920 US Censuses. Since the immigrants in the panel data are the same ones throughout time – as opposed to pseudo-cohorts used in the repeated cross-sectional methodology – selective return migration does not bias the estimates. Abramitzky et al. (2014) recreate the finding of positive assimilation using repeated cross sections but then demonstrate that this pattern disappears once using the panel data, which implies that the positive rate of assimilation in earlier studies was primarily driven by negatively selected return migrants.[4]

One limitation of these results is that the 1900–1920 censuses do not record wages, but only occupational status; it is possible that immigrant assimilation profiles were steeper for earnings. But a lack of relative occupational upgrading after arrival is still surprising given the human capital model of assimilation, suggesting either that post-migration human capital was not valuable or that immigrants were negatively selected on ambition or drive relative to native-born Americans. Ward (2018) explores the importance of one piece of post-migration human capital that is highly valuable today – English fluency – and finds that acquisition of English skills was associated with only a small upgrade in occupation. While the reason for a low occupational return to English skill is unclear, it may be related to the structure of the early twentieth century economy: the skill premium was relatively low, and jobs which required communication skills were not as prevalent.

Despite a low occupational return to English fluency, people at the time were convinced of its importance (Jenks and Lauck 1926). This was a contributing factor to the "Americanization" movement of the 1910s and 1920s, which aimed to assimilate immigrants into American society typically through education programs (United States Immigration Commission 1911; Bandiera et al. 2018). However, it is unclear how effective the Americanization movement was: Lleras-Muney and Shertzer (2015) argue that compulsory schooling and English-only laws were ineffective for raising immigrants' income in 1940. Moreover, these laws also led to a backlash from the targeted population. For example, German Americans in states that banned German instruction were more likely to give their children German names and less likely to volunteer in World War II (Fouka 2015).[5] While

[4]Abramitzky et al. (2014) focus on European arrivals given their importance in the flow. For non-European sources, Kosack and Ward (2018) estimate the assimilation rates of Mexican immigrants, and Hilger (2016) estimates outcomes for descendants of Asian immigrants.

[5]This behavior of resisting assimilation is also explored for French Americans in New England by MacKinnon and Parent (2012).

programs aiming to assimilate immigrants may not have been effective, there is evidence that the appearance of being American was economically important: Biavaschi et al. (2017) show that adopting a more American-sounding name yielded a large occupational-based return, and Abramitzky et al. (2016a) show that the second generation with less foreign-sounding names had higher incomes relative to their brothers with more foreign-sounding names (see also Goldstein and Stecklov (2016) and Carneiro et al. (2015)); moreover, becoming a citizen was associated with improved occupational status (Catron 2017). This is consistent with Alexander and Ward (2018) showing that younger-arriving brothers were both more socially assimilated and had a smaller wage gap with natives compared with their older-arriving brothers.

The assimilation literature for the United States is limited in that the US Census does not record incomes until 1940; therefore, while immigrants had similarly paid occupations as white natives, they may not have had similar incomes. Fortunately, the 1911–1931 Canadian Censuses did measure income, which Inwood et al. (2016) use to show that immigrants upgraded at faster rates than natives through the life cycle – a result that differs from the lack of occupational-based assimilation in the United States. The Canadian result is consistent with a positive return to Canadian-specific human capital and a positive return to English skills. Yet this result must be interpreted with caution because the Canadian random samples form a repeated cross section, making it unclear whether their assimilation result is due to true upgrading or selective return migration.

In contrast to the lack of upward occupational mobility for immigrants relative to natives in the United States, there was more success for Europeans who headed to South America. Using data that links immigrants and natives across censuses, Pérez (2017) shows that immigrants to Argentina upgraded their occupations at faster rates than natives, ending up higher skilled than natives after decades of stay. Importantly, this pattern exists when tracking the same individual over time, which avoids the problem that selective return migration may bias rates of upward mobility. This relative advantage for the first generation also persisted to the second generation of Europeans, who were more likely to be literate and own property than the sons of natives.

This persistence of occupational outcomes across generations of immigrants has been shown not only in Argentina but also in the United States (Abramitzky et al. 2014); moreover, it contrasts with the "melting pot" analogy which suggests that parents' or grandparents' country of birth does not strongly predict outcomes. Ward (2017b) extends the intergenerational literature beyond the second generation to the third generation and shows that immigrants' grandchildren in 1940 tended to be in the same occupational categories as the first generation in 1880. Mean skill level by country of origin persisted more strongly than would be predicted from a standard multigenerational model, suggesting that it took more than three generations for occupational differences across source countries to disappear. Therefore, while the melting pot may have been hot in terms of social assimilation, strong persistence in occupational outcomes meant that it was cooler for economic assimilation (Abramitzky et al. 2016a).

The Effects of Migration

One of the most contentious issues in the immigration debate has been the effect of immigration on the economy in general and on the wages and incomes of non-immigrants in particular. A variety of methods have been used to estimate the impact of migration on average wages in origin and destination countries during the age of mass migration. Taylor and Williamson (1997) estimate the effects of international migration from 1870 to 2010 for 5 New World countries and 12 Old World countries. They first calculate the cumulative effects of immigration or emigration on the country's labor force and then use labor demand elasticities to estimate the long-term effects on real wages. For the New World, they find that, in the absence of immigration after 1870, real wages would have been higher in 1910 by 9% in the United States, 17% in Australia, and 27% in Argentina. Emigration from European countries increased real wages relative to the counterfactual but by smaller amounts, corresponding to the more modest effect of emigration on labor force growth. According to these estimates, the real wage gap between the New and Old Worlds fell by 11%, whereas under the no-migration counterfactual, it would have increased by 10%.

Some studies have allowed for a richer set of adjustments by using multi-sector computable general equilibrium models that include three factors of production (land, labor, and capital) and also account for international trade. To give one example, O'Rourke et al. (1994) estimated that, in the absence of international migration after 1851, urban wages would have been higher in 1910 by as much as 34.0% in the United States and lower by 12.2% in Britain.[6] To the extent that factor intensities become more similar, migration tends to be a substitute for trade and so trade grows less rapidly than it would have in the absence of migration. But more important is the assumption that is made about international capital mobility. With perfect capital mobility rather than totally inelastic capital (as assumed above), the effects of immigration on wages are much smaller as capital chases labor to exploit the incipient increase in its returns. Thus with full capital mobility, the no-migration counterfactual US wage in 1910 would have been just 9.2% higher (as less capital flowed in) and the UK wage would have been 6.6% lower than the actual (as capital stayed at home).

Exercises such as these hinge on assumptions about the structure of the economy, in particular the assumption of diminishing returns to labor and therefore downward sloping labor demand curves. In the economic literature, attempts to test these effects have met with mixed results and have led to substantial disagreement about how to model and estimate the labor market impacts of inflows within an economy. Various approaches have been taken, including exploiting the variation in skill mix of inflows at a national level (Borjas 2003), the variation in the number of immigrants across geographical regions (Altonji and Card 1991), or a combination of skill mix and region (Card 2001). More recently, Ottaviano and Peri (2012) assume that immigrants and natives are imperfect substitutes even within skill groups, which

[6]Other studies in this vein are reviewed in more detail by Hatton (2011).

tends to reduce the direct negative effect of immigrants on the wages of comparable natives as estimated by Borjas (2003).[7]

Rather than directly estimating the effect of immigration on native outcomes, several studies explore the effect of immigration on average earnings overall (including both natives and immigrants). Goldin (1994) uses average wages by occupation and city to demonstrate that a 1% point increase in the foreign-born share led to a 1–1.5% decrease in wages for unskilled workers – an effect that is larger than one would expect purely from a compositional shift to more immigrants. Examining state-level time series from 1929 to 1957, Biavaschi (2013) finds that immigrant inflows had a negative effect on earnings, while outflows had a positive effect. Exploiting the reduction of inflows from the 1920s immigration quotas, Xie (2017) finds that a one percentage point decline in the foreign-born share at the county level increases manufacturing wages by 2% – once again, an effect that cannot be explained by compositional changes. Studies of national time series have estimated that the effect of mass migration was to widen skill premiums in the United States and narrow them in Europe (Anderson 2001; Betrán and Pons 2004), results that are consistent with the effects on inequality found in general equilibrium models (Betrán et al. 2010).

Some studies have followed the modern-day literature in seeking to identify the effect of immigration on the earnings of natives (rather than on average earnings) using microdata. Green and Green (2016) use microdata on earnings by occupation from the Canadian Censuses of 1911–1931 to estimate the impact of immigration on natives, following the methodology of Ottaviano and Peri (2012). Their estimates show that immigration did not have a large impact on native earnings and thus did not cause the increase in Canadian inequality during the early twentieth century. This contrasts with the negative effects of immigration found for the United States (using different approaches). But given the similarity in the immigration inflow between Canada and the United States, it seems likely that the same method applied to US data would yield the same null effect. Unfortunately there are no data on earnings in the US Census until 1940, so any analysis of immigration's effect on wages must use wage data at a level higher than the individual.

Studies that can separate native and immigrant outcomes with microdata in the United States must resort to occupational status, and these tend to find a small negative effect or even a positive effect of immigration on native outcomes. For example, Ferrie (1999) uses 1850 and 1860 US linked data and estimates that immigration caused unskilled natives to move faster up the occupational ladder, while immigration had a negative effect on skilled natives. Ager and Hansen (2017) show that a reduction of inflows due to the 1920s quotas caused natives to enter lower-skilled jobs, suggesting

[7]Ottaviano and Peri's (2012) method is to specify a structural model of the economy where newly arrived immigrants, long-established immigrants, and natives are different types of labor. Further, labor is differentiated by skill in terms of education and labor market experience. The elasticity of substitution between labor types is measured with the data, and then the structural model is estimated. Dustmann et al. (2016) critique this methodology because, due to occupational downgrading at arrival, immigrants' effective skill level is lower than their observed skill level, which implies a misspecification when allocating immigrants to education-experience cells.

that more immigration would have positively affected native jobs (as defined by occupational score), a finding which Tabellini confirms (2017). In addition to the measurement issue of lacking microdata on native wages, these studies all use different empirical methodologies, leaving the effect of immigration on natives unclear, similar to the modern-day literature (Dustmann et al. 2016).

There are several other margins besides wages and occupation on which immigration affects the economy, such as outflows from high-immigration areas, changes in the output mix, and technological adoption or invention. Hatton and Williamson (1998, Chap. 8) estimate that immigration to the northeastern states displaced around 40 natives to other states for every 100 of the migrant inflow. And Collins (1997) demonstrates that the fall of inflows due to World War I and the immigration quotas contributed to the Great Migration of African-Americans from the South to the North. While immigration is often associated with urban areas in the early twentieth century, immigration also had important effects in rural areas. For example, a reduction of inflows to a county between 1910 and 1940 shifted the crop mix toward less labor-intensive and more capital-intensive crops, although it had no effect on the adoption of labor-saving technologies (Lafortune et al. 2015). Besides agriculture, there were important effects on innovation: Moser et al. (2014) show that the inflow of Jewish scientists escaping from Germany to America during the 1930s led to a large increase in chemical patents, with no evidence of negative or positive spillovers on incumbent scientists.

The large volume of inflows into the New World implies that there were also significant impacts on the sending countries. For Ireland, Hatton and Williamson (1998, Chap. 9) find that from the 1860s onward, the population loss induced by emigration led to more rapid wage growth. Rising labor costs also led to a shift within agriculture away from tillage and fostered mechanization (O'Rourke 1991). And in Sweden, higher out-migration rates were associated with more inventions per capita (Andersson et al. 2016). The effect of outflows from the source country may not simply be due to a loss of labor; those who left could still have important effects in the source due to remittances or eventually returning home with savings acquired abroad. As noted earlier, remittances were often used to finance further emigration, but the flow of remittances also enhanced financial development in countries of the European periphery (Esteves and Khoudour-Castéras 2011). Qualitative evidence also suggests that return migration to Italy increased farm prices due to savings brought back home (Wyman 1993; Cinel 1991), and linked census data from Norway shows that return migrants improved their occupational standing due to their temporary trip abroad, likely because of bringing back savings (Abramitzky et al. 2016b). Thus while the effects of immigration on the working population in the New World seem to have been rather mixed, for those that remained in the Old World, the economic effects were largely positive.[8]

[8]On the other hand the contemporary social effects are less clear. For the extreme case of Sweden, which lost nearly a quarter of its population to the New World, Karadja and Prawitz (2016) show that areas with higher outflows had higher strike participation, welfare expenditures, and support for left-wing parties, which they argue is due to greater political power for citizens.

The Legacy of Historical Immigration

The immigration literature is primarily concerned with estimating the immediate consequences of inflows, but it is possible that historical migrations have important long-run effects, over 100 years later. It is often asserted that immigrants leave a distinctly positive legacy to subsequent generations. One recent book carried the title: "Exceptional people. How migration shaped our world and will define our future" (Goldin 2011). While such descriptive accounts tend to accentuate the positive, the economic literature has offered a more nuanced interpretation of the long-term effects of immigration on productivity and economic growth, as well as on inequality and on social cohesiveness.

The literature stresses two enduring effects. On one hand ethnic and birthplace diversity brings skills and knowledge that are complementary in production and may lead to higher incomes. For instance, in cross-country analysis, Alesina et al. (2016) find that, in the presence of a range of controls, greater birthplace diversity, particularly within the immigrant population, leads to higher income per capita and more patenting activity. On the other hand, diversity could reduce intergroup communication and overall levels of trust, and it may also imply divergence in attitudes toward inequality and in preferences over welfare policy and public goods. Alesina et al. (1999) find that, across US cities and counties in 1990, ethnic fractionalization is associated with higher public expenditure but a lower share of expenditure on public goods. Similarly, for the age of mass migration, Ager and Brückner (2013) estimate a positive effect of fractionalization on income at the county level, consistent with gains from specialization, but on the other hand polarization tended to reduce income. And Tabellini (2017) finds that cities in 1910 to 1930 with more immigration had higher employment and manufacturing output but lower tax rates and public spending; the more so the greater the cultural distance between immigrants and natives.[9]

To the extent that culture, human capital, or other traits are correlated across generations, there may be long-run effects stemming from the ancestries of today's population and ultimately from the countries that their immigrant forebears came from. Although disentangling the effects of past immigration from other mechanisms is difficult, a recent literature investigates the effects inherited from immigration over the very long run. This reflects the wider trend in economics toward estimating the causal effects of events and characteristics from deep in the past that shape culture and institutions and influence economic outcomes up to the present. As Spolaore and Wacziarg (2013) note, this literature has increasingly focused on the qualities that

[9]Output is measured as the sum of manufacturing value added plus agricultural value added. Fractionalization measures the likelihood that two randomly selected individuals are from different backgrounds, while polarization measures the difference between the distributions of ethnicity from a bimodal distribution. See Ager and Brückner (2013) for a further discussion. Rodríguez-Pose and von Berlepsch (2017) extend the analysis to 2000 and argue that fractionalization in the late nineteenth century is positively correlated with county income in 2000, while polarization is negatively correlated.

migrants brought with them; thus, what matters for economic development is not a country's geographic location or resources but the characteristics of its population.

For a cross section of countries, Putterman and Weil (2010) examine the effect of origin-country political organization and economic development since 1500 using as weights the ancestral origins of the current population. They find that GDP per capita is higher in countries where the ancestral composition is associated with more advanced state organization but that the greater the degree of fractionalization in ancestry the higher the inequality. Interestingly, they find that the effects of ancestry are stronger than those stemming from the current foreign-born mix. This is consistent with a range of studies finding that past levels of human capital positively influence institutional quality and predict living standards in the present (Glaeser et al. 2004; Easterly and Levine 2016; Chanda et al. 2014; Spolaore and Wacziarg 2013). Human capital persistence is one mechanism through which immigration shocks have long-lasting effects but cultural traits also matter. Algan and Cahuc (2010) show that trust levels persist across generations; for example, Swedish immigrants had high levels of trust in the past, which they transmitted to their descendants today.

Much of the mixing of ancestries occurred during the age of mass migration and, even within countries, this could have differential effects in the present that reflect the settlement patterns of the past. Fulford et al. (2015) explore these effects by estimating the ancestries of the population of the United States for each county in each census between 1850 and 2010. Analyzing the data as a panel, they find that incomes are higher in counties which received immigrants from origins with high levels of income, education, and trust. And the effects of ancestry persist beyond the age of mass migration up to the present, with little change in intensity. Consistent with other studies, they find that while the degree of ancestral fractionalization alone has a positive effect on GDP per capita, ancestry, when weighted by attitudes toward trust, thrift, and cooperation, has a negative effect. This is consistent with the notion that while diversity itself has a positive effect, diversity in attitudes does not.

One channel for a long-run impact of immigration is stronger economic connections abroad. Burchardi et al. (2017) explore the correlation between a county's population ancestry and foreign direct investment to or from the ancestor's origin country. Using an instrumental variables strategy to predict ancestries across counties, they show that doubling the number of descendants from a foreign country increases the likelihood of present-day foreign direct investment with that country by 4%.[10] Provocatively, they argue that FDI with China would be much larger if not for the Chinese Exclusion Act of 1882 and the Asiatic Barred Zone in 1917. This finding parallels the shorter-run work on international trade of Dunlevy and Hutchinson (1999), who show that between 1870 and 1910 more immigration from a country

[10]Burchardi et al. (2017) instrument for the number of foreign-born in a county using a mix of push factors, as proxied by the total outflow from the country to the United States, and pull factors to the specific county, which is proxied by the number of foreign-born entering the county from other continents.

was associated with more international trade. One reason why FDI is correlated with the number of descendants from a country is the reduction of information frictions between countries (Burchardi et al. 2017).

Rather than connecting historical migrations with modern outcomes through the ancestral channel, Sequeira et al. (2017) gauge the relationship between a county's foreign-born share (no matter the source) between 1860 and 1920 with county-level outcomes in 2010.[11] Using an instrumental variables strategy to predict the foreign-born share across counties, the authors find that counties with a higher foreign-born share between 1860 and 1920 had higher levels of income and education in 2010, an outcome that is not simply due to reallocation across counties.[12] This effect did not appear without precedent in 2010 but can be traced through intervening decades. Counties with more immigrants had higher manufacturing output per capita, greater farm value per acre, and more patents also in 1930. Akcigit et al. (2017) confirm this positive effect of immigration on inventions by showing that areas with more foreign-born expertise between 1880 and 1940 had more patenting and citations between 1940 and 2000.

Some of the persistence in the effects of immigration during the age of mass migration may be due to the persistence of immigration itself. Long-established traditions of migration from a particular origin, by eroding the communication gap with the host population, allow subsequent immigrants from the same source to assimilate more easily (Hatton and Leigh 2011). But perhaps more significant is the effect of immigration on education and human capital. As noted earlier, Bandiera et al. (2018) find that the introduction of compulsory state schooling took place earlier in states with higher proportions of immigrants and especially those from origins with no experience of compulsory schooling. To the extent that immigration stimulated such nation building projects, these could have persistent effects.

It is possible that long-term effects might be different for countries where immigrants met rather different host-country conditions than in the United States. However, there is evidence that immigration from Europe to Brazil and Argentina led to higher incomes in the long run. Rocha et al. (2017) examine long-term outcomes of European immigration across municipalities in Brazil, focusing on the establishment from the 1870s of state sponsored settlements, designed to attract immigrants. These municipalities experienced faster transitions to manufacturing

[11]See Rodríguez-Pose and von Berlepsch (2014) on a similar long-run relationship, where instead of instrumenting immigrants with an interaction with the railroad and decade of arrival, they test the relationship with a variety of different instruments, such as distance from New York and the standard shift-share instrument. They also find that immigration had a long-run positive effect on modern-day outcomes, which they explore further by examining the effect by immigrant country of origin in a separate paper (Rodríguez-Pose and von Berlepsch 2015).

[12]To separate the effect of immigration from other factors on long-run growth, they interact the volatile time series of immigration with variation in when a county was connected to the railroad network; they argue that some counties received more or less immigrants not due to county-specific factors but because counties happened to be connected to the railroad during a boom decade of immigration.

and subsequently to services due to human capital accumulation. By 2000, they had income per capita 15% higher than municipalities without settlements. For Argentina, the economic effects seem to have been even larger. Looking across counties, and using frontier military campaigns as an instrument for immigration to localities, Droller (2017) finds that immigration from 1895 to 1914 was associated with more than double the level of income per capita in the present. Again, persistence in human capital was the mechanism, as reflected in the present by more years of education and a higher proportion of skilled workers.

Immigration may have effects on a range of social outcomes, transmitted down the generations by the persistence of cultural norms (Nunn 2012). One example is the effects of immigration on sex composition. For the United States from 1920 to 1940, Angrist (2002) finds that where inflows were heavily skewed toward men, second-generation females had lower labor participation rates and were more likely to be married and to have a more-skilled spouse (Angrist 2002). And Lafortune (2013) shows that males responded to a scarcity of potential marriage partners with more investment in education in order to improve their attractiveness in the marriage market. These imbalances tended to dissipate over time but their effects lingered. Grosjean and Khattar (2015) study the legacy of the high male to female ratio in mid-nineteenth-century Australia, owing originally to convict transportation. They find that across counties, high historic sex ratios are associated in the present with lower female participation rates, fewer women in professional occupations, and more conservative attitudes toward women.

Overall the literature is almost unanimous in finding that immigration during the age of mass migration generated positive economic outcomes both within and between countries. This is all the more remarkable as immigrant arrivals during the age of mass migration do not appear to be exceptional in terms of their selection and assimilation. The main transmission mechanism was by stimulating human capital formation not only among immigrants and their descendants but also as an externality to the wider population. We may speculate that human capital diffusion also helps to explain why the negative effects of fractionalization, which are sometimes observed as contemporary effects, become muted in the longer term.

Conclusion

In this chapter, we have examined a wide range of studies focusing on the age of mass migration and its later demise. The literature on the time series determinants of migrant streams is relatively settled and has not been the focus of the bulk of recent research. On the other hand, the cliometric literature on the political economy of restriction is still comparatively small and deserves further attention. The most recent wave of research on immigrant selection and assimilation has revised pre-existing views, downgrading the view of immigrants as somehow exceptional. And while their effects on host-country labor markets are still contested, there is at least some evidence that they had characteristics broadly comparable with the native-born and competed on relatively equal terms. In that light, it is all the more remarkable that the

long-term legacy of the age of mass migration should have been so positive, especially for the United States.

The literature has produced many advances, not only in methodology and data but also in the questions asked and the subtlety of the answers given. Clearly, there is room for further research to deepen these insights. But the focus has been heavily skewed toward migrants to, their experience in, and their effects on, the United States, something that is strongly reflected in the studies cited here. The largest dividends for future research probably lie in other New World countries on which less attention has been lavished. And our account completely omits migrations in Asia, which were quantitatively significant and have left enduring legacies but have not received comparable attention. Perhaps a future Cliometric Handbook chapter could be devoted to that, but if it were written now, it would be very short.

References

Abramitzky R, Platt Boustan L, Eriksson K (2012) Europe's tired, poor, huddled masses: self-selection and economic outcomes in the age of mass migration. Am Econ Rev 102(5): 1832–1856

Abramitzky R, Platt Boustan L, Eriksson K (2013) Have the poor always been less likely to migrate? Evidence from inheritance practices during the age of mass migration. J Dev Econ 102:2–14

Abramitzky R, Platt Boustan L, Eriksson K (2014) A nation of immigrants: assimilation and economic outcomes in the age of mass migration. J Polit Econ 122(3):467–506

Abramitzky R, Platt Boustan L, Eriksson K (2016a) Cultural assimilation during the age of mass migration. NBER working paper 22381

Abramitzky R, Platt Boustan L, Eriksson K (2016b) To the New World and back again: return migrants in the age of mass migration. NBER working paper 22659

Ager P, Brückner M (2013) Cultural diversity and economic growth: evidence from the US during the age of mass migration. Eur Econ Rev 64(C):76–97

Ager P, Hansen CW (2017) Closing heaven's door: evidence from the 1920s U.S. immigration quota acts. Unpublished paper: University of Copenhagen

Akcigit U, Grigsby J, Nicholas T (2017) Immigration and the rise of American ingenuity. Am Econ Rev 107(5):327–331

Akerman S (1976) Theories and methods of migration research. In: Rundblom H, Norman H (eds) From Sweden to America: a history of the migration. University of Minnesota Press, Minneapolis

Alesina A, Baqir R, Easterly W (1999) Public goods and ethnic divisions. Q J Econ 114(4): 1243–1284

Alesina A, Harnoss J, Rapoport H (2016) Birthplace diversity and economic prosperity. J Econ Growth 21(2):101–138

Alexander R, Ward Z (2018) Age at arrival and assimilation during the age of mass migration. J Econ Hist. (forthcoming)

Algan Y, Cahuc P (2010) Inherited trust and growth. Am Econ Rev 100(5):2060–2092

Altonji JG, Card D (1991) The effects of immigration on the labor market outcomes of less-skilled natives. In: Abowd JM, Freeman RB (eds) Immigration, trade, and the labor market. University of Chicago Press, Chicago, pp 201–234

Anderson E (2001) Globalisation and wage inequalities, 1870–1970. Eur Rev Econ Hist 5(1): 91–118
Andersson D, Mounir K, Prawitz E (2016) Mass migration, cheap labor, and innovation. Unpublished paper, Uppsala University
Angrist J (2002) How do sex ratios affect marriage and labor markets? Evidence from America's second generation. Q J Econ 117(3):997–1038
Armstrong A, Lewis FD (2017) Transatlantic wage gaps and the migration decision: Europe–Canada in the 1920s. Cliometrica 11(2):153–182
Balderas JU, Greenwood MJ (2010) From Europe to the Americas: a comparative panel-data analysis of migration to Argentina, Brazil and the United States, 1870–1910. J Popul Econ 23 (4):1301–1318
Bandiera O, Rasul I, Viarengo M (2013) The making of modern America: migration flows in the age of mass migration. J Dev Econ 102(C):23–47
Bandiera O, Mohnen M, Rasul I, Viarengo M (2018) Nation-building through compulsory schooling during the age of mass migration. Econ J. (forthcoming)
Bertocchi G, Strozzi C (2008) International migration and the role of institutions. Public Choice 137(1–2):81–102
Betrán C, Pons MA (2004) Skilled and unskilled wage differentials and economic integration, 1870–1930. Eur Rev Econ Hist 8(1):29–60
Betrán C, Ferri J, Pons MA (2010) Explaining UK wage inequality in the past globalisation period, 1880–1913. Cliometrica 4(1):19–50
Biavaschi C (2013) The labor demand was downward sloping: disentangling migrants' inflows and outflows, 1929–1957. Econ Lett 118(3):531–534
Biavaschi C, Facchini G (2017) Immigrant franchise and immigration policy: evidence from the progressive era. Unpublished paper: University of Nottingham
Biavaschi C, Giulietti C, Siddique Z (2017) The economic payoff of name Americanization. J Labor Econ 35(4):1089–1116
Bohlin J, Eurenius A-M (2010) Why they moved – emigration from the Swedish countryside to the United States, 1881–1910. Explor Econ Hist 47(4):533–551
Borjas GJ (1985) Assimilation, changes in cohort quality, and the earnings of immigrants. J Labor Econ 3(4):463–489
Borjas GJ (1987) Self-selection and the earnings of immigrants. Am Econ Rev 77(4):531–553
Borjas GJ (2003) The labor demand curve is downward sloping: reexamining the impact of immigration on the labor market. Q J Econ 118(4):1335–1374
Borjas GJ, Bratsberg B (1996) Who leaves? The outmigration of the foreign-born. Rev Econ Stat 78(1):165–176
Boustan LP (2007) Were Jews political refugees or economic migrants? Assessing the persecution theory of Jewish emigration, 1881–1914. In: Hatton TJ, O'Rourke KH, Taylor AM (eds) The new comparative economic history: essays in honor of Jeffrey G. Williamson. MIT Press, Cambridge, MA, pp 267–290
Burchardi KB, Chaney T, Hassan TA (2017) Migrants, ancestors, and investments. NBER working paper 21847
Card D (2001) Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. J Labor Econ 19(1):22–64
Carneiro PM, Lee S, Reis H (2015) Please call me John: name choice and the assimilation of immigrants in the United States, 1900–1930. London: Cemmap working paper 28/15
Catron P (2017) The citizenship advantage: immigrant socioeconomic attainment across generations in the age of mass migration. Unpublished paper, University of Pennsylvania
Chanda A, Cook CJ, Putterman L (2014) Persistence of fortune: accounting for population movements, there was no post-Columbian reversal. Am Econ J Macroecon 6(3):1–28
Chiswick BR (1978) The effect of Americanization on the earnings of foreign-born men. J Polit Econ 86(5):897–921

Cinel D (1991) The national integration of Italian return migration, 1870–1929. Cambridge University Press, Cambridge
Cohn RL (1995) Occupational evidence on the causes of immigration to the United States, 1836–1853. Explor Econ Hist 32(3):383–408
Cohn RL (2005) The transition from sail to steam in immigration to the United States. J Econ Hist 65(2):469–495
Cohn RL (2009) Mass migration under Sail: European immigration to the antebellum United States. Cambridge University Press, Cambridge
Collins WJ (1997) When the tide turned: immigration and the delay of the great black migration. J Econ Hist 57(3):607–632
Connor D (2016) The cream of the crop? Inequality and migrant selectivity in Ireland during the age of mass migration. University of California Los Angeles: Unpublished paper
Covarrubias M, Lafortune J, Tessada J (2015) Who comes and why? Determinants of immigrants' skill level in the early 20th century U.S. J Demogr Econ 81:115–155
Deltas G, Sicotte R, Tomczak P (2008) Passenger shipping cartels and their effect on transatlantic migration. Rev Econ Stat 90(1):119–133
Droller F (2017) Migration, population composition, and long run economic development: evidence from settlements in the pampas. Econ J. (forthcoming)
Dunlevy JA, Hutchinson WK (1999) The impact of immigration on American import trade in the late nineteenth and early twentieth centuries. J Econ Hist 59(4):1043–1062
Dustmann C, Schönberg U, Stuhler J (2016) The impact of immigration: why do studies reach such different results? J Econ Perspect 30(4):31–56
Easterly W, Levine R (2016) The European origins of economic development. J Econ Growth 21(3):225–257
Engerman SL, Sokoloff KL (2005) The evolution of suffrage institutions in the New World. J Econ Hist 65(4):891–921
Esteves R, Khoudour-Castéras D (2011) Remittances, capital flows and financial development during the mass migration period, 1870–1913. Eur Rev Econ Hist 15(3):443–474
Faini R, Venturini A (1994) Italian emigration in the pre-war period. In: Hatton TJ, Williamson JG (eds) Migration and the international labor market, 1850–1939. Routledge, London, pp 72–90
Ferrie JP (1999) Yankeys now: immigrants in the Antebellum US 1840–1860. Oxford University Press, New York
Foreman-Peck J (1992) A political economy of international migration, 1815–1914. Manch Sch 60(4):359–376
Fouka V (2015) Backlash: the unintended effects of language prohibition in US schools after World War I. Stanford University: Unpublished paper
Fulford SL, Petkov I, Schiantarelli F (2015) Does it matter where you came from? Ancestry composition and economic performance of US counties, 1850–2010. IZA, Bonn. Discussion Paper 9060
Glaeser EL, La Porta R, Lopez-De-Silanes F, Shleifer A (2004) Do institutions cause growth? J Econ Growth 9(3):271–303
Goldin CD (1994) The political economy of immigration restriction in the United States. In: Goldin C, Libecap G (eds) The regulated economy: a historical approach to political economy. University of Chicago Press, Chicago, pp 223–258
Goldin I (2011) Exceptional people: how migration shaped our world and will define our future. Princeton University Press, Princeton
Goldstein JR, Stecklov G (2016) From Patrick to John F.: ethnic names and occupational success in the last era of mass migration. Am Sociol Rev 81(1):85–106
Gould JD (1979) European inter-continental emigration: patterns and causes. J Eur Econ Hist 8(3):593–679
Gould JD (1980) European inter-continental emigration. The road home: return migration from the USA. J Eur Econ Hist 9(1):41–112
Green AG, Green DA (2016) Immigration and the Canadian earnings distribution in the first half of the twentieth century. J Econ Hist 76(2):387–426

Green AG, MacKinnon M, Minns C (2002) Dominion or republic? Migrants to North America from the United Kingdom, 1870–1910. Econ Hist Rev 55(1):666–696
Greenwood MJ (2007) Modeling the age and age composition of late 19th century U.S. immigrants from Europe. Explor Econ Hist 44(2):255–269
Greenwood MJ, Ward Z (2015) Immigration quotas, World War I, and emigrant flows from the United States in the early 20th century. Explor Econ Hist 55(1):76–96
Grosjean P, Khattar R (2015) It's raining Men! Hallelujah? University of New South Wales: Unpublished paper
Hanes C (1996) Immigrants' relative rate of wage growth in the late 19th century. Explor Econ Hist 33(1):35–64
Hatton TJ (1995) A model of U.K. emigration, 1870–1913. Rev Econ Stat 77(3):407–415
Hatton TJ (1997) The immigrant assimilation puzzle in late nineteenth-century America. J Econ Hist 57(1):34–62
Hatton TJ (2000) How much did immigrant 'quality' decline in late nineteenth century America? J Popul Econ 13(3):509–535
Hatton TJ (2011) The cliometrics of international migration: a survey. J Econ Surv 24(5):941–969
Hatton TJ, Leigh A (2011) Immigrants assimilate as communities, not just as individuals. J Popul Econ 24(2):389–419
Hatton TJ, Williamson JG (1998) The age of mass migration: causes and economic impact. Oxford University Press, New York
Hatton TJ, Williamson JG (2005) Global migration and the World economy: two centuries of policy and performance. MIT Press, Cambridge, MA
Hatton TJ, Williamson JG (2007) A dual policy paradox: why have trade and immigration Policies always differed in labor scarce economies? In: Hatton TJ, O'Rourke KH, Taylor AM (eds) The new comparative economic history: essays in honor of Jeffrey G. Williamson. MIT Press, Cambridge, MA, pp 217–240
Hilger N (2016) Upward mobility and discrimination: the case of Asian Americans. NBER working paper 22748
Inwood K, Minns C, Summerfield F (2016) Reverse assimilation? Immigrants in the Canadian labour market during the Great Depression. Eur Rev Econ Hist 20(3):299–321
Jenks JW, Lauck WJ (1926) The immigration problem: a study of American immigration conditions and needs, 6th edn. Funk and Wagnalls, New York
Jerome H (1926) Migration and business cycles. National Bureau of Economic Research, New York
Karadja M, Prawitz E (2016) Exit, voice, and political change: evidence from Swedish mass migration to the United States. Uppsala University: Unpublished paper
Keeling D (1999) The transport revolution and trans-Atlantic migration, 1850–1914. Res Econ Hist 19:39–74
Kirk D (1946) Europe's population in the interwar years. League of Nations, Geneva
Kosack E, Ward Z (2018) El Sueño Americano? The generational progress of Mexican Americans in US history. Australian National University: Unpublished paper
Kuznets S, Rubin E (1954) Immigration and the Foreign born. National Bureau of Economic Research, Cambridge, MA
Lafortune J (2013) Making yourself attractive: pre-marital investments and the returns to education in the marriage market. Am Econ J 5(2):151–178
Lafortune J, Tessada J, González-Velosa C (2015) More hands, more power? Estimating the impact of immigration on output and technology choices using early 20th century US agriculture. J Int Econ 97(2):339–358
Larsen UM (1982) A quantitative study of emigration from Denmark to the United States, 1870–1913. Scand Econ Hist Rev 30(2):101–128
Lleras-Muney A, Shertzer A (2015) Did the Americanization movement succeed? An evaluation of the effect of English-only and compulsory schooling laws on immigrants. Am Econ J Econ Pol 7(3):258–290
Lubotsky D (2007) Chutes or ladders? A longitudinal analysis of immigrant earnings. J Polit Econ 115(5):820–867

MacKinnon M, Parent D (2012) Resisting the melting pot: the long term impact of maintaining identity for Franco-Americans in New England. Explor Econ Hist 49(1):30–59
Magee GB, Thompson AS (2006) The global and local: explaining migrant remittance flows in the English-speaking world, 1880–1914. J Econ Hist 66(1):177–202
Massey CG (2016) Immigration quotas and immigrant selection. Explor Econ Hist 60(1):21–40
Minns C (2000) Income, cohort effects, and occupational mobility: a new look at immigration to the United States at the turn of the 20th century. Explor Econ Hist 37(4):326–350
Mokyr J, Ó Gráda C (1982) Emigration and poverty in pre-famine Ireland. Explor Econ Hist 19(4):360–384
Moser P, Voena A, Waldinger F (2014) German Jewish émigrés and US invention. Am Econ Rev 104(10):3222–3255
Nunn N (2012) Culture and the historical process. Econ Hist Dev Reg 27(S1):S108–S126
O'Rourke KH (1991) Rural depopulation in a small open economy: Ireland 1856–1876. Explor Econ Hist 28(4):409–432
O'Rourke KH, Williamson JG, Hatton TJ (1994) Mass migration, commodity market Integration and real wage convergence. In: Hatton TJ, Williamson JG (eds) Migration and the international labor market, 1850–1939. Routledge, London, pp 203–220
Ottaviano GIP, Peri G (2012) Rethinking the effect of immigration on wages. J Eur Econ Assoc 10(1):152–197
Pérez S (2017) The (South) American dream: mobility and economic outcomes of first- and second-generation immigrants in 19th-century Argentina. J Econ Hist 77(4):971–1006
Pope DH (1981) Modelling the peopling of Australia, 1900–1930. Aust Econ Pap 20:258–282
Pope D, Withers G (1994) Immigration and wages in late nineteenth century Australia. In: Hatton TJ, Williamson JG (eds) Migration and the international labor market, 1850–1939. Routledge, London, pp 240–262
Putterman L, Weil DN (2010) Post-1500 population flows and the long-run determinants of economic growth and inequality. Q J Econ 125(4):1627–1682
Quigley JM (1972) An economic model of Swedish emigration. Q J Econ 86(1):111–126
Richardson G (2005) The origins of anti-immigrant sentiments: evidence from the heartland in the age of mass migration. B. E. Press, Topics in Economic Analysis & Policy, 5, Art 11
Rocha R, Ferraz C, Soares RR (2017) Human capital persistence and development. Am Econ J 9(4):105–136
Rodríguez-Pose A, von Berlepsch V (2014) When migrants rule: the legacy of mass migration on economic development in the United States. Ann Assoc Am Geogr 104(3):628–651
Rodríguez-Pose A, von Berlepsch V (2015) European migration, national origin and long-term economic development in the United States. Econ Geogr 91(4):393–424
Rodríguez-Pose A, von Berlepsch V (2017) Does population diversity matter for economic development in the very long-term? Historic migration, diversity and county wealth in the US. CEPR, London. Discussion Paper 12347
Sánchez-Alonso B (2000a) European emigration in the late nineteenth century: the paradoxical case of Spain. Econ Hist Rev 53(2):309–330
Sánchez-Alonso B (2000b) Those who left and those who stayed behind: explaining Emigration from the regions of Spain, 1880–1914. J Econ Hist 60(3):730–755
Sánchez-Alonso B (2007) The other Europeans: immigration into Latin America (1870–1914). Revista de Historia Economica. J Iberian Latin Am Econ Hist 25(3):395–426
Sánchez-Alonso B (2013) Making sense of immigration policy: Argentina, 1870–1930. Econ Hist Rev 66(2):601–627
Sequeira S, Nunn N, Qian N (2017) Migrants and the making of America: the short- and long-run effects of immigration during the age of mass migration. NBER working paper 23289
Shertzer A (2016) Immigrant group size and political mobilization: evidence from European migration to the United States. J Public Econ 139:1–12

Shughart W, Tollinson R, Kimenyi M (1986) The political economy of immigration restrictions. Yale J Regul 51(4):79–97
Spitzer Y (2014) Pogroms, networks, and migration: the Jewish migration from the Russian Empire to the United States 1881–1914. Brown University: Unpublished paper
Spitzer Y, Zimran A (2017) Migrant self-selection: anthropometric evidence from the mass migration of Italians to the United States, 1907–1925. Unpublished paper, Northwestern University
Spolaore E, Wacziarg R (2013) How deep are the roots of economic development? J Econ Lit 51(2):325–369
Steckel RH (1995) Stature and the standard of living. J Econ Lit 33(4):1903–1940
Steckel RH (2008) Biological measures of the standard of living. J Econ Perspect 22(1):129–152
Stolz Y, Baten J (2012) Brain drain in the age of mass migration: does relative inequality explain migrant selectivity? Explor Econ Hist 49(2):205–220
Tabellini M (2017) Gifts of the immigrants, woes of the natives: lessons from the age of mass migration. MIT: Unpublished paper
Taylor AM (1994) Mass migration to distant southern shores: Argentina and Australia. In: Hatton TJ, Williamson JG (eds) Migration and the international labor market, 1850–1939. Routledge, London
Taylor AM, Williamson JG (1997) Convergence in the age of mass migration. Eur Rev Econ Hist 1(1):27–63
Thomas DS (1941) Social and economic consequences of Swedish population movements. Macmillan, New York
Timmer AS, Williamson JG (1998) Immigration policy prior to the 1930s: Labor markets, policy interactions, and globalization backlash. Popul Dev Rev 24(4):739–771
Todaro MP (1969) A model of labor migration and urban unemployment in less developed countries. Am Econ Rev 59(1):138–148
Tollnek F, Baten J (2016) Age-heaping-based human capital estimates. In: Diebolt C, Haupert M (eds) Handbook of Cliometrics. Springer, pp 131–154
United States Immigration Commission (1911) Reports, 61st congress, 3rd session. Government Printing Office, Washington, DC
Ward Z (2016) There and back (and back) again: repeat migration to the United States, 1897–1936. Unpublished paper, Australian National University
Ward Z (2017a) Birds of passage: return migration, self-selection and immigration quotas. Explor Econ Hist 64(1):37–52
Ward Z (2017b) The not-so-hot melting pot: the persistence of outcomes for descendants of the age of mass migration. Unpublished paper, Australian National University
Ward Z (2018) Have language skills always been so valuable? The low return to English fluency during the age of mass migration. Unpublished paper, Australian National University
Wegge SA (1998) Chain migration and information networks: evidence from nineteenth- century Hesse-Cassel. J Econ Hist 58(4):957–986
Williamson JG (1998) Globalization, labor markets and policy backlash in the past. J Econ Perspect 12(4):51–72
Wyman M (1993) Round-trip to America: the Immigrants Return to Europe, 1880–1930. Cornell University Press, Ithaca
Xie B (2017) The effect of immigration quotas on wages, the great Black migration, and industrial development. Unpublished paper, Rutgers University

Cliometrics and the Concept of Human Capital

Charlotte Le Chapelain

Contents

Abstract

The role played by human capital in the historical process of economic development is an important issue in cliometrics. This chapter traces the history of the human capital concept and underlines that it has undergone criticism since its origins. Among these criticisms, we stress that the human capital research program was plagued with measurement difficulties. This issue has recently fostered a sense of strong reluctance toward the concept. Portraying a growing debate in cliometrics – the role of human capital in the first stage of the industrialization process in the late eighteenth and early nineteenth centuries in Western Europe – we emphasize that recent cliometric contributions on the issue open up stimulating lines of thought about what human capital is and how it can adequately be measured.

C. Le Chapelain (✉)
Centre Lyonnais d'Histoire du Droit et de la Pensée Politique, Université de Lyon, Lyon, France
e-mail: charlotte.le-chapelain@univ-lyon3.fr

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_45

Keywords
History of human capital · Education · Industrialization · Cliometrics

Introduction

Analysis of the economic role of individuals' skills and abilities long preceded that intellectual moment which is today referred to as the "human capital revolution." But the formalization of these ideas and their introduction into the core of economic theory emerged only in the 1960s, under the impulse of Theodore Schultz, Gary Becker, and Jacob Mincer. Prior to the 1960s, human beings or their skills were indeed considered in certain analyses as a particular form of capital, but they appeared therein only in a metaphorical way. The human capital theorists of the 1960s provided solid and precise theoretical foundations for that metaphor by using the analytical framework of standard capital. By considering human skills and abilities in analogy with physical capital, they laid the foundations of human capital theory.

Several moments are habitually recognized as fundamental steps in the building of the human capital research program. Schultz's articles published in 1959, 1960, and 1961a, the issue of *the Journal of Political Economy*, "Investments in Human Beings" in October 1962, Gary Becker's book *Human Capital* (1964), and the work of Mincer (1957, 1958), are identified as crucial contributions (see, for instance, Blaug 1976, Teixeira 2000, Ehrlich and Murphy 2007). Schultz, Becker, and Mincer, through different approaches, are generally seen as the fathers of the "human capital revolution." Since that time, a huge literature has developed.

Although this revolution was a success in the sense that the concept came into wide use both in economics and in various social sciences, human capital theory has come in for a great deal of criticism. These critiques are diverse, but the thrust of a large number of them concerns the major empirical challenges faced by the theory: at root, they address the difficulty of measuring human capital, especially at the macro level.

This chapter traces the history of the human capital concept by providing an overview of the way the idea of human capital was conceptualized, the historical background in the late 1950s against which the human capital revolution took place, and the critiques that have been levelled at it since. In particular, we point out that the concept has recently come under severe pressure through a resurgence of criticisms based on the problem of measuring human capital.

The cliometric approach quickly embraced the theoretical insights provided by human capital theorists. The role of human capital in the process of industrialization is a prominent issue for today's economic historians. Early analysis on the issue had come to the paradoxical conclusion that human capital was not important as regards the first stage of the industrialization process, but this thesis is currently under re-evaluation. By reviewing the cliometric contributions on the role of human capital in the first phase of industrialization, we highlight that this re-evaluation opens new and innovative avenues to tackling the issue of human capital measurement.

The chapter is organized as follows. Section "The Development of the Human Capital Research Program" provides an account of the human capital revolution of the late 1950s. We highlight the context in which the revolution took place, stress the importance of the figure of Schultz in promulgating the research program, and finally describe the theoretical synthesis it achieved. Yet while Schultz, Becker, and Mincer succeeded in imposing human capital theory in the mainstream economics, it nevertheless met criticisms from competing methodological perspectives also to be found within the mainstream. Section "Reservations, Old and New, About the Concept of Human Capital" retraces these criticisms and underscores that the empirical difficulty in measuring human capital is an issue that has been and is still currently a source of misgivings about the concept. In section "Human Capital and the Industrialization Process," we describe a growing debate within the field of cliometrics, namely, the (re-)evaluation of the role of human capital in the first stage of the industrialization process. The view that human capital played only a minor role in the first stage of industrialization is currently under reconsideration. The revisionist literature builds upon the idea that early literature was plagued by shortcomings concerning the way the human capital endowments were evaluated. This view has spawned new and innovative reflections within cliometrics about which type of knowledge was truly productive in eighteenth- and nineteenth-century Western Europe, which types of skills can be assimilated to a form of (human) capital, and thus how to accurately measure human capital.

Section I. The Development of the Human Capital Research Program

Forerunners

Analysis of the economic role of individuals' skills and abilities long preceded the human capital revolution of the 1950s and 1960s. Many authors have considered human beings, and their skills, as capital. Kiker's survey (1966) refers in particular to Petty, Smith, Say, Senior, List, von Thünen, Roscher, Bagehot, Ernst Engel, Sidgwick, Walras, and Fisher.[1] For his part, Sweetland (1996) considers that "the most prominent economists to address issues of human capital were Adam Smith, John Stuart Mill, and Alfred Marshall. Irving Fisher, prominent in his own right, expressed the pivotal arguments connecting early economic thought to contemporary human capital methodologies" (Sweetland 1996, p. 343). As pointed out by Kiker (1966), some authors attempted to provide a monetary evaluation of human beings or their skills via cost-of-production (costs of "producing a human being") or capitalized-earnings procedures, which refer to the estimation of the present value of future income streams. Different motives led them to the following monetary evaluation of what is recognized today as human capital: demonstrating the power

[1]For an exhaustive analysis of the authors who treated human beings as capital, see Kiker (1966).

of a nation, estimating the cost of the war, and assessing the economic effects of education, health, or migration. Other authors never attempted to provide monetary evaluations yet still considered human beings and their skills as a form of capital, whether explicitly or implicitly. This is notably the case for Adam Smith, Jean-Baptiste Say, John Stuart Mill, William Roscher, Walter Bagehot, and Henry Sidgwick.

Smith's thought on the issue deserves special attention. The treatment of human skills within the analytical framework of capital is in fact very explicit in Smith's thought. Recalling that Smith was "concerned with education for the betterment of men, not for the creation of human resources," Bowman (1966, p. 113) rightly notes that Smith's analysis of human skills is by far the closest to the concept developed in the 1960s.

For this reason, Smith is generally recognized as having laid out the premises on which the modern concept of human capital has drawn. Spengler (1977) and Schultz (1992) explicitly attributed it to him, while Rosen (2008) refers to Petty and Marshall but acknowledges Smith as the forerunner of the concept. Becker (1962) himself,[2] in his formulation of the analytical framework explaining investment choices in human capital, invokes the Smithian analysis.

In *the Wealth of Nations* (1776, p. 282), Smith wrote:

> The acquisition of such talents, by the maintenance of the acquirer during his education, study, or apprenticeship, always costs a real expense, which is a capital fixed and realized, as it were, in his person. Those talents, as they make a part of his fortune, so do they likewise that of the society to which he belongs. The improved dexterity of a workman may be considered in the same light as a machine or instrument of trade which facilitates and abridges labor, and which, though it costs a certain expense, repays that expense with a profit.
>
> When any expensive machine is erected, the extraordinary work to be performed by it before it is worn out, it must be expected, will replace the capital laid out upon it, with at least the ordinary profits. A man educated at the expense of much labor and time to any of those employments which require extraordinary dexterity and skill, may be compared to one of those expensive machines. The work which he learns to perform, it must be expected, over and above the usual wages of common labor, will replace to him the whole expense of his education, with at least the ordinary profits of an equally valuable capital. It must do this, too, in a reasonable time, regard being had to the very uncertain duration of human life, in the same manner as to the more certain duration of the machine. (Smith 1776, pp. 118–119)

The idea that acquiring skills is costly but that it contributes to the accumulation of a particular form of capital that yields future returns is clear in Smith's analysis. Yet that "investment" orientation is precisely what, in Blaug's words (1976, p. 829), forms "the 'hard core' of the human-capital research program" developed at the end of the 1950s. More particularly, Smith's analysis comes very close to the analytical framework Becker developed in his landmark 1964 book. Despite these early developments, the acceptation and generalization of this theoretical perspective made no progress until the end of the 1950s.

[2]Becker also mentions Mill and Marshall's legacy.

Schultz's Early Developments and the Opportunity of Solow's Residual

As we have emphasized, the human capital revolution does not mark the emergence of a new analytical perspective but rather the acceptation and the concrete deployment of this perspective at the core of economic analysis. The 1950s saw a renewed interest in the issue of economic growth and growth accounting exercises, which are commonly recognized as the decisive factors in the success of the human capital revolution: "Much of the original work that led to the 'human capital revolution' of the late 1950s and early 1960s followed an earlier revolution in economic thought spawned by the neoclassical growth model (Solow 1956)" (Ehrlich and Murphy 2007, p. 1).

Teixeira (2000) cites some other features of the intellectual context of the 1960s which created favorable conditions for the human capital revolution. Besides the flourishing of analysis concerned with the sources of economic growth – growth accounting research after World War II and the rise of development studies – he points outs the methodological and institutional dimensions that underpinned the human capital research program and ensured its rapid development. He underlines the role of the Chicago Department of Economics and the rise of Milton Friedman's positive economics, which promoted an empirical trend in economics more widely. The political context was also favorable: the "spread of the Keynesian gospel" had shaped the idea that public education expenditures are economically meaningful and, according to him, created an environment conducive to the emergence of human capital theory. Finally, he underlines the influence of some international institutions, which rapidly endorsed human capital ideas (especially the World Bank and the OECD).

Growth theories, in fact, played a critical role in the development of human capital theory, notably through the puzzle of the residual which was brought to light by the growth accounting exercises of the 1950s and 1960s. Showing that much of the US economic growth was not explained by the rate of growth of inputs but by the residual contribution of total factor productivity (TFP), Tinbergen (1942), Fabricant (1954), Abramovitz (1956), and in particular Solow's contribution of 1957[3] sets the stage for new challenging research opportunities. Considerations about the quality of the factors of production, especially the labor input, were raised in order to resolve the puzzle, and these made a crucial contribution to the rise of human capital thinking. The problem of the residual highlighted the inconsistency between the theoretical insights and the empirical evidence in growth approaches, which in fact provided Theodore Schultz, who is recognized as "the father of the human investment revolution in economic thought" (Bowman 1980, see also Blaug 1966), a favorable context in which to promote the reflections he had developed since the 1940s in his works on agriculture. While Schultz's articles published in 1958, 1959, 1960, and 1961a are often regarded as a starting point for the human

[3]Solow's contribution of 1957 was a landmark in the development of growth accounting.

capital research program (see Bowman 1966, 1980, Sobel 1978), his thought on human capital traces its origin back to his analysis of agricultural poverty. The great influence these works on agricultural economics exerted on the later development of human capital theory has been outlined by Nerlove (1999) and has been examined in further detail by Le Chapelain and Matéos (2018).

As early as 1943, in his book *Redirecting Farm Policy*, Schultz identifies differences in educational expenditures as lying at the origin of labor productivity inequalities in agriculture and, consequently, of income inequalities. From this analysis, he sketched out the idea that educational expenditures can be regarded as a form of investment. These first forays into the idea of human capital, developed in a rather piecemeal fashion in the 1940s and 1950s, evolve into a full-blown research program from the end of the 1950s onward.

In his signal contributions of 1959, 1960, 1961a, and 1962 to the "human capital revolution," Schultz in fact explicitly introduced the idea of considering education as a form of capital in the light of the puzzle of the residual[4]:

> The principal hypothesis underlying this treatment of education is that some important increases in national income are a consequence of additions to the stock of this form of capital. Although it will be far from easy to put this hypothesis to the test, there are many indications that some, and perhaps a substantial part, of the unexplained increases in national income in the United States are attributable to the formation of this kind of capital. (Schultz 1960, p. 571)

Denison's work of 1962 provided empirical support for the hypothesis formulated by Schultz (Schultz never properly proceeded to growth accounting research). His approach relies on an estimation of the quality of labor measured through education, and it shows that taking this "educational capital" into account sharply reduced the value of the residual. But Schultz is recognized as one of the founding fathers of human capital theory because he explicitly linked the empirical analysis of aggregate input-output series with the theme of investment in human beings.

The Synthesis Provided by Schultz, Becker, and Mincer

> I propose to treat education as an investment in man and to treat its consequences as a form of capital. Since education becomes a part of the person receiving it, I shall refer to it as human capital. (Schultz 1960, p. 571)

The human capital revolution marks a shift in the way economic analysis examines individual skills and capacities (accumulated through education, experience, health, etc.). These skills would henceforth be considered a form of capital and therefore be seen as arising as the result of an investment, whereas education-related spending

[4]The residual was not the only puzzle that led Schultz to build the foundations of human capital theory. We must also cite the Leontief paradox (see Schultz 1972a). But the residual appears to be by far the most decisive element in the development of human capital theory.

had hitherto only been addressed through the prism of consumption. For Bowman (1966), this investment orientation is the prime characteristic of the human capital revolution ("investment in human beings").

Besides Schultz's prominent role in the early conceptualization of the human capital idea, Becker (1964) and Mincer (1958) are recognized as two other great leaders of the human capital revolution. Their respective works exerted great influence on the diffusion of the concept of human capital within economic analysis at the end of the 1950s.

Becker's 1964 monograph *Human Capital* is well known for providing a framework for analysis of individual investment decisions in human capital, based on a cost-benefit assessment. Relying on rational choice theory as a framework of analysis (Teixeira 2014), Becker gave solid micro foundations to Schultz's proposals,[5] and his text is now recognized as one of the most decisive contributions to the emerging research program: indeed, his 1964 monograph is considered the *locus classicus* of human capital theory (Blaug 1976). In Becker's canonical model, education yields future income flows under the form of wage increases coming from the enhanced productivity due to higher individual human capital. Education also involves direct and indirect costs, which correspond, respectively, to expenditures directly related to education – such as education fees, transportation costs, or books – and to earnings forgone due to continuing schooling. Through this cost-benefit approach, and by an analogy with physical capital, the rate of return to human capital investment is defined as the rate of discount that equalizes the stream of discounted benefits with the stream of costs. This analytical framework explaining the individual's choices of investment in education has progressively become the most representative piece of a theory of human capital grounded on methodological individualism,[6] that is on "the view that all social phenomena should be traced back to their foundation in individual behavior" (Blaug 1976, p. 830).

Mincer's works were also motivated by an initial interest in an issue that was distinct from Schultz's macroeconomic concerns with economic development and growth.[7] His 1957 and 1958 contributions to human capital theory analysis dealt with the question of wage inequalities and are, as with Becker's approach, grounded on the application of rational choice theory to provide a theory of personal income distribution:

> The starting point of an economic analysis of personal income distribution must be an exploration of the implication of the theory of rational choice. (Mincer 1958, p. 283)

[5]Schultz pays explicit tribute to Becker's theoretical model of investment in human capital, developed in the 1960s. See, for instance, Schultz, 1961b, p. 1037; at about the same time he also wrote: "I have placed the paper by Gary S. Becker first because it gives the reader an overview of the pervasiveness of human capital and because it reveals many vistas awaiting to be explored" (Schultz, 1962, p. 2).

[6]For a critical review of the meanings associated with methodological individualism, as well as its contradictions, see Hodgson (2007).

[7]"Becker and Mincer were not engaged in making aggregative estimates of the contribution of education to income growth" (Bowman, 1964, 453).

Mincer's theoretical proposal treated education as a determinant for wages. In this way it contributes – albeit unintentionally, since Mincer's first developments were isolated[8] – to the development of human capital theory. His work was deployed mainly in the field of labor economics (Teixeira 2007, 2011).

While Schultz, Becker, and Mincer developed a body of work research arising from distinct questions and motives and took different routes of analysis, their respective works constituted the human capital research program, which rapidly spread to the various fields of economic analysis.

Section II. Reservations, Old and New, About the Concept of Human Capital

Although the success of human capital theory is undeniable – as attested by the broad usage of the concept in economics and beyond – the human capital program has faced criticism ever since its inception. This section reviews the history of these concerns.

Section "Theoretical and Methodological Objections" offers an overview of major criticisms stemming from a reluctance to apply standard capital theory to the question of individual skills and capabilities or based on a rejection of the methodological foundation of human capital theory, i.e., the rejection of methodological individualism.

We then turn to questions that arise from the mixed results of the empirical literature on economic growth. We show in the section "Skepticism About the Measurement of Human Capital" that the ambiguous findings of analyses exploring the contribution of education to economic growth have led to the resurgence of doubts about the concept of human capital. The criticism here is distinct from those stemming from theoretical objections or methodological disputes. It arises out of the issue of the measurement of human capital that has bedeviled the human capital research program since its origins. This issue has more recently led to a radical revision of the concept, prompted by the so-called "quality of education" approach.

Theoretical and Methodological Objections

Marshall's critique of the idea of human capital was the most influential in the period preceding the human capital revolution itself. The status of the concept of human capital in Marshall's thought was controversial. The debate between Kiker (1966,

[8]As underlined by Teixeira (2005, p. 137): "A peculiar aspect was that the initial development was isolated and only after Schultz saw Mincer's dissertation and decided to invite him to Chicago for a post-doctoral fellowship (1957–1958), they interacted more closely. Then, they became aware of the closeness of their research and its *unplanned* complementarity." See also Biddle and Holden (2016).

1968) and Blandy (1967) is enlightening in this respect. The thrust of this debate lies in the question of whether Marshall analyzed human skills within the analytical schema of capital – which is precisely what was done by the proponents of the "human capital revolution." According to Kiker (1966, p. 481), "Marshall discarded the notion as 'unrealistic'." Yet Marshall did not neglect the importance of human investments and their economic consequences, quite the contrary. As quoted by Becker in the introduction to his landmark book (1964), Marshall wrote: "The most valuable of all capital is that invested in human beings" (Marshall 1890, p. 469).

Marshall built upon Smith's legacy to assert:

> The motives which induce a man and his father to invest capital and labour in preparing him for his work . . . are similar to those which lead to the investment of capital and labour in building up the material plant and the organization of a business. In each case the investment (so far as man's action is governed by deliberate motive at all) is carried up to that margin at which any further investment appears to offer a balance of gain, no excess or surplus of utility over 'disutility'. (Marshall 1890, p. 514)

But, according to Kiker, Marshall was reluctant to place human skills within the definitional scheme of capital; a view which Blandy (1967) opposed, even though he recognized that Marshall "felt that its [the idea of human capital] inclusion in the notions of wealth and capital in all circumstances was not in harmony with the usage of ordinary life and would lead to confusion if it were included in his general definition" (Blandy 1967, p. 875). Sweetland (1996) also supports Kiker's view, claiming that Marshall discarded the notion as unrealistic. Without a market for human capital, determining its value was a gamble. He was therefore not convinced that there was any possibility of providing a reliable estimation of the monetary value of human beings or their skills.

According to Schultz, Marshall's reluctance about the application of capital theory to individual skills and capabilities played a major role in the belated acceptance of the concept of human capital.

> But Fisher's approach to capital was not accepted by the mainstream of economists, mainly because of Marshall's adverse reactions to it, backed by his great prestige. Although Marshall at many points in his own work refers to the abilities man acquires by schooling and by working as an apprentice and to the economic role of knowledge, his view was that, whereas human beings are incontestably capital from an abstract and mathematical point of view, it would be impractical to extend the traditional market place concept of capital to include human capital. (Schultz 1972a, pp. 6–7)

Shaffer's commentary (1961) on Schultz's presidential address delivered at the annual meeting of the American Economic Association in 1960[9] offers a snapshot of the criticisms addressed to the idea of human capital at the time. These criticisms did not question the proposal that educational expenditures enhance productivity nor

[9]Schultz's presidential address to the American Economic Association in 1960 is seen as one of the fundamental moments in the emerging theory of human capital (see, for instance, Teixeira 2000).

did they deal with the fact that considering men or their abilities as a form of capital raised moral issues. The core of Shaffer's criticism is that disentangling the consumption and the investment part of total expenditures on man (education, health, etc.) is unfeasible.

As with Marshall's qualms developed 50 years earlier, Shaffer doesn't ignore the importance of human skills, and he accepts that it can be considered as a form of human capital but only in a metaphorical sense. He claims, in fact, that considering men and their abilities as a form of capital should be regarded as a metaphor but can be analyzed within the framework of and with the tools of capital theory.

> Any attempt to show that rational individuals tend to undertake expenditure on education up to the point where the marginal productivity of the human capital produced by the process of education equals the rate of interest – a point at which the marginal expenditure on education yields a return equal to the return on marginal expenditure for any other factor of production – would be a mockery of economic theory. (Shaffer 1961, p. 1028)

On this occasion, Schultz, Becker, and Mincer definitively won the battle against human capital's skeptics. But new offensives arose at the turn of the 1970s.

The new criticisms came, on one side, from competing methodological approaches. The radical school articulated its criticism around the rejection of the individual choice framework and took a firm stance against the methodological individualism that characterized human capital theory (see Bailly 2016). The objection it raised concerned the fact that the human capital research program traces the dynamics of the education system back to individual choices alone and neglects the role of class conflicts in determining inequalities in education: "Human capital theory is the most recent, and perhaps ultimate, step in the elimination of class as a central economic concept" (Bowles and Gintis 1975, p. 74).

The institutionalist approach – which had dominated labor economics before the rise of human capital theory (see White 2017) – also rejects the methodological orientation of the human capital research program. Institutionalist-oriented analyses of the dynamics of the education system reject the individual choice model as a basis for a theory of the supply of education (see Chirat and Le Chapelain 2017) and, in opposition to human capital theory, confer a leading role to corporations and the requirements of technology. "Nothing is more alien to the human capital research program than the manpower forecasters' notion of technically- determined educational requirements for jobs" (Blaug 1976, p. 846).

On the other side, within this set of challenges to human capital theory that flourished in the 1970s, criticisms also emerged from within mainstream economics. The filter theory of education (Arrow 1973) and the theory of "screening" (Stiglitz 1975) didn't attack the neoclassical methodological basis of the research program, but both put forward a criticism targeting a pivotal argument of the human capital theorists: namely, the positive link between education and productivity. These theories consider that education reflects an individual's productivity potential, but they oppose the view that educational investment contributes to the accumulation of skills that enhance agents' productivity.

Henceforth, as emphasized by Sobel at the beginning of the 1980s, "Human capital, with its individualistic approach, while still the dominant theory, is not the only game in town" (Sobel 1982, p. 268).

Skepticism About the Measurement of Human Capital

The first questions about the positive relationship between human capital and growth arose at the turn of the 1970s (see Demeulemeester and Diebolt 2011). In the economic context of the 1970s, marked by falling economic growth rates, doubts began to be expressed about the virtues of mass higher education (see notably Freeman 1976).

Endogenous theories of growth (Romer 1986, 1990, Lucas 1988) revived the idea of great importance for human capital on the growth process. This strand of literature expects there to be a positive relationship between growth and human capital *accumulation* (mainly, in these models, through education) (Lucas 1988) or – emphasizing the role of human capital on technological imitation and innovation – predicts that *stocks* of human capital are decisive as regards growth inequalities. In the 1990s, this literature, as well as the exogenous growth models incorporating human capital (Mankiw et al. 1992, for instance), led to a series of empirical exercises aimed at improving the understanding of the relationship between human capital (education) and growth. This empirical literature has given rise to puzzling results. Benhabib and Spiegel (1994) and Pritchett's (2001) studies, notably, have called into question the contribution of education to economic growth and hence have questioned the relevance of the human capital theoretic approach itself.

Several contributions have outlined the paucity of the educational proxies used in the empirical macro literature and their weakness in correctly reflecting different amounts of stocks and flows of knowledge and skills (see Woessmann 2003, Cohen and Soto 2007, Folloni and Vittadini 2010, and Prados de la Escosura and Roses 2010). According to these approaches, the ambiguous empirical findings about the contribution of human capital to economic growth can directly be explained by the poor quality of the data (as advanced by Cohen and Soto (2007)) or by the human capital proxies inappropriately reflecting the nature of the theoretical concept.

Woessmann (2003) has stressed the weaknesses of the three most commonly used human capital proxies in the empirical literature: the literacy rate, the enrollment rate, and the average years of schooling of the active population. According to the author, an appeal to literacy rates leads to an underestimation of the stock of human capital in an economy, as this proxy ignores all human capital investments that go beyond reading and writing itself. Nor do enrollment rates provide a more suitable appraisal of the endowments in human capital: as the students do not belong to the active population, assessing the effective stock of human capital of this population by using this indicator is likely to induce significant gaps. Appeal to the average number of years of schooling of the workforce does overcome the previously mentioned limitation. Different methods, based either on the enrollment rate (when sufficiently long-term data allow for it) or on censuses, are used to assess the average years of schooling of the active population (see, for instance, Lau et al. (1991) and

Nehru et al. (1995)), and this seems to constitute the least imperfect indicator of human capital among the indicators previously mentioned. This measure, however, has its own weakness: it considers that a year of education leads to an equivalent accumulation of human capital, independent of the grade which the year of education covers, i.e., its rank in the schooling hierarchy,[10] and independent of the quality of the education system itself.

This last channel, i.e., the effect of the unequal quality of teaching on the effective accumulation of human capital, has a vast literature devoted to it. The seminal work of Hanushek[11] contributed significantly to promoting the idea that purely quantitative educational variables were only poorly able to illuminate the role of human capital on wages at the individual level or on economic growth at the macroeconomic level.

This field of research firstly considered different educative outputs (education-related spending, size of classrooms, level of training of the teachers, etc.) to reflect the differences in the quality of education systems and their impact on the accumulation of human capital (see, for instance, Behrman and Birdsall (1983)). Progressively, the ability of these indicators to accurately reflect differences in the quality of schooling has been called into question. The use of scores in international tests (e.g., PISA, TIMSS) now permits a more direct evaluation of these differences by measuring the skills accumulated through education.

The research program on the quality of education has made salient the difficulty of measuring human capital. First, due to inequalities in the quality of education, the same amount of education can lead to different levels of human capital accumulated. Second, research has shown that education quality matters for growth and, accordingly, that the human capital proxies commonly used have failed to properly estimate human capital endowments at the macro level.

On this basis, the research program on education quality has taken a somewhat radical turn. In their latest book, Hanushek and Woessmann (2015) come down hard on the state of the human capital research program. Their criticism now centers on the view that human capital endowments at the macro level are very badly calculated due to the use of inappropriate measures. As a result, the message sent by the literature on the role of education on economic growth is fundamentally distorted. They claim that the concept of human capital should be abandoned in favor of a new concept of "knowledge capital."

> The conclusion of the analysis we develop in this book is that Adam Smith was right: human capital, as we now call it, is extraordinarily important for a nation's economic development. The significance of education, however, has been obscured by measurement issues. (Hanushek and Woessmann 2015, p. 2)

[10]Microeconometric approaches have highlighted the existence of decreasing private returns of education (see notably Card (1999) and Psacharopoulos and Patrinos (2004)).

[11]Eric Hanushek significantly contributed to the development of the so-called approaches of the "quality of education." His early work on the subject began in 1970. Regarding the role of the quality of education in the contribution of human capital to growth, see, for instance, Hanushek and Kimko (2000), Jamison et al. (2007), or Hanushek and Woessmann (2008, 2011, 2012).

Hanushek and Woessmann's critique of the concept of human capital does not turn on doubts about the relevant theoretical framework – i.e., the theoretical logic driving the human capital program – but rather on the inability of this framework to accurately contribute to economic policy recommendations due to the overwhelming empirical difficulties it faces. They emphasize that the measurement problem thus seriously threatens the credibility of the human capital research program.

The measurement of human capital is a major issue for cliometric approaches to the question of the industrialization process. These approaches have recently opened up new avenues of thought on the measurement problem, which though not so radical as to propose the abandonment of the concept tout court, nevertheless raise knotty and fascinating issues.

Section III. Human Capital and the Industrialization Process

The human capital revolution at the end of the 1950s coincided with another revolution, albeit a distinct one, namely, the development of the *New Economic History,* or cliometrics, also initiated in the United States (Diebolt and Haupert 2017). In its analysis of the determinants of growth, the cliometric approach quickly came to appropriate the theoretical synthesis fostered by Schultz, Becker, and Mincer.

The issue of the growth process, which motivated the development of human capital theory at the end of the 1950s, is a prime question for cliometric studies (see Goldin 2016). The understanding of the determinants of economic growth has in fact been a main challenge for cliometrics ever since its origin. In tackling this field of inquiry, cliometricians have naturally built on the contributions provided by human capital theorists, and the concept of human capital is widely used in the examination of national or regional growth trajectories. The concept has also been appealed to in support of cliometrics's new interest in unified growth theories, which builds on the idea of the inverse relationship between fertility and human capital to provide a unified explanatory model of the transition from Malthusian stagnation to modern growth (Galor and Weil 2000, Galor and Moav 2002, Galor 2011).

In the investigation of the causal relationships linking human capital and growth, namely, in its attempt to illuminate the contribution of the endowments in human capital to the growth processes, the cliometric synthesis was primarily applied in a specific field of investigation: the industrialization process experienced by Western European countries in the eighteenth and nineteenth centuries. It is in fact out of the industrial revolution, and the associated massive accumulation of human capital, that the complex question of the underlying relationships linking these two phenomena arises.

The role of human capital at the time of the second industrial revolution is close to the predictions of contemporary theory, which considers human capital as a driver of growth and defends the idea of "skill-biased technological change." In contrast, and paradoxically, the role of human capital in the first stage of the industrialization

process has long been regarded as minor (Human Capital and Early Industrialization: A Paradox in Economic History). This dominant view is currently under review (The Paradox Under Review: Re-evaluating What Human Capital Is).

Human Capital and Early Industrialization: A Paradox in Economic History?

As regards human capital and early industrialization, cliometricians have addressed two main issues. On one side, a large strand of literature has examined the importance of human capital as an explanatory factor in the industrial takeoff and of the subsequent transition to the modern regime of economic growth. On the other, the effect of the industrialization process on the formation of human capital has prompted attention. By showing that human capital was not a key factor in the industrial takeoff and that the industrialization process did not trigger human capital accumulation, these two interrelated strands of research have emphasized the inconsistency between neoclassical growth models – and of human capital theory – and the features that had characterized the first stage of the industrialization process.

Indeed, the early literature in this field didn't see human capital as a key factor in the industrial takeoff. On the basis that Britain experienced a low and stagnant level of education at the time of its industrial takeoff, Mitch (1999) was one of the first to assert that human capital endowments didn't play a major role in the British industrial revolution. Allen (2003) and Clark (2005) also support the view that human capital theory cannot provide a convincing explanation of the British transition from the Malthusian era to modern growth. Sandberg (1979) helped diffuse the same idea by highlighting that despite a large stock of human capital, in the mid- nineteenth century, Sweden experienced low and stagnant income levels ("the theory of the impoverished sophisticate"). Emphasizing the role of culture and values, Mokyr (2016) and McCloskey (2006, 2010, 2016) – though on different grounds – provide a more nuanced account of the role of human capital in the transition to modern growth. For McCloskey, the great enrichment experienced by Western European countries since the eighteenth century is rooted in the evolution of culture and ideas, and more precisely in changes to expressed attitudes to capitalism and in changes in thought about the bourgeois ("The bourgeois 'Revaluation'"), but not in capital accumulation nor human capital accumulation (nor in institutions, coal, etc.). Mokyr's argument, while distinct from McCloskey's, also stresses the importance of culture – the culture of the intellectual elites, not the role of bourgeois culture – as a key root of the great enrichment. Mokyr points out that the transition to modern economic growth in eighteenth-century Europe must be regarded as the result of the development in Europe between 1500 and 1700 of new values and new ideas that diffused notably through the influence of the *Republic of letters*. This new culture – and not the narrow differences in literacy rates – is regarded as decisive for having set the stage for the industrial revolution.

Besides the thesis that human capital was unimportant to the industrial takeoff, early cliometric contributions dealing with the effect of the industrialization process of the eighteenth and nineteenth centuries on human capital accumulation came to

the conclusion that industrialization was, first, a deskilling process. Early views on the issue have argued that industrialization and technological change were not skill-demanding but, on the contrary, at least in the first stage increased the relative demand for unskilled labor (see, for instance, Nicholas and Nicholas 1992, Mokyr 1993, Mitch 1999). Expansion of production and the mechanization of industry notably increased the demand for female and child unskilled labor (Sanderson 1972), thus impairing school participation and human capital accumulation. Early contributions on the issue have all suggested that the tasks required by booming production during the first stage of the industrialization process didn't require a workforce with a high level of skills. According to Goldin and Katz (1998), the complementarity between innovation and skills took place in the early twentieth century with the technological shift from steam power to electricity.

Overall, whether the issue is the determinants of the takeoff or the complementarity between industrialization and skills, cliometric analyses have spread the message that human capital mattered little in the first phase of industrialization. But this conclusion is currently under reconsideration.

The Paradox Under Review: Re-evaluating What Human Capital Is

A new wave of analysis is currently challenging the idea that human capital is not a relevant dimension in the analysis of the industrialization process. Interestingly, this recent literature has developed on the basis that early accounts of the human capital-industrialization relationship are subject to limitations because of their unsatisfactory evaluation of human capital endowments. This has opened up new paths of reflection about what human capital is and hence how to measure it.

On the one hand, efforts have been made to improve the quality of the human capital estimates. Jacob (2014) stresses the weaknesses of British educational records prior to 1850. According to her, the weak role attributed to education in the first industrial revolution partly derives from the poor state of the relevant historical sources. Work by De Pleijt (2018) provides a reassessment of the evolution of formal education in Britain between 1307 and 1900 by estimating – on the basis of information on literacy rates, primary, secondary, and tertiary schooling – average years of education over the period. These estimations bring to light that literacy rates alone are misleading, since they underestimate human capital accumulation in the period of the industrial revolution. Trends in the evolution of educational attainment are rather poorly reflected in literacy rates on the eve of the industrial revolution.

Madsen and Murtin (2017) also provide a reassessment of British human capital over the long run, based on educational attainment at primary, secondary, and tertiary levels. Analyzing the determinants of GDP growth over the period 1270–2010, they come to the conclusion that education was the prime driver of British economic growth over the period, a result which contrasts starkly with early contributions. They also put forward the fact that primary education was of prime importance in explaining growth in the pre-1750 period and that the role of secondary and tertiary education became predominant from 1750 onward.

More generally, faced with the difficulties raised by the nonavailability of historical data, the cliometric approach has developed new and innovative strategies to measure human capital. Within the spirit of the procedure based on marriage registers, the *age heaping method* has recently been developed (see, for instance, A'Hearn et al. 2009, Crayen and Baten 2010, Baten et al. 2014, Baten and Fourie 2015, Tollnek and Baten 2016, Cappelli and Baten 2017). This method relies on "the tendency of poorly educated people to round their age erroneously. For example, they answer more often '40', if they are in fact 39 or 41, compared with better educated people" (Crayen and Baten 2010, p. 452). Age heaping uses accuracy of age reporting as a proxy for numeracy.

On the other hand, another path of research builds on the idea that human capital endowments are badly evaluated, not because of the quality of the data or of the estimates used in empirical exercises but because assessing human capital through literacy rates or crude enrolment rates only (the common practice in empirical exercises at the macro level) shows just one side of a larger picture.

In the context of measuring human capital endowments, the effort required for an extensive understanding of the historical context necessarily highlights the need to keep the institutional and organizational features of the education systems under close review. From this knowledge, highly aggregated educational attainment measures of human capital, which are commonly used in early literature, quickly become unsatisfactory as they cannot reveal the differences in national experiences – i.e., the orientation of knowledge taught (general vs. vocational), the importance of apprenticeships, the presence of knowledge elites, the degree of centralization of the education, the way it was financed, etc. All of these dimensions may have an impact on the level of human capital effectively accumulated at the aggregate level but are ignored as soon as empirical approaches come to rely on aggregate enrolment rates or literacy rates only.

The need for a proper assessment of human capital has thus triggered reflections on the education systems per se and more specifically about their specificities at the institutional and organizational levels. The specific understanding of these specificities and their manner of inclusion in the establishment of relevant indicators of human capital are derived from the methodological nature of cliometrics itself. Because its methodology necessitates developing linkages between theory and measurement within the analysis of a specific historical context, the cliometric approach has opened new perspectives on the human capital empirical problem.

> Clearly human capital as a concept is indispensable, but we need to be far more specific as to what kind of human capital was produced, for and by whom, what was the source of the demand for it, and how it was distributed over the population. (Mokyr 2005, p. 1155)

At stake in this revisionist literature, which builds on the idea that there are different types of human capital, is the determination of what type was important for industrialization. Yet the early literature, by the way it measured human capital, has only considered differences in basic knowledge, namely, differences in literacy, without considering the other forms of accumulated skills.

Some approaches underline the idea that the human capital of the elite was decisive in the industrialization process, not the basic or average level of human capital (of the country), i.e., the human capital of the mass of workers. Mokyr (1990, 2005) has emphasized the need to "focus not on the mean level of human capital [. . .], but just on the *density in the upper tail* of the distribution, that is, the level of education and sophistication of a small and pivotal elite of engineers, mechanics, and chemists" (Mokyr 2005, p. 1157). Emphasizing that science and technique were concretely connected in the industrialization process, Jacob (1997, 2014) places at the forefront the role of scientists' and innovators' human capital. In that vein, Meisenzahl and Mokyr (2012) outline the key influence of the human capital of the elite in the British industrialization process. Because innovation and technological change were at the core of the industrial revolution, they argue that prime importance must be attributed to "the top 3 to 5 percent of the labor force in terms of skills: engineers, mechanics, millwrights, clock and instrument makers, skilled carpenters and metal workers, and wheelwrights."

Recently, Squicciarini and Voigtländer (2015) have developed a similar thesis about French industrialization. Relying on subscription density to the *Encyclopédie* (Diderot, d'Alembert) to measure the repartition and the density of French scientific elites (*knowledge elite)*, they show that the human capital of the elites in the eighteenth century, unlike the average human capital, played a significant role on the subsequent French industrial takeoff, starting in the first half of the nineteenth century.

Besides approaches that focus on the elites, analysis concentrating on the role of the apprenticeship system has also played a role in reviving the idea that human capital was of importance in the first stage of the industrialization process. A key argument behind the idea of a key role for apprenticeships is that the skills accumulated through it – the skills of the technical or mechanical workers – were decisive in fostering technological change, not because they promote technological innovation but because they were crucial in promoting the capacity to implement devices and maintain them while also promoting incremental innovation on the shop floor. "These characteristics did not require much science or even originality, but they needed people who were good with their hands and had been taught how to use them. It is that resource that the apprenticeship institution in Britain supplied better than anywhere else" (Zeev et al. 2017, p. 245). Zeev et al. (2017) suggest that the flexibility and effectiveness of Britain's system of apprenticeship is an important factor in explaining why England entered first into industrialization. Kelly et al. (2014, p. 364) assert explicitly that if the early literature has rejected human capital as a key driver of industrialization, it is because it "focuses on the wrong variables." For them too, Britain had a decisive advantage in terms of the quality of labor. The high level of technical competence of artisans and workers was linked, according to them, to the effective British apprenticeship system, and this explains the rapid industrialization. Humphries (2003) also sees in apprenticeship a decisive explanatory dimension behind early industrialization in Britain, pointing out that, in connection with the Elizabethan Poor Law, it played a role in reallocating labor to industry in eighteenth-century England. Focusing on pre-industrial

growth, De La Croix et al. (2018) emphasize the role of apprenticeship as a driver for *tacit* knowledge diffusion and stress its importance in the technological and economic advance of Western Europe in the century before the industrial revolution.

The importance of the British apprenticeship system in the second stage of the British industrialization process has been emphasized by Broadberry (2005, 2006), who opposes the widely held view that low levels of education in the British labor force, as compared to Germany or the United States, explained the British productivity decline at the end of the nineteenth century. In order to provide a convincing analysis of the relationship between human capital and productivity performance, he sought to disaggregate productivity indicators (according to the different sectors of the British economy) from education variables. By considering formal education and vocational training, he shows that "the key institutions of human capital accumulation in Britain were the system of apprenticeships and the body of professional associations" (Broadberry 2006, p. 128). Through the examination of the whole education system, he rejects the view that human capital endowment in Britain was low or that it explains the British productivity decline from the end of the nineteenth century until the eve of World War I. He therefore puts emphasis on the necessity of considering "human capital accumulation strategies" in order to provide proper measures of human capital. In fact, commonly used proxies of human capital, such as literacy rates or average years of education of the workforce, would have completely concealed what historical analysis has here brought to light: the key importance of specific vocational paths of education in Britain at the end of the nineteenth century.

Turning to the still open debate on the role that industrialization played on the accumulation of human capital, investigations into other forms of human capital (i.e., other than literacy standards and enrolment rates in primary education) have called into question the early consensus that industrialization was a deskilling process. Franck and Galor's (2017) seminal analysis was the first to question the idea that the path of economic development during the first industrial revolution was characterized by unskilled-based technological change. Examining early French industrialization (1839–1847), they show that technological change did indeed foster human capital accumulation. Showing that the number of apprentices and their share in the cohort of 15-year-olds increased in response to inventions, Feldman and Van der Beek (2016) also claim that technological progress was conducive to skills acquisition in eighteenth-century England. De Pleijt and Weisdorf (2017) show that there was a large decrease in average skills in agriculture and industry from the end of the sixteenth century to the beginning of the nineteenth century in England. They claim that deskilling generally occurred along with technological progress, despite a modest increase in the share of "high-quality" workers. This finding gives support to the view, already defended by Mokyr (1990, 2005) and more recently for the French case by Squicciarini and Voigtländer (2015), that upper-tail knowledge played a prominent role in early industrialization. For De Pleijt et al. (2016), the effect of industrialization on human capital is mixed. They show that

in the first phase of British industrialization, more steam engines were associated with a higher skilled labor force, and thus they claim that technological adoption was skill-demanding. But they also highlight that adopting new technologies was not conducive to elementary education (where this is approximated by literacy rates and enrolment rates). Diebolt et al. (2017) distinguish the accumulation of basic human capital (basic literacy skills) from intermediate human capital (basic scientific and technical knowledge and basic knowledge in law and trade) in nineteenth-century France. They show that French industrialization was not deskilling but that a shift in the kind of skills required occurred in the second half of the nineteenth century. Focusing on lifelong training, Diebolt et al. (2018) provide evidence that technological change contributed significantly to the development of adult education during the period 1850–1881 in France. They show that steam technology adoption was skill-demanding, highlighting that it triggered the development of evening classes for workers and apprentices.

Finally, the education systems at the time of the first industrial revolution evolved following different patterns, which privileged different types of institutional organizations. The focus on general vs. vocational education, the nature of the skills taught, the degree of centralization, types of funding – private or public – but also their attitudes toward social and gender inequalities (girls' access to education) are important dimensions that single out the education systems of the countries that entered industrialization in the eighteenth and nineteenth centuries. These dimensions also inflect the respective paths of human capital accumulation.

Goldin (1998, 2001, 2016) and Goldin and Katz (1999, 2008) have shown that the US educational model differed profoundly from European elitism, having been based since the nineteenth century on the principle of egalitarianism. Several characteristics reflect this principle of egalitarianism: public financing, the separation of church and state, the decentralized functioning, and the access of girls to education (gender neutrality). Goldin and Katz (2008) point out that this orientation of the education system constitutes an incremental ingredient of the accumulation of human capital in the United States and of the economic success of the country in the twentieth century.

Contrary to Mitch's (1999) findings, which call into question the contribution of formal education to the British takeoff, Becker et al. (2011) show that the Prussian education system, characterized by the large and early diffusion of elementary education, created favorable conditions for the adoption of existing technologies in Prussia and therefore contributed to its industrial takeoff. Consequently, they assess the importance of the education system in the process of industrial catch-up in Prussia in the nineteenth century. Analyzing primary, secondary, and higher education separately, they argue that secondary and higher education were not of significant importance in this process.

Cinnirella and Streb (2017) consider different types of human capital and their influence on innovation at the time of the second industrial revolution in Prussia. They outline how highly skilled craftsmen were still important – as recent literature has maintained in the case of the first stage of industrialization – for the innovation

process in the second stage of Prussian industrialization. Contrary to the first phase, they claim that the quality of basic education became important at the end of the nineteenth century, since it fostered both labor productivity and firm innovation. In addition, Cappelli (2016) has recently shown that the centralization that characterized the organization of the Italian education system in the nineteenth century negatively influenced human capital accumulation. His analysis shows that the move to decentralized primary schooling in the late nineteenth century was an important driver for the increase in Italy's human capital endowment.

Conclusion

This chapter has sketched the development of the study of human capital. We have noted some of the challenges to the human capital revolution. There were theoretical objections to the idea of analyzing human skills within the framework of capital theory and opposition to the methodological orientation of the program. The radical school and the institutionalist approaches both criticized the failure of human capital theory to adequately describe the mechanisms and driving forces behind the dynamics of human capital accumulation.

Somewhat surprisingly, however, the most severe criticism directed against the human capital program comes from what we might describe as an operational problem of measurement. Currently, the deadlock faced by human capital as regards the issue of empirically measuring stocks of skills (at the macro level) severely threatens the achievements of the late 1950s. We have emphasized in this chapter that, in its analysis of the role played by human capital in the first phase of the industrialization process, the cliometric revisionist literature paves the way toward renewed insights into the measurement of human capital. Due to its particular methodological character, lying at the crossroads between theory, measurement, and history, the cliometric synthesis promotes a reformulation of the question of the measurement of human capital in a unique and, we believe, successful way, as regards the empirical challenges faced by the concept. The question of measurement is raised through the question of national and/or regional strategies of human capital accumulation and of their efficacy regarding economic growth. This distinction is not trivial. It indicates that the difficulties of measuring human capital should not be reduced to a purely operational problem but are linked with a more fundamental challenge: the accurate and precise understanding of education spending, at the aggregate level, that effectively contributes to the accumulation of human capital, i.e., by relying precisely on investments, and not on a form of consumption. This perspective brings us back to an original challenge of the human capital revolution, which was to disentangle the investment and the consumption part of education expenditure. This challenge has never been fully taken up; which led Schultz to assert that "a satisfactory theory of economic growth should explain the mechanism that determines the formation of human and nonhuman capital, including the accumulation of knowledge" (Schultz 1972b, p. S6). Cliometric approaches have recently embarked on this route.

References

A'hearn B, Baten J, Crayen D (2009) Quantifying quantitative literacy: age heaping and the history of human capital. J Econ Hist 69(3):783–808
Abramovitz M (1956) Resource and output trends in the United States since 1870. Resource and output trends in the United States Since 1870, NBER, pp 1–23
Allen RC (2003) Progress and poverty in early modern Europe. Econ Hist Rev 56(3):403–443
Arrow K (1973) Higher education as a filter. J Public Econ 2(3):193–216
Bailly F (2016) The radical school and the economics of education. J Hist Econ Thought 38(03):351–369
Baten J, Fourie J (2015) Numeracy of Africans, Asians, and Europeans during the early modern period: new evidence from Cape Colony court registers. Econ Hist Rev 68(2):632–656
Baten J, Crayen D, Voth H (2014) Numeracy and the impact of high food prices in industrializing britain, 1780–1850. Rev Econ Stat 96(3):418–430
Becker GS (1962) Investment in human capital: a theoretical analysis. J Polit Econ 70(5):9–49
Becker, Gary (1964) Human Capital. A Theoretical and Empirical Analysis with Special Reference to Education, Columbia University Press, New York
Becker SO, Hornung E, Woessmann L (2011) Education and catch-up in the industrial revolution. Am Econ J Macroecon 3(3):92–126
Behrman J, Birdsall N (1983) The quality of schooling: quantity alone is misleading. Am Econ Rev 73:928–946
Benhabib J, Spiegel MM (1994) The role of human capital in economic development evidence from aggregate cross-country data. J Monet Econ 34(2):143–173
Biddle J, Holden L (2016) The introduction of human capital theory into education policy in the United States. Working paper
Blandy R (1967) Marshall on human capital: a note. J Polit Econ 75:874–875
Blaug M (1966) Economics of education: a selected annoted bibliography. Pergamon Press, Oxford
Blaug M (1976) The empirical status of human capital theory: a slightly jaundiced survey. J Econ Lit 14(3):827–855
Bowles S, Gintis H (1975) The problem with human capital theory – a Marxian critique. Am Econ Rev 65(2):74–82
Bowman JM (1964) Schultz, Denison, and the contribution of "Eds" to national income growth. J Polit Econ 72(5):450–464
Bowman JM (1966) The human investment revolution in economic thought. Sociol Educ 39(2):111–137
Bowman JM (1980) On Theodore W. Schultz's contributions to economics. Scand J Econ 82(1):80–107
Broadberry SN (2005) The productivity race: British manufacturing in international perspective, 1850–1990. Cambridge University Press, Cambridge, UK
Broadberry SN (2006) Human capital and productivity performance: Britain, the United States and Germany, 1870–1990. In: David PA, Thomas M (eds) The economic future in historical perspective. Oxford University Press, Oxford
Cappelli G (2016) Escaping from a human capital trap? Italy's regions and the move to centralised primary schooling, 1861–1936. Eur Rev Econ Hist 20(1):46–65
Cappelli G, Baten J (2017) European trade, colonialism and human capital accumulation in Senegal, Gambia and Western Mali, 1770–1900. J Econ Hist 77(3):920–951
Card D (1999) The causal effect of education on earnings. In: Handbook of labor economics, vol 3. Elsevier, Amsterdam, pp 1801–1863
Chirat A, Le Chapelain C (2017) Some 'unexpected proximities' between Schultz and Galbraith on human capital. Working paper BETA, 2017–2018
Cinnirella F, Streb J (2017) The role of human capital and innovation in economic development: evidence from post-Malthusian Prussia. J Econ Growth 22(2):193–227
Clark G (2005) The condition of the working class in England, 1209–2004. J Polit Econ 113(6):1307–1340

Cohen D, Soto M (2007) Growth and human capital: good data, good results. J Econ Growth 12(1):51–76
Crayen D, Baten J (2010) New evidence and new methods to measure human capital inequality before and during the industrial revolution: France and the US in the seventeenth to nineteenth centuries. Econ Hist Rev 63(2):452–478
De la Croix D, Doepke M, Mokyr J (2018) Clans, guilds, and markets: apprenticeship institutions and growth in the preindustrial economy. Q J Econ 133(1):1–70
De Pleijt AM (2018) Human capital formation in the long run: evidence from average years of schooling in England, 1300–1900. Cliometrica 12(1):99
De Pleijt AM, Weisdorf J (2017) Human capital formation from occupations: the 'deskilling hypothesis' revisited. Cliometrica 11(1):1–30
De Pleijt AM, Nuvolari A, Weisdorf J (2016) Human capital formation during the first industrial revolution: evidence from the use of steam engines (No. 294). Competitive Advantage in the Global Economy (CAGE)
Demeulemeester J-L, Diebolt C (2011) Education and growth: what links for which policy? Hist Soc Res 36:323–346
Denison EF (1962) Sources of economic growth in the United States and the alternatives before us. Committee for Economic Development, New York
Diebolt C, Haupert M (2017). A cliometric counterfactual: what if there had been neither Fogel nor North?. Cliometrica 12(3):407–434
Diebolt C, Le Chapelain C, Ménard A (2017) Industrialization as a deskilling process? Steam engines and human capital in XIXth century France. Working paper BETA, 2017-17, Working paper AFC 7–2017
Diebolt C, Le Chapelain C, Ménard A (2018) Learning outside the factory: the impact of technological change on the rise of adult education in nineteenth-century France. Working paper AFC 2-2018
Ehrlich I, Murphy KM (2007) Why does human capital need a journal? J Hum Cap 1(1):1–7
Fabricant S (1954) Economic progress and economic change. National Bureau of Economic Research, New York
Feldman NE, van der Beek K (2016) Skill choice and skill complementarity in eighteenth century England. Explor Econ Hist 59(January):94–113
Folloni G, Vittadini G (2010) Human capital measurement: a survey. J Econ Surv 24:248–279
Franck R, Galor O (2017) Technology-skill complementarity in the early phase of industrialization. IZA discussion paper series no 9758
Freeman R (1976) The overeducated American. Academic, New York
Galor O (2011) Unified growth theory. Princeton University Press, Princeton
Galor O, Moav O (2002) Natural selection and the origin of economic growth. Q J Econ 117(4):1133–1191
Galor O, Weil DN (2000) Population, technology, and growth: from Malthusian stagnation to the demographic transition and beyond. Am Econ Rev 90(4):806–828
Goldin C (1998) America's graduation from high school: the evolution and spread of secondary schooling in the twentieth century. J Econ Hist 58(02):345–374
Goldin C (2001) The human capital century and American leadership: virtues of the past. J Econ Hist 61:263–291
Goldin C (2016) Human capital. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer, Heidelberg, pp 55–86
Goldin C, Katz LF (1998) The origins of technology-skill complementarity. Q J Econ 113(3):693–732
Goldin C, Katz L (1999) The shaping of higher education: the formative years in the United States, 1890 to 1940. J Econ Perspect 13:37–62
Goldin C, Katz L (2008) The race between education and technology. Harvard University Press, Cambridge

Hanushek E, Kimko D (2000) Schooling, labor force quality, and the growth of nations. Am Econ Rev 90(5):1184–1208
Hanushek E, Woessmann L (2008) The role of cognitive skills in economic development. J Econ Lit 46(3):607–668
Hanushek E, Woessmann L (2011) The economics of international differences in educational achievement. In: Hanushek E, Machin S, Woessmann L (eds) Handbook of the economics of education, vol 3. North Holland, Amsterdam, pp 89–200
Hanushek E, Woessmann L (2012) Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. J Econ Growth 17(4):267–321
Hanushek E, Woessmann L (2015) The knowledge capital of nations: education and the economics of growth. MIT Press, Cambridge, MA
Hodgson GM (2007) Meanings of methodological individualism. J Econ Methodol 14(2):211–226
Humphries J (2003) English apprenticeship: a neglected factor in the first industrial revolution. In: David PA, Thomas M (eds) The economic future in historical perspective. Oxford University Press, Oxford
Jacob MC (1997) Scientific culture and the making of the industrial west. Oxford University Press, New York
Jacob MC (2014) The first knowledge economy: human capital and the European economy, 1750–1850. Cambridge University Press, Cambridge
Jamison E, Jamison D, Hanushek E (2007) The effects of education quality on income growth and mortality decline. Econ Educ Rev 26:772–789
Kelly M, Mokyr J, Gráda CÓ (2014) Precocious Albion: a new interpretation of the British industrial revolution. Annu Rev Econ 6(1):363–389
Kiker BF (1966) The historical roots of the concept of human capital. J Polit Econ 74(5):481–499
Kiker BF (1968) Marshall on human capital: comment. J Polit Econ 76(5):1088–1090
Lau LJ, Jamison DT, Louat FF (1991) Education and productivity in developing countries: an aggregate production function approach, WPS 612. World Bank Publications
Le Chapelain C et Matéos S (2018) Schultz et le capital humain: une trajectoire intellectuelle, working paper AFC 2018
Lucas RE (1988) On the mechanics of economic development. J Monet Econ 22:3–42
Madsen JB, Murtin F (2017) British economic growth since 1270: the role of education. J Econ Growth 22(3):229–272
Mankiw NG, Romer D, Weil DN (1992) A contribution to the empirics of economic growth. Q J Econ 107(2):407–437
Marshall A (1890) Principles of economics. Macmillan, London, (eight edition, 1920)
McCloskey DN (2006) The bourgeois virtues: ethics for an age of commerce. University of Chicago Press, Chicago
McCloskey DN (2010) Bourgeois dignity: why economics can't explain the modern world. University of Chicago Press, Chicago
McCloskey DN (2016) Bourgeois equality: how ideas, not capital or institutions, enriched the world, vol 3. University of Chicago Press, Chicago
Meisenzahl R, Mokyr J (2012) The rate and direction of invention during the industrial revolution: incentives and institutions. In: Lerner J, Stern S (eds) The rate and direction of inventive activity revisited. University of Chicago Press, Chicago, pp 443–479
Mincer J (1957) A study on personal income distribution. PhD Dissertation, Columbia University
Mincer J (1958) Investment in human capital and personal income distribution. J Polit Econ 66(4):281–302
Mitch D (1999) The role of education and skill in the British industrial revolution. In: Mokyr J (ed) The British industrial revolution: an economic perspective, 2nd edn. Westview, Boulder, pp 241–279
Mokyr J (1990) The lever of riches: technological creativity and technological progress. Oxford University Press, Oxford

Mokyr J (1993) The new economic history and the industrial revolution. In: Mokyr J (ed) The British Industrial Revolution: an economic perspective. Westview, Boulder, Editor's Introduction
Mokyr J (2005) Long-term economic growth and the history of technology. In: Aghion P, Durlauf S (eds) Handbook of economic growth, vol 1. Elsevier, Amsterdam, pp 1113–1180
Mokyr J (2016) A culture of growth: the origins of the modern economy. Princeton University Press, Princeton
Nehru V, Swanson E, Dubey A (1995) A new database on human capital stock in developing and industrial countries: sources, methodology, and results. J Dev Econ 46(2):379–401
Nerlove M (1999) Transforming economics: Theodore W. Schultz, 1902–1998: in memoriam. Econ J 109:F726–F748
Nicholas SJ, Nicholas JM (1992) Male literacy, "Deskilling," and the industrial revolution. J Interdiscip Hist 23(1):1–18
Prados de la Escosura LP, Rosés JR (2010) Human capital and economic growth in Spain, 1850–2000. Explor Econ Hist 47(4):520–532
Pritchett L (2001) Where has all the education gone ? World Bank Econ Rev 15:367–391
Psacharopoulos G, Patrinos HA (2004) Returns to investment in education: a further update. Educ Econ 12(2):111–134
Romer PM (1986) Increasing returns and long-run growth. J Polit Econ 94(5):1002–1037
Romer PM (1990) Endogenous technological change. J Polit Econ 98(5):S71–S102
Rosen S (2008) Human capital. In: Durlauf SN, Blume LE (eds) The new Palgrave dictionary of economics, 2nd edn. Palgrave, Basingstoke
Sandberg LG (1979) The case of the impoverished sophisticate: human capital and Swedish economic growth before World War I. J Econ Hist 39(1):225–241
Sanderson M (1972) Literacy and social mobility in the industrial revolution in England. Past Present 56:75–104
Schultz TW (1943) Redirecting farm policy. MacMillan, New York, États-Unis
Schultz TW (1958) The emerging economic scene and its relation to high-school education. In: Chase FS, Anderson HA (eds) The high school in a new era. University of Chicago Press, Chicago, pp 97–109
Schultz TW (1959) Investment in man: an economist's view. Soc Serv Rev 33(2):109–117
Schultz TW (1960) Capital formation by education. J Polit Econ 60:571–583
Schultz TW (1961a) Investment in human capital. Am Econ Rev 51(1):1–17
Schultz TW (1961b) Investment in human capital: reply. Am Econ Rev 51(5):1035–1039
Schultz TW (1962) Reflections on investment in man. J Polit Econ 70(5):1–8
Schultz TW (1972a) Human capital: policy issues and research opportunities. National Bureau of Economic Research, New York, pp 1–84
Schultz TW (1972b) Optimal investment in college instruction: equity and efficiency. J Polit Econ 80(3, Part.2):S2–S30
Schultz TW (1992) Adam smith and human capital. In: Fry M (ed) Adam Smith's legacy. His place in the developement of modern economics. Routledge, London
Shaffer HG (1961) Investment in Human Capital: comment. Am Econ Rev 51(5):1026–1035
Smith A (1776) An inquiry into the nature and causes of the Wealth of Nations, 2 vol [éd. par R.H. Campbel et A.S. Skinner]. Clarendon Press, Oxford, 1976
Sobel I (1978) Human capital revolution in economics development: current status, expectations and realities. Comp Educ Rev, 25th Anniversary Issue
Sobel I (1982) Human capital and institutional theories of the labor market: rivals or complements? J Econ Issues 16(1):255–272
Solow RM (1956) A contribution to the theory of economic growth. Q J Econ 70:65–94
Solow RM (1957) Technical change and the aggregate production function. Rev Econ Stat 39 (3):312–320
Spengler JJ (1977) Adam Smith on human capital. Am Econ Rev 67(1):32–36

Squicciarini M, Voigtländer N (2015) Human capital and industrialization: evidence from the age of enlightenment. Q J Econ 30(4):1825–1883
Stiglitz JE (1975) The theory of "screening," education, and the distribution of income. Am Econ Rev 65(3):283–300
Sweetland SR (1996) Human capital theory: foundations of a field of inquiry. Rev Educ Res 66(3):341–359
Teixeira PN (2000) A portrait of the economics of education, 1960–1997. Hist Polit Econ 32:257–288
Teixeira PN (2005) The 'Human capital revolution' in economics. Hist Econ Ideas:129–148
Teixeira PN (2007) Jacob Mincer. Oxford University Press, Oxford, UK
Teixeira PN (2011) A reluctant founding father: placing Jacob Mincer in the history of (labor) economics. Eur J Hist Econ Thought 18(5):673–695
Teixeira PN (2014) Gary Becker's early work on human capital – collaborations and distinctiveness. IZA J Labor Econ 3(12):1–20
Tinbergen J (1942) Zur theorie der langfristigen wirtschaftsentwicklung. Weltwirtschaftliches Arch:511–549
Tollnek F, Baten J (2016) Age-heaping-based human capital estimates. In: Diebolt and Haupert (eds) Handbook of Cliometrics, Springer Verlag, pp 131–154
White LH, (2017) Human capital and its critics: Gary Becker, institutionalism, and anti- neoliberalism. GMU working paper in economics no 17-02
Woessmann L (2003) Specifying human capital. J Econ Surv 17(3):239–270
Zeev NB, Mokyr J, Van Der Beek K (2017) Flexible supply of apprenticeship in the british industrial revolution. J Econ Hist 77(1):208–250

Age-Heaping-Based Human Capital Estimates

Franziska Tollnek and Joerg Baten

Contents

F. Tollnek (✉)
University of Tuebingen, Tuebingen, Germany
e-mail: franziska.tollnek@uni-tuebingen.de

J. Baten
University of Tuebingen and CESifo, Tuebingen, Germany
e-mail: joerg.baten@uni-tuebingen.de

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_24

Abstract

In this chapter, we provide comprehensive insights into the implementation and the use of the age-heaping method. Age heaping can be applied to approximate basic numerical skills and hence basic education. We discuss the advantages and potential issues of different indicators, and we show the relationship of those indicators with literacy and schooling. The application of age-heaping-based indicators enables us to explore various topics on basic education such as the gender gap and the divergence of countries in the very long run. This well-established technique has been used by a great variety of authors who also show that numeracy has a large impact on growth.

Keywords

Age-Awareness · Development · Education · Numeracy

Introduction

Education is one of the driving factors for the development and long-term economic growth of countries. Many projects in development aid are set up to increase school enrollment rates or years of schooling to improve education and thus the prospects of future generations. Nowadays, there are plenty of measures and indexes at hand to quantify different levels of education among children, adolescents, and adults. Through various tests and methods, the levels of education or human capital are comparable on an international basis. In the famous Programme for International Student Assessment (PISA), scholars compare cognitive skills of students from various countries around the world. On the one hand, the impact of such a program is enormous: The countries with lower scores invest financial means or restructure their schedules to push forward in the range. On the other hand, the results build one of the largest databases on students' education worldwide with which scholars are able to conduct analyses and draw conclusions for the future.

However, if we go some decades further back in time, we have to rely on other measures of human capital such as years of schooling, enrollment rates, or literacy because we simply lack other indicators. The differentiation between different years of schooling, for example, is slightly less exact than that of the cognitive skills tests of the PISA study. Moreover, there are other issues that might occur with these indicators. If a child is enrolled in school, it does not necessarily mean that he or she acquires a certain level of reading or mathematical skills before potentially dropping out. Literacy rates are often self-reported or even have to be constructed from people's ability to append their signatures to documents, such as marriage registers or wills, which does not necessarily imply that the person is able to read and write. Reis (2005) reports such estimated literacy rates for a number of European countries around 1800. The English database implemented by Schofield (1973) reaches back to the middle of the eighteenth century. By analyzing wills, Gregory Clark (2007) constructed another large database on English literacy that even dates back to 1585.

The construction of databases on literacy reaching back to the sixteenth century is, of course, an exceptional case and only possible for a country such as England where the availability of sources is much better than in most of the other countries in the world. In most countries, data sources are scarce and do not provide literacy or enrollment rates until after the Industrial Revolution. For some less developed countries or world regions, we do not even find comprehensive enrollment rates for the past 50 years because schooling was not obligatory or there were no schools nearby for children to attend. But how can we measure human capital in times in which education was only available for the rich or in regions where data sources are very scarce?

In numerous surveys, church registers, or census lists, people reported information from which scholars are able to derive a basic indicator of human capital: their age. The underlying concept for calculating such an indicator is the so-called age heaping: In earlier times, when people did not have birth certificates or passports, they were often not aware of their true age or they simply did not know it because no one kept record of their exact date of birth. As a consequence, when people were asked for their age and they did not know it, they tended to state a "popular" number. For instance, they claimed to be 35 when they were in reality 34 or 36. Hence, the age distribution shows "heaps" or "spikes" at these popular digits that are mainly multiples of 5. Why does this clearly not reflect the true distribution of ages? We can explore that with a small example: If in the year 1935, for example, 100 people stated to be 35 years old but only 50 people reported being 34 or 36 years of age, this would mean that twice as many children were born in 1900 compared to the years 1901 and 1899. This is a very unlikely scenario and most probably due to age non-awareness. This phenomenon causes problems for demographers because they have difficulties estimating the true distribution of males and females in certain age groups or the life expectancy of a population (see, e.g., A'Hearn et al. 2009). But, while being a disadvantage to the accuracy of demographic research, this pattern is actually a benefit for the research on basic education: By implementing an indicator such as the Whipple, we can calculate the ratio of the individuals who were able to report their own ages exactly in contrast to those who stated rounded numbers. Consequently, an indicator based on age heaping enables us to conduct studies on basic numeracy or human capital for a great variety of countries and in the very long run.

Many authors used the by now well-established age-heaping method on various topics related to basic education: Myers (1954); Mokyr (1983); Zelnik (1961); Duncan-Jones (1990); Budd and Guinnane (1991); Ó Gráda (2006); Manzel et al. (2012); as well as Crayen and Baten (2010a, b), among others, studied differences in numeracy of various countries, world regions, and time periods. A'Hearn et al. (2009) demonstrated the strong relationship between age-heaping-based indicators and literacy. De Moor and Van Zanden (2010), Manzel and Baten (2009), and Friesen et al. (2013) assessed gender inequalities in numeracy in different world regions, whereas Juif and Baten (2013) compared the numeracy levels of Inca Indios before and after the Spanish conquest. Stolz and Baten (2012) analyzed the effects of migration on human capital selectivity – hence, they measured the extent

of "brain drain" or "brain gain" of countries through migration.[1] Charette and Meng (1998), for instance, assessed the impact of literacy and numeracy on labor market outcomes.

In the following section we will explain in greater detail the advantages and potential caveats of the age-heaping method. We also discuss the indicators that are commonly used to approximate basic numeracy, and we describe in which way they are calculated. Furthermore, we explore the relationship between age-heaping-based indicators and other measures such as literacy and schooling. In section "Applied Age-Heaping Indicators in Various Research Topics," we describe different research topics that have been assessed by implementing the age-heaping method, while in section "The Development of Numerical Skills in Different World Regions and Time Periods," we discuss studies that explore differences in numeracy levels across various world regions. In section "Numeracy Trends of Women and the Gender Gap in Different World Regions," we present the development of women's numeracy and the gender gap. Section "Conclusion: The Impact of Numerical Abilities on Growth" provides concluding remarks concerning the impact of basic numeracy.

Age-Heaping-Based Indicators: Advantages, Potential Biases, and Indexes

Advantages, Potential Biases, and Heaping Patterns

The requirement for employing numeracy as an indicator for human capital is that a certain share of people in earlier times – especially before the Industrial Revolution – was not aware of their actual age because they did not know their date of birth or they were not able to calculate the number of years from their date of birth to the actual year.[2] Consequently, when individuals were asked for their age and could not state it exactly, they did not choose any number randomly, but they typically tended to report a number divisible by 5 such as 35, 40, 45, and so on (Duncan-Jones 1990; A'Hearn et al. 2009).

[1]Brain drain means that highly educated people emigrate from their country of origin to another. Brain gain means the opposite effect.

[2]However, we have to keep in mind that there are individuals still living today, predominantly in the least developed countries, which are not aware of their true age when they are asked for it (Juif and Baten 2013).

While the aforementioned is the most commonly detected heaping pattern, there is also some heaping on multiples of 2 – hence even numbers.[3] In the Chinese culture, one might also think of a different heaping pattern, for example, the avoidance of the number 4, which when pronounced sounds similar to the word for "death," or the preference of the number 8, which can be associated with fortune (Crayen and Baten 2010a). However, Baten et al. (2010) found that Chinese migrants to the United States (US) heaped considerably more on multiples of 5 than on the birth year of the dragon, for instance, which is a very popular animal sign in China.

One great advantage of an age-heaping-based indicator is that it enables us to assess basic numeracy for a large number of countries over a very long period of time because this phenomenon presumably appeared in most societies until a certain point in time (Duncan-Jones 1990). The second advantage is that there exist a large number of sources that can be employed to calculate numeracy indexes. In principle, we can use any list for which people had to report their age including census lists, ecclesiastical surveys, tax lists, marriage registers, death registers, and shipping lists, just to name a few. Of course, selection biases need to be studied. One very early census in the history of mankind that we are aware of is the population census decreed by Emperor Augustus – around the birth of Christ – for which Maria and Joseph were heading to their place of birth to be enumerated. Duncan-Jones (1990, p. 79), however, reveals another way to measure age awareness in ancient times: the inscriptions on tombstones in the Roman world. Age heaping on multiples of 5 was very common in the first centuries after Christ, with levels of age misreporting of up to 60%.

[3] De Moor and Van Zanden (2010) even report a preference for multiples of 12 in different medieval and early modern sources, among them a census from Tuscany in 1427 and another from Reims in 1422. This phenomenon could be the result of religious orientations and the underlying usage of the number 12 as a holy number. Interestingly, this heaping pattern was more often adopted by women than by men, especially during early modern times in the South Netherlands. This could be due to a stricter adherence of religious practices or beliefs by women than by men, though this is not scientifically proven so far.

Another pattern might also occur if a certain share of the population was surveyed and the results were written down in year t, whereas the rest of the data collection was performed in the following year t + 1. After the census was finished, the census official compiled the results in a clean and comprehensive list in year t + 1. Because he or she was aware of the age statements that had been reported in year t, he added 1 year to those ages. As a result, we find heaping on the terminal digits one and six in these lists. If this pattern can be identified without reasonable doubt, the additional year should be subtracted from all of the affected age statements.

In a similar way, the authors of some studies have found that numeracy estimates based on age statements of marriage lists tend to be upwardly biased (which is partly due to the fact that marriage was restricted to those who earned a living and could nourish a family in many historical societies). Death registers on the other hand tend to yield downwardly biased estimates. This type of bias could happen if the deceased person did not have any relatives or close friends whom the recorder could ask for an age statement. Consequently, he or she estimated the age by himself. Adjustment factors for these types of sources are available from the authors.

The most important factor when calculating age-heaping levels derived from the aforementioned lists is that the ages of the individuals are self-reported and not counterchecked.[4] In some cases, particularly church survey data, such as marriage registers, it is possible that an ambitious priest counterchecked the ages of the bride and groom by their respective birth dates in a birth or baptism register. In the case that ages are counterchecked, we usually cannot detect any age heaping at all. Hence, if numeracy levels are extremely high, particularly in the case of very early samples of rural parishes, we should either eliminate the sample from the dataset or check the possibility of high numeracy levels. We could, for example, compare the numeracy levels to the corresponding literacy rates of the parish or to the numeracy levels of regions or villages with a similar infrastructure, education system, and so on (A'Hearn et al 2009). Generally, we can say that the further back in time the period of interest lies and the higher the age heaping is, the more likely it is that ages are not counterchecked. In censuses executed by governmental authorities and in times in which obligatory identification did not exist, we can assume that ages are not counterchecked.

Another possible objection could be the question: Whose age heaping do we measure after all? Do the statements truly reflect the pattern of the respondents or is the observed age heaping actually caused by the census taker? Critics could argue that the census taker might have estimated the ages of the people by himself or herself or corrected those that seemed implausible to him or her. This potential issue has to be examined carefully for each data source. However, there are various hints that this is not the case in the studies under discussion. According to Manzel and Baten (2009), some of the executive authorities explicitly required the census takers to interrogate the people individually.[5] Moreover, if the age-heaping results were influenced by the individual numeracy level of the census taker, the results of different censuses should vary within one region or country for the same birth cohorts. The authors, however, find that the results of different censuses display very similar levels of age heaping for the respective birth decades.

Another strong argument in favor of the self-reporting of surveyed individuals is the difference in numeracy levels that we find between occupational and social groups. Baten and Mumme (2010) as well as Tollnek and Baten (2013) reveal that better educated groups of professionals, such as merchants, show significantly higher levels of basic numeracy than unskilled or partly skilled individuals. Furthermore, A'Hearn et al. (2009) show that the correlation between literacy and numeracy rates is very strong on a regional or countrywide basis. Clearly, we are only able to

[4]Self-reporting is, of course, not an option if we consider tombstones or death registers. The ages provided in these sources reflect the heaping pattern of the individual who reported the age in place of the respective person. But even in such cases, there are gender- or social group-specific differences observable (Duncan-Jones 1990, p. 83). It is most likely that the persons providing the ages for the tombstones were related to the deceased person or at least of similar social or educational status.

[5]They found information on censuses from which it becomes clear that the authorities required the census takers of surveying each person individually.

detect such considerable region- or occupation-specific differences if people stated their ages by themselves.

Related to information about households or married couples, there is a further possible question to discuss: Did women report their ages themselves or did their husband help them – or even answer for them? How reliable are comparisons between male and female numeracy originating from the same source? In various studies, scholars suggest that we can rely on the age statements made by or assigned to women: According to De Moor and Van Zanden (2010), the indexes of women and men in a Belgian census, for example, were actually not that different. Hence, it seems plausible that the individuals responded by themselves. Furthermore, they find that women sometimes displayed preferences for different numbers than men – such as multiples of the number 12 – which can only occur if the women stated their ages by themselves.

Manzel et al. (2012) also find evidence in favor of the self-reporting of household members, which is based on results from the 1744 census of Buenos Aires: If it was the case that the head of household stated the ages in place of the other family members, there should be substantial differences in the numeracy levels, because one might assume that the heads were better educated than the other members, given that he or she provided the family income and in most of the cases had an occupation. However, the difference is almost negligible. Moreover, the authors report sources in which the interviewer made complementary remarks. Related to a certain person who reported to be 30 years old, he noted, "[. . .] but looked considerably older" (Manzel et al. 2012, p. 940). Such statements strengthen the assumption that census takers asked the people individually for their ages and did not accept someone else answering in their place. With all the results of the aforementioned studies and the information provided on the procedure of various censuses, we can assume that the studies discussed in this paper deliver reliable information on the basic education of the respective population.

Whipple, ABCC, and Other Indexes

There are various indexes we can adopt for measuring age heaping. In some cases the employed scheme varies from one study to another, depending on the author. What many of the indexes have in common, though, is the assumption that ages, stated as integers, follow a discrete uniform distribution. For example, 10% of the people in the 10-year age group from 30 to 39 are expected to report their age as 31, i.e., with "1" as the terminal digit since it is the only number ending with this digit in this 10-number interval. Applied to heaping on multiples of 5, this implies that 1/5 (two out of ten) or 20% of the ages in this age group end in the digit "0" or "5." Ó Gráda (2006), for example, implements a simple index by observing the frequency of the numbers divisible by 10 in the age groups 30–34, 40–44, etc. Observing five ages in each group should, in the simplest case, deliver the same frequency for each digit. A value greater than 0.2 (which equals 1/5) indicates a rounding pattern of the respondents. As a consequence, we expect each age to be

reported by about the same number of individuals. However, we have to be careful concerning the assumptions of age distributions in general. Especially in older age groups, it is most likely that a higher share of people is alive at age 60 in contrast to those aged 69 (Crayen and Baten 2010a, p. 84).

When it comes to measuring the actual degree of age heaping, there are some desired properties that can improve the results of the indicator, as described by A'Hearn et al. (2009). First, the index should be scale independent, which means that it delivers comparable results for two samples with the same heaping patterns but different sample sizes. The second valuable feature is the linear response to the degree of heaping, which implies that the indicator increases linearly when heaping rises. Finally, the coefficient of variation should be as small as possible across different random samples.[6]

There are several established measures with at least some of the desired properties such as the indexes suggested by Mokyr (1983); Bachi (1951); Myers (1954).[7] A'Hearn et al. (2006) state that the indicators proposed by Mokyr and Bachi are not calculated on the basis of specific expected frequencies. Hence, they do not rely on a particular assumption about which terminal digit appears with a certain frequency. However, there is a common procedure also discussed by Myers (1954) that implies the expected proportion of each terminal digit to be 10%. For this procedure it is necessary to sum up all of the ages ending in zero, then those ending in one, and so on, starting at age 20, for example. In the next step, the share of the population stating the respective terminal digit (zero to nine) relative to the whole population is calculated.[8] Consequently, each percentage share greater than 10% means an overrepresentation of the ages with the respective digit. The "blended" index proposed by Myers (1954) works in a similar way as this procedure but with some adjustments: Instead of starting the aggregation at age 20, he uses the terminal digits at each age between 23 and 32, for example, as the starting point. He then proceeds with the aggregation of the ages with each terminal digit (zero to nine), but instead of counting each unit digit once, it is counted several times, according to the "leading" digit.[9] The result of this procedure represents the relative share of the people that reported ages with the respective last digit. If there is no age heaping in the data, the percentage share of each figure should not differ largely from 10% (Myers 1954, p. 827).[10]

While the Bachi and Myers indexes are scale independent at least in the mathematical sense, none of the indexes turns out to be scale independent in the statistical

[6]Please see A'Hearn et al. (2006, pp. 11–21) for a more detailed discussion on the properties.

[7]The Mokyr index we refer to in this section is also called the Lambda index (A'Hearn et al. 2006).

[8]The digit "0" includes all ages ending in zero, hence 30, 40, 50, etc. The digit "1" includes all ages ending in one, hence 31, 41, 51, and so on.

[9]Myers criticizes that starting the aggregation at a certain age, for example, 20, increases the share of people with a digit ending in zero because ". . . the 'leading' digits naturally occur more frequently among the persons counted than the 'following' ones." (Myers 1954, p. 826).

[10]For a more detailed description of the "blended" method, see Myers (1954).

sense, meaning that the mathematical scale independency does not hold in random sample settings, as A'Hearn et al. (2006) show.[11] Each of the three indexes discussed in this section can be adopted to reveal any kind of heaping, be it rounding on multiples of 5 or the preference for any other of the 10 digits. This might be a small advantage in contrast to indicators that can only detect a preference for multiples of 5. However, there is an indicator that exceeds all of the others in terms of its properties: the Whipple index. The Whipple is statistically scale independent, its expected value rises linearly with the degree of heaping, and its coefficient of variation is lower than that for the other indicators discussed (A'Hearn et al. 2009). The Whipple index is calculated as presented in the following formula (1):

$$Wh = \frac{\sum (n_{25} + n_{30} \ldots + n_{65} + n_{70})}{\frac{1}{5}\sum_{i=23}^{72} n_i} \times 100 \tag{1}$$

In the enumerator, the number of people reporting ages ending in zero or five is aggregated. This is divided by all of the reported ages in the age range 23–72. Subsequently, we multiply the sum of the reported ages by 1/5 in the denominator. This is based on the assumption that 20% of all the people correctly report an age ending with zero or five. The whole term is then multiplied by 100 for convenience. Hence, the Whipple can take on values usually ranging between 100 and 500. If exactly 1/5 of all the individuals state an age ending in a multiple of 5, the Whipple takes on the value 100. In the case that all of the people report a multiple of 5, the Whipple increases to 500. However, we have to be careful when interpreting this figure: A value of 500 would still mean that 1/5 of the individuals who state a rounded age were doing so correctly. Admittedly, with an age-heaping effect of this size, we might as well assume that these individuals did not report their correct age because of age awareness. In theory, the Whipple can also take on the value zero, if no person reports a multiple of 5 – this would be the case of perfect "anti-heaping" (A'Hearn et al. 2009). The Whipple increases linearly, which means that it rises by 50% whenever the proportion of people reporting a multiple of 5 increases by 50% (Crayen and Baten 2010a, p. 84)

Because of its design, the Whipple index obviously does not account for the fact that fewer people are alive at higher ages. Thus, there is naturally a higher number of people reporting the age of 60 than the age of 69, even if there was no age heaping in the population otherwise. We are able to reduce this potential bias by calculating the Whipple for age groups of 10-year steps. Additionally, we arrange the age groups such that the multiples of 5, and especially the numbers ending with zero, are more evenly distributed within the age groups: The first age group starts at age 23 and ends with age 32. The other age groups are arranged accordingly: 33–42, 53–62, and so

[11] Statistical scale dependency means that the assumed mathematical scale independency can change when applying an indicator to random samples of different sizes. For more information on this topic, see A'Hearn et al. (2006, pp. 11–21).

on. It is more reliable to exclude individuals older than 72 years because they tend to exaggerate their age. In principle, the survivor bias effect could also play a role because people with a higher basic education might have a higher life expectancy due to a higher expected income, for example. However, Crayen and Baten (2010a) showed that it did not have an empirical impact.

It is also common to exclude the individuals younger than 23 years of age from the analysis for two reasons: First, young people often married around the age of 20 or entered military service at that time. As they often had to report their ages at such occasions, their age awareness is expected to be better than that of older individuals. Second, younger people tended to round their ages to a much greater degree on multiples of 2 than of 5. Additionally, for children still living with their parents, we do not know if they reported their ages themselves or if their parents answered for them (Manzel and Baten 2009). To account for a higher degree of heaping on multiples of 2 among this group, which is not captured directly by the Whipple, Crayen and Baten (2010a, Appendix A) propose an upward adjustment of the Whipple index. With this adjustment, the value of the youngest age group increases, and hence, the estimated numeracy decreases.[12]

The Whipple index combines a number of desired properties and is – after making some adjustments – a reliable measure for the degree of age heaping. However, the adopted scale and the interpretation of its outcomes are not particularly intuitive. A'Hearn et al. (2009) solved this issue by introducing another indicator which they called the "ABCC."[13] The calculation works as shown in the following formula (2):

$$ABCC = \left(1 - \frac{(Wh - 100)}{400}\right) \times 100 \text{ if } Wh \geq 100; \text{ else } ABCC = 100 \qquad (2)$$

The ABCC is a simple linear transformation of the Whipple and ranges between 0 and 100. For the case of "perfect" heaping and thus a Whipple of 500, the ABCC takes on the value 0. If every person states their age correctly, the ABCC value increases to 100. Hence, the ABCC can intuitively be interpreted as the share of people reporting their age correctly. This measure has been successfully used in a variety of studies so far (Manzel and Baten 2009; Baten and Mumme 2010; Manzel et al. 2012; Stolz and Baten 2012; Juif and Baten 2013 as well as Baten and Juif 2013).

[12]If the Whipple indicator is larger than 100, they suggest adding 0.2 units to the value of the age group 33–42 for every Whipple unit above 100. The resulting value is aggregated to the value of the age group 23–32, which delivers the new estimate for this group. For example, if the value of the age group 23–32 is 150 and that of the age group 33–42 is 160, then the digit above 100 has to be multiplied by 0.2 (60 * 0.2 = 12). The result is added to the original value of those aged 23–32 (150 + 12). Consequently, the new estimate for the youngest age group is 162 (Crayen and Baten 2010a, Appendix A, pp. 95–96).

[13]The name of the index is constructed by the initials of the last names of the three authors plus Gregory Clark's.

Because age-heaping indicators such as the Whipple and the ABCC index are employed to approximate basic education if other indicators are not available, it is very important that these indexes correlate with other measures. It turns out that there is a strong correlation between the share of people reporting their correct age and indicators such as literacy or schooling. Myers (1954) finds a correlation of high literacy rates and low levels of age misreporting for Australia, Canada, and Great Britain. Duncan-Jones (1990) also reports a significant correlation between age heaping and illiteracy in a number of developing countries in the twentieth century, among them Egypt (1947), Morocco (1960), and Mexico (1970). Furthermore, A'Hearn et al. (2006, p. 21) perform analyses on the relationship between age heaping and illiteracy in various countries. They detect a very strong, significant, and robust correlation between the two indicators for almost all of the 52 countries in their dataset. In the very detailed analysis for the United States, the correlation is particularly strong, even when controlling for birthplace, ethnic group, and gender balance; and it is evident for both pooled and regional fixed effects regressions (A'Hearn et al. 2009).

Moreover, Crayen and Baten (2010a) tested the impact of several factors such as primary schooling, height, and state antiquity on age heaping.[14] For a global dataset, they found that school enrollment is one of the driving factors for the development of numerical abilities among societies. In all of the modifications and independent of the factors controlled for, it is always highly and significantly correlated with age heaping. Consequently, we assume that age-heaping-based indicators are valid estimators for basic education.

Applied Age-Heaping Indicators in Various Research Topics

Reconstructing Very Early Numeracy Differences: The Example of Inca Indios

Acemoglu et al. (2001, 2002) studied the differences between former European colonies. They compare the former colonies that are rich today to those that are poor. Acemoglu et al. argue that the Europeans created exploitative institutions in the colonies that had an adverse disease environment for Europeans. In contrast, they implemented growth-promoting types of institutions in those colonies in which Europeans settled. Examples for the latter would be the United States, Australia, Argentina, and South Africa in part, whereas a classical example for the former would be West Africa. The more or less growth-promoting nature of colonial institutions translated into better or worse institutions during the late twentieth century. This had an impact on today's difference in real income per capita because

[14]Height is employed as a proxy indicator for infant malnutrition because the smaller a person is, the more likely it is that he or she did not have access to protein-rich nutrition which also hinders the development of numerical skills. State antiquity approximates the quality of institutions.

institutions tend to remain similar for a longer period. Applying the age-heaping technique to this topic is particularly useful because alternative views suggest a strong role of human capital channels (Glaeser et al. 2004). A related question is, for example, whether there were "precolonial legacies": How much did the ancient economies and societies invest before the colonialists arrived?

A paper by Juif and Baten (2013) employs an early Spanish census that was taken directly after the invasion of the Incan Empire. It makes use of the fact that basic numeracy is usually attained during the first decade of life. Clearly, the question needs to be considered whether such a birth cohort-specific analysis could be distorted by later learning processes. However, the numeracy values of the cohort born before the invasion are close to zero and thus cannot be upwardly biased. The numeracy levels of the cohorts born after the invasion, in contrast, were slowly rising. Consequently, the most important result of this study was that in fact some precolonial legacy – or burden – existed in Andean America. This legacy has not been reduced during colonial times, as colonial institutions such as the Peruvian "Mita" reinforced educational inequality (Dell 2010). During the early period, it is interesting that some Indio groups that were allied with the Spanish during the invasion (and received tax exemptions and a slightly less terrible standard of living after the invasion in return) also displayed a better numeracy. A likely interpretation is that their slightly higher net income allowed more investments in the basic numeracy of their children. This observation also stands in contrast to the suspicion that cultural attitudes could have implied a different number rounding behavior. Another problem considered by the authors is whether colonial officials did not ask the Indios for their age, but tended to estimate it without asking (if they estimated after asking, this would not be a problem for the age-heaping procedure because in this case the respondent did most likely not know his or her age either). Juif and Baten rejected these doubts in their study with arguments based on the effect that the social difference of numeracy within the Indio groups was substantial. In addition, the colonial officials sometimes explicitly noted thoughts about the appearance of a person if the self-reported age and the official's impression differed. This clearly indicates that the Indios were in fact asked for their age. As a result, this earliest numeracy study for a non-European country revealed that a negative precolonial legacy was in fact very likely.

Religion and Numeracy

A number of scholars have recently studied potential religious determinants of human capital formation (see Becker and Woessmann 2009 for a widely cited study and a good overview). The relative exogenous character of religious rules has been stressed by this literature because beliefs about the necessity to read religious texts are considered to be less influenced by economic factors and profit-maximizing educational investment decisions. Botticini and Eckstein (2007) explained how religious rules for the provision of education of one's (male) offspring appeared in the Jewish faith. In the first century BCE, a conflict between two

influential religious factions of Judaism took place. One of these factions, the Pharisees, stressed the religious duty to educate, and they gained stronger influence on Judaism than the other group. Botticini and Eckstein emphasize that the education rule was not economically motivated because the large majority of the Jewish were farmers and rural day laborers, for whom a substantial educational investment would not yield sufficient returns during this period. Only with the substantial urban growth in Mesopotamia during the eighth and ninth centuries CE could the Jewish population living there use their religiously determined education to achieve profitable positions as merchants and, later on, as bankers. Medieval Western Europe actually first tried to attract this religious and occupational group because the kings of England and France assumed correctly that government revenues might increase. The famous restriction of Jewish population groups to being exclusively merchants, bankers, and other traders - occupations that were forbidden to the Christian population - was only created later, during the High Middle Ages. Botticini and Eckstein (2007) therefore reject the hypothesis that this restriction caused high Jewish educational levels.

The debate over religious differences of education and numeracy in particular has important implications for history and for our understanding of human capital formation. For that reason, Juif and Baten (2014) studied the differences between the average population and the persons who were accused by the inquisition of practicing Jewish beliefs in Iberia and Latin America. The period under study runs from the fifteenth to the eighteenth centuries. The sources that are available for this early period were primarily created by the inquisition. A question about the age of the accused was included for identification purposes. Besides the evidence from the inquisition lists, we also included census-based numeracy evidence to compare the average population in the same regional units. We studied potential selectivities and biases intensively and dismissed them ultimately. The most important result of this study of religion and numeracy is that persons who were accused of being Jewish had a substantially higher numeracy than the average population. If we accept the working hypothesis that most of the persons accused of Judaism came from families of a different educational behavior (and a different educational self-selection), the religious factor appears to be of important influence. However, the authors also find that the catholic elites (such as priests) had a substantially higher numeracy compared to the average Iberian and Latin American population.

Path Dependency of Early Numeracy and Land Inequality as Determinants of Modern Math and Science Skills?

Within the framework of Unified Growth Theory, Galor et al. (2009) have focused on land inequality as one of the crucial obstacles to human capital formation. They describe the political economy of regions and countries with higher and lower land inequality, assuming an influential role of two different elite groups: large landowners and industrial capitalists. In regions with lower land inequality, industrialists wielded larger relative power in the decision-making process concerning educational

investments. In contrast, in regions with high land inequality, large landowners remained in power and were not particularly interested in spending their taxed income for primary schooling: First of all, their agricultural day laborers did not have to be educated to fulfill their manual tasks (at least that is the traditional view). Secondly, additional primary schooling would have increased their burden of taxation. Thirdly, educated workers might have moved to cities or may even have initiated land reforms. In a study of this land inequality effect on modern math and science skills, Baten and Juif (2013) also include early numeracy (around 1820) as the second main determinant. They find that early numeracy has a large explanatory share, even after controlling for land inequality and a number of other factors. It seems that this path dependency worked via economic specialization: If an economy specialized early on the production of human capital-intensive products, the relatively high income allowed investing in education for the next generation. In addition, such human capital-intensive production methods probably resulted in substantial switching costs – hence, the countries specialized in this type of production and developed a branding and reputation for their products. As a consequence, they were most likely entering a high degree of path dependency.

Numeracy Differences Between Occupational Groups in Preindustrial Times

When it comes to the question of who stated the ages written down on a census list – the enumerator or the respondent – the analysis of numeracy between occupational groups is crucial. If the age-heaping levels between occupational groups vary significantly, this might indicate that the respondents stated their age themselves. De Moor and Van Zanden (2010, p. 204) were able to verify differences in numeracy between three occupational groups for the seventeenth century in Amsterdam: professionals, craftsmen, and unskilled laborers. While the highly skilled professionals had relatively low age-heaping levels (with an ABCC index of 100), the opposite was the case for the non-skilled individuals, who displayed a high degree of age heaping.[15] The craftsmen had slightly better values than the unskilled group.

Tollnek and Baten (2013) assess the numeracy of occupational groups for four countries in early modern Europe (Austria, Spain, Southern Italy, and Germany) as well as for Uruguay. Additional information is provided by literacy data from Switzerland. In total, the comprehensive dataset includes nearly 30,000 observations with information on age, sex, and occupation of individuals. The authors distinguish between six occupational groups, adapting the Armstrong scheme (Armstrong 1972): the professionals (doctors, lawyers, etc.), the intermediate (administrators and higher clerks), the skilled group (craftsmen and shopkeepers), the partly skilled (herdsmen and carriage drivers), the unskilled group (day laborers), and the farmers

[15]De Moor and Van Zanden (2010) use the Whipple index for their calculations. We translated the numbers into ABCC values for convenience.

(smallholders and farmers with medium-sized or larger farms).[16] The descriptive analysis already reveals large differences between the groups. In all of the European countries, the professionals have the highest numeracy values (ABCC index between 86 and 96), followed by the intermediate and skilled groups that reflect lower numerical abilities (Tollnek and Baten 2013, p. 33). The two lowest groups of society, the partly skilled and the unskilled groups, have the lowest values of numeracy in all of the countries.[17] Interestingly, the farmers have low age-heaping levels, with numeracy values similar to the skilled group. In Germany and Uruguay, the farmers' values are even close to the groups with the highest levels, which are the professionals in Germany and the skilled in Uruguay.

The authors also assess these differences in a logistic regression with "numerate" as the dependent variable that assumes the value of one if the individual stated an exact age and zero otherwise.[18] They control for the birth half century, the country, the age (because younger people might know their age more exactly), and, most importantly, the occupational groups. The regression results strongly confirm the descriptive results for all of the countries in the sample. The three upper groups and the farmers have a significantly higher probability of being numerate than the partly skilled and unskilled groups (Tollnek and Baten 2013, p. 28). The values of the coefficients range between roughly 18 for the professional groups and 8 for the skilled. The farmers have the third highest chance for success (hence, "numerate" takes on the value one) with a coefficient of nearly 9, which can be translated into a higher probability of being numerate of about 9% in contrast to the two lowest groups. These results are also confirmed by regression results using literacy evidence from Switzerland.

The Development of Numerical Skills in Different World Regions and Time Periods

A Human Capital Revolution in Europe

A'Hearn et al. (2009) discuss the development of numeracy all over Europe from the late middle ages to the early modern period. The European countries experienced a striking increase in numeracy during this time period, which can be identified as a "human capital revolution." While the numeracy values rose in all of the European countries, there was variation between the different parts of Europe. The Western European countries showed an exceptional development. As early as around 1450,

[16]The occupations in brackets are only examples. In total, there are hundreds of occupations in the dataset that were arranged according to the Armstrong scheme.

[17]Germany is an exceptional case because the values of the intermediate, skilled, partly skilled, and unskilled groups differ only slightly.

[18]The coefficients are subsequently multiplied by 125 to correct for the 20% of the people who state a multiple of 5 correctly. For further information, please see Appendix B in Tollnek and Baten (2013).

the Netherlands represented numeracy values (approximated by the ABCC index) of roughly 70% (A'Hearn et al. 2009, pp. 801/804).[19] Britain and France surpassed this value at around 1600 and 1650, respectively. Britain and Denmark, on the other hand, already experienced numeracy rates of 90% or more in the period of 1700. While Denmark's rates grew continuously until the end of the period at around 1800, Britain's values remained at the same level.

The picture looks similar if we look at Central Europe (A'Hearn et al. 2009, pp. 801/804). Austria and Protestant Germany already had high numeracy levels of between 78% and 87% around the period of 1600. Catholic Germany had lower values (68% in circa 1700), but it converged strongly thereafter. The Eastern European countries, in contrast to the rest of Europe, lagged slightly behind: Around 1600, Bohemia represented numeracy values of only 44%. One period later, around 1650, Russia and Hungary showed levels of 43% and 32%, respectively. However, toward the end of the early modern era at approximately 1800, the overwhelming majority of the European countries managed to increase their human capital values significantly. Even the regions that lagged behind, such as Bohemia and Russia, reached numeracy levels well above 80% or close to 90%.

Numeracy Levels in Latin America

Manzel et al. (2012) analyze long-term trends in numeracy for a number of Latin American countries from the seventeenth to the beginning of the twentieth century. Some of the countries, such as Argentina and Uruguay, experienced strong increases of human capital throughout the whole time period that are comparable to those of some European countries. While Argentina started with an ABCC value of less than 20% in the birth decade 1680, it reached values of almost 70% around 1800 (Manzel et al. 2012, p. 954).[20] With an exceptional increase during the nineteenth century, Argentina reached almost full numeracy at the beginning of the twentieth century. The development of Uruguay is similar, showing even higher numeracy levels than Argentina in parts of the nineteenth century. Despite such great examples of convergence, some of the Latin American countries underwent a process of divergence during the nineteenth century: In Colombia, Mexico, and Ecuador, the ABCC levels stagnated. While Mexico started off well with continuously growing numeracy levels from 1680 to 1790, there was almost no improvement throughout the

[19]The data are arranged in age groups and then transferred into birth half centuries. Hence, the value of the respective age group is subtracted from the census year. The resulting values are rounded to 50-year-intervals. For example, if the census year was 1740, then the age group 23–32 was born in the half century 1700.

[20]The values for Argentina and Mexico are estimates based on regression results. They are controlled for capital effects and male share. For further information, please see Manzel et al. (2012). The data of all of the countries are arranged in birth decades. Hence, the value of the age group is subtracted from the census year, and the resulting values are rounded to 10-year intervals. For example, if the census year was 1940, then the age group 23–32 was born in the decade 1910.

nineteenth century. Ecuador's levels even worsened slightly during the nineteenth century. Brazil was a particular case because it began with increasing levels of numeracy during the eighteenth century, then experienced a short period of stagnation at the first half of the nineteenth century and managed to increase human capital again in the following decades. Toward the beginning of the twentieth century, numeracy levels rose considerably in all of the observed countries.

Industrialized Countries Versus the Rest of the World?

Crayen and Baten (2010a) assess long-term trends of numeracy in 165 countries all over the world. The development of some industrialized countries not discussed so far is of interest: The United States started with ABCC values below 87% at the beginning of the nineteenth century, which are among the lowest numbers compared to the other industrialized countries in the same period (Crayen and Baten 2010a, p. 85).[21] Toward the middle of the nineteenth century, the values of the country increased significantly to around 94%. The United States converged continuously in the following decades and reached values of circa 98% at the end of the nineteenth century. Spain had values of about 88% around 1830. The increase of Spain's numeracy developed more slowly than that of the United States, but it also reached levels close to 100% at the beginning of the twentieth century. Exceptional cases are also Greece and Cyprus, which had values below 75% and 78%, respectively, at the end of the nineteenth century. However, their rates increased dramatically throughout the twentieth century. Ireland is one of the few industrialized countries in which the ABCC index decreased slightly in the 1870s, which is likely due to the behavior after the Great Famine that took place two decades earlier.

The comparison of world regional numeracy trends reveals some crucial differences. South Asian countries had the highest age-heaping levels with ABCC values of less than 13% around 1840 (Crayen and Baten 2010a, p. 87). The numbers increased steadily throughout the following decades, reaching an ABCC index of above 55% toward the 1940s. The Middle East and North Africa had the second lowest levels of numeracy with values lower than 25% in the 1820s. Egypt most likely had the highest age-heaping level in this region with an ABCC of almost 0 (the case of "perfect" heaping) (Crayen and Baten 2010a, p. 86). But similar to South Asia, the Middle Eastern and North African countries managed to increase their numeracy levels continuously (Crayen and Baten 2010a, p. 87). The industrialized countries were on the upper range of the strata with the highest numeracy levels. East Asia still had ABCC levels of below 88% at the beginning and toward the middle of the nineteenth century.[22] In only a few decades, though, age heaping in China

[21]Crayen and Baten (2010a) use the Whipple index for all of their calculations. We translated all of the numbers into ABCC values for convenience.

[22]East Asia is dominated by Chinese data, since Japan is considered part of the industrialized countries.

decreased strongly and vanished around 1880. Southeast Asia and Latin America ranged between the regions with fairly high and relatively low levels of age heaping.

Numeracy Trends of Women and the Gender Gap in Different World Regions

Numeracy Trends of Women in Some Industrialized Countries

Gender equality in education and wages is a controversial topic. Even in countries with relatively high levels of income and education, such as the European countries or the United States, there is an ongoing debate about wage differentials between men and women. Women with the same degree of education and experience often receive considerably lower wages than their male counterparts working in the same field or position.

But what about educational differences between men and women before formal schooling became accessible for most people? When did the gender gap open and did it worsen or improve over time? Duncan-Jones's (1990, p. 86) analysis of inscriptions on tombstones reveals a numeracy difference between men and women in Roman times that is most likely the earliest measurable gender gap. Although the age reported on the tombstone supposedly reflects the numerical abilities of a relative, the ages of women show a higher heaping pattern than those of men. The indicator implemented by Duncan-Jones represents the percentage share of people who report a rounded age, relative to those who state their age correctly.[23] While in some regions, such as Moesia or Pannonia, the women had considerably higher heaping levels than the men (28.1% and 17.1%), the differences were relatively small in most of the other regions: In Mauretania, for instance, the women's index was only 4.8% higher than the men's index. In Rome, the difference was 6.8%. However, there were also regions in which women had lower heaping values, such as Italy outside Rome (−1.9%).

De Moor and Van Zanden (2010) assess human capital levels in the medieval and early modern Low Countries. The results of the numeracy levels of Bruges in Belgium (1474–1524) suggest that the differences between women and men were relatively small in total: The men have an ABCC index of about 85% and the women 83% (De Moor and Van Zanden 2010, p. 194).[24] In the city of Bruges, the women even surpassed the men slightly.[25] The authors also found similar results for Holland

[23]He subtracts the 20% of the people who report a multiple of 5 correctly from the total number of people who state a rounded age. Hence, the reported percentage share contains those who incorrectly state a rounded age.

[24]De Moor and Van Zanden (2010) use the Whipple index. We translate the results from the Whipple index into ABCC levels for convenience.

[25]The women, however, represent higher values at the "dozen index" that detects rounding behavior on multiples of 12. This is likely due to religious practices among Catholics (De Moor and Van Zanden 2010).

during the sixteenth to eighteenth centuries. The gender gaps were small then and women sometimes had higher numeracy levels than men.

For the United States, Myers (1954 p. 830) reports that women showed significantly higher levels of age heaping than men in the 1950s. For the other countries included in his study – Australia, Canada, and Great Britain – he detects only very slight differences in age misreporting between women and men in the late 1940s or early 1950s. In Great Britain, women reported their ages even more precisely than men in that time period.

The Gender Gap in Latin America

The previous examples suggest that in particular regions and time periods, women's access to basic education was not as limited as one might have expected. However, we have to keep in mind that the Low Countries, for example, are different from many other countries with respect to the position of women in the society. Men and women already seemed to have had a relatively equal standing in the household in early modern times (De Moor and Van Zanden 2010). But what about the basic education of women in the rest of the world?

Manzel and Baten (2009) assess the development of women's basic education for a large number of countries in Latin America via age-heaping-based indicators. They perform their analyses following a fundamental theory about labor force participation developed by Goldin (1995). Goldin argues that female labor force participation follows a U-shaped pattern over time. In societies with low income and low levels of education, women engaged to a large extent in home production of agricultural goods and work on family farms. At this stage of the process, labor force participation shares are high for both men and women. With increasing levels of income and market integration, more women are tied to household activities and childcare, while men work in factories, for example, where new production techniques overcome the traditional home production. Hence, women's level of labor market participation decreases. One possible reason for that development could be that women's work in factories is socially stigmatized. The third stage of the process is observable in countries that have reached a high level of income and education. Women are able to achieve higher degrees of education and enter white-collar occupations that are less stigmatized than manufacturing work. In this last phase of the U-shape, women participate actively in the labor force again.

Manzel and Baten (2009) were able to confirm this pattern based on numeracy estimates for 28 countries in Latin America and the Caribbean from 1880 to 1949.[26] Instead of testing the relative labor force participation of women, they implement "the relationship between average education and the ratio between female and male education" as an indicator to demonstrate the U-shaped development. As a general

[26]The data are arranged in birth decades.

measure of educational equality between men and women, they subtract the Whipple index of men from that of women and divide the result by the Whipple of men. This is subsequently multiplied by −100 for convenience. If the outcome is positive, the women have a numeracy advantage over the men (and the other way round, if the index is negative). The positive index is defined as "gender equality" in basic education. It turns out that the equality index is negative for most of the countries. However, for some of the countries with high levels of basic numeracy throughout the time period, the equality is relatively high as well, indicating the last stage of the U-shaped hypothesis. This is the case for Argentina, Uruguay, Guyana, and Suriname, meaning that gender equality increases if basic education is well established in the society in general (Manzel and Baten 2009, p. 50/51). The ABCC values for Argentina, to state an example, reach from about 95% to 100% and the equality index is slightly above zero (Manzel and Baten 2009, p. 51 and Appendix p. 69). In Guatemala and the Dominican Republic, for example, the authors find the opposite effect, namely, low levels of basic numeracy and low equality indexes. Colombia, however, has ABCC levels between roughly 80% and 90%, while the equality index ranges between approximately −26 and −10, meaning that women have large educational disadvantages in Colombia at the beginning of the period, which decrease over time (Manzel and Baten 2009, p. 50/51 and Appendix pp. 69–71). But there are also cases such as Haiti where numeracy is low, whereas gender inequality is not observable, indicating the first stage of the U-shaped hypothesis. In general, the non-Hispanic parts of the Caribbean represent considerably higher equality indexes as well as higher ABCC levels than the Latin American countries during the whole time period.[27] Toward the end of the period, equality rises with increasing levels of basic numeracy in all of the countries. In Latin America, the ABCC values increase from roughly 78% in 1880 to about 93% in the 1940s and in the non-Hispanic Caribbean from about 90% to 99% (Manzel and Baten 2009, p. 52). The equality values increase from less than −12 to about −5 in Latin America and from roughly −3 to slightly above zero in the non-Hispanic Caribbean (Manzel and Baten 2009, p. 55). The ABCC and equality values of the Hispanic Caribbean are mainly lower compared to the values of Latin America.

To test the U-shaped hypothesis, Manzel and Baten perform a regression analysis with the equality index as the dependent variable, controlling for a number of other factors such as female voting rights and a democracy index. The most important factors for the U-shape are the ABCC values to approximate basic education: They are included as a linear parameter to control for initial levels of education, and they are added as squared values to test for higher levels of education. As a result, the linear (and hence lower) ABCC values have a significant and negative impact on equality, while higher levels of education (squared ABCCs) have a significant and

[27]The low inequality of non-Hispanic countries might be due to the institutional framework created by slavery. As both men and women were torn away from their home countries and had to work equally, the "traditional" gender roles did not evolve as they did in other countries. Besides, Caribbean women tended to work outside the household more often than Latin American women (Manzel and Baten 2009).

positive impact on gender equality. The authors also plot the estimated values to illustrate the U-shape: The downward slope tends to be smooth, whereas the upward sloping part is strongly observable in the data. Hence, they demonstrated that Goldin's hypothesis also applies to basic education in Latin America and the Caribbean.

The Gender Gap in Asia

Friesen et al. (2013) test the U-shaped hypothesis for 14 countries in Asia from 1900 to the 1960s.[28] They use the ABCC index to approximate basic numeracy. Furthermore, they employ the educational gender equality index based on the Whipple index in the same way as Manzel and Baten (2009) did. Besides the age-heaping-based indicators, Friesen et al. (2013, p. 7) discuss literacy and school enrollment rates in the Asian countries in the dataset that clearly indicate high levels of inequality between men and women.

The analysis of the ABCC values provides further information on basic education between the sexes, especially when enrollment rates are not available for some of the regions. The authors find different results for the women's ABCC indexes among the observed regions: The vast majority of Southeast Asian women were already numerate around 1900, especially in Hong Kong and Thailand, while Indonesia lagged slightly behind (Friesen et al. 2013, p. 18). However, the picture looks different for women in South and West Asia: While Sri Lanka began with ABCC values of around 59% in 1900 and reached almost full numeracy in the 1950s, all of the other countries in this region reflected values far below. Women from Pakistan and Bangladesh had the lowest levels, not even reaching values of 50% toward the end of the period (Friesen et al. 2013, p. 16).

The equality index primarily reflects the different stages of the U-hypothesis. In the countries with very low human capital values for both women and men, such as Pakistan, Bangladesh, and India, equality values are only slightly below zero, indicating relative equality between women and men (Friesen et al. 2013, p. 23). This is also the case for the countries with high numeracy values, for example, Hong Kong and Thailand, for which the equality values range slightly below or above the zero line (Friesen et al. 2013, p. 25). The equality indexes of most of the other countries lie considerably below zero (e.g., in Indonesia or Sri Lanka). Most of the countries with negative values experienced an increase toward the end of the period, which in some cases even turned the negative into a positive index, such as in the Philippines. The opposite effect takes place in Afghanistan, for instance. While the inequality is not as high around 1910 (about −12), it decreases continuously until reaching a value below −60 in the 1950s (Friesen et al. 2013, p. 23).

[28]Included countries are Afghanistan, Bangladesh, India, Iran, Sri Lanka, Nepal, Pakistan, Hong Kong, Indonesia, Cambodia, Federation of Malaya, Sarawak, the Philippines, and Thailand.

In the next step, the authors test the U-hypothesis in different regression models in which the equality index is the dependent variable. They control for factors such as female voting rights and religion. The most important determinant, the ABCC index, is included as a linear and a squared parameter (as in Manzel and Baten 2009). The results for the ABCCs are always highly significant, and the correlation is negative for the linear ABCCs and positive for the squared ones. Furthermore, Friesen et al. (2013, p. 35) plot the regression results to illustrate the fitted values. The scatterplot shows an exact U-shaped pattern. Hence, the assumption of low gender inequality at low levels of human capital, rising inequality at increasing levels of education and, in the last phase, high levels of education and equality is fulfilled in the analysis of the 14 Asian countries under study.

Conclusion: The Impact of Numerical Abilities on Growth

In this chapter we showed that the age-heaping technique provides a unique opportunity to approximate basic education, especially in preindustrial times. One might argue, though, that the mere knowledge of numeracy levels between different countries, for example, does not contribute to achieve a higher goal. However, although numeracy correlates strongly with literacy, number discipline might even have a larger impact on the development of market exchange (see, e.g., De Moor and Van Zanden 2010). In many cases, we do not even know what literacy measures exactly: A broad range reaching from "is able to read and write" to "is only able to sign with his or her name" is possible. On the other hand, numeracy, or the ability to count, is the basis for participating actively in market mechanisms and for the emergence of capitalism. Crayen and Baten (2010a) show that numerical skills, in fact, have a strong impact on growth patterns across different world regions. In their analysis, the authors regress GDP growth rates on various factors, "growth capabilities," such as initial GDP levels and numeracy, approximated by the Whipple index, as well as a number of other control variables. It turns out that numeracy has not only a significant but also an economically meaningful impact on the growth rates of the included countries. Hence, the economy of those countries displaying higher levels of numeracy also grows at a faster pace than the economy of the countries with lower numeracy. All in all, we showed that age-heaping-based human capital estimates provide the opportunity to track potential reasons for the divergence of countries or world regions in the very long run.

References

A'Hearn B, Baten J, Crayen D (2006) Quantifying quantitative literacy: age heaping and the history of human capital. Economics working paper no.996, Universitat Pompeu Fabra

A'Hearn B, Baten J, Crayen D (2009) Quantifying quantitative literacy: age heaping and the history of human capital. J Econ Hist 69:783–808

Acemoglu D, Johnson S, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. Am Econ Rev 91:1369–1401
Acemoglu D, Johnson S, Robinson JA (2002) Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q J Econ 117:1231–1294
Armstrong A (1972) The use of information about occupation. In: Wrigley EA (ed) Nineteenth-century society: essays in the use of quantitative methods for the study of social data. Cambridge University Press, Cambridge, pp 191–310
Bachi R (1951) The tendency to round off age returns: measurement and correction. B Int Statist Inst 33:195–221
Baten J, Juif D (2013) A story of large land-owners and math skills: Inequality and human capital formation in long-run development 1820–2000. J Comp Econ. https://doi.org/10.1016/j.jce.2013.11.001
Baten J, Ma D, Morgan S, Wang Q (2010) Evolution of living standards and human capital in China in the 18–20th centuries: evidences from real wages, age-heaping, and anthropometrics. Explor Econ Hist 47:347–359
Baten J, Mumme C (2010) Globalization and educational inequality during the 18th to 20th centuries: Latin America in global comparison. Rev Hist Econ 28:279–305
Becker SO, Woessmann L (2009) Was Weber wrong? A human capital theory of protestant economic history. Q J Econ 124:531–596
Budd JW, Guinanne T (1991) Intentional age-misreporting, age-heaping, and the 1908 Old Age Pensions Act in Ireland. Popul Stud 45:497–518
Botticini M, Eckstein Z (2007) From farmers to merchants, conversions and diaspora: human capital and Jewish history. J Eur Econ Assoc 5:885–926
Charette MF, Meng R (1998) The determinants of literacy and numeracy, and the effect of literacy and numeracy on labour market outcomes. Can J Econ 31:495–517
Clark G (2007) A farewell to alms: a brief economic history of the world. Princeton University Press, Princeton
Crayen D, Baten J (2010a) Global trends in numeracy 1820–1949 and its implications for long-term growth. Explor Econ Hist 47:82–99
Crayen D, Baten J (2010b) New evidence and new methods to measure human capital inequality before and during the industrial revolution: France and the US in the seventeenth to nineteenth centuries. Econ Hist Rev 63:452–478
De Moor T, Van Zanden JL (2010) "Every woman counts": a gender-analysis of numeracy in the low countries during the early modern period. J Interdiscipl Hist 41:179–208
Dell M (2010) The persistent effects of Peru's mining mita. Econometrica 78:1863–1903
Duncan-Jones R (1990) Structure and scale in the Roman economy. Cambridge University Press, Cambridge
Friesen J, Baten J, Prayon V (2013) Women count: gender (in-)equalities in the human capital development in Asia 1900–60. Working paper, University of Tuebingen
Galor O, Moav O, Vollrath D (2009) Inequality in landownership, the emergence of human-capital promoting institutions, and the Great Divergence. RevEcon Stud 76:143–179
Glaeser EL, La Porta R, Lopez-de-Silanes F, Shleifer A (2004) Do institutions cause growth? J Econ Growth 9:271–303
Goldin C (1995) The U-shaped female labor force function in economic development and economic history. In: Schultz TP (ed) Investment in women's human capital. The University of Chicago Press, Chicago, pp 61–90
Juif D-T, Baten J (2013) On the human capital of 'Inca' Indios before and after the Spanish conquest. Was there a "pre-colonial legacy"? Explor Econ Hist 50:227–241
Juif D-T, Baten J (2014) Dangerous education? The human capital of Iberian and Latin American Jews and other minorities during the Inquisition. Working paper, University of Tuebingen
Manzel K, Baten J, Stolz Y (2012) Convergence and divergence of numeracy: the development of age heaping in Latin America from the seventeenth to the twentieth century. Econ Hist Rev 65:932–960

Manzel K, Baten J (2009) Gender equality and inequality in numeracy: the case of Latin America and the Caribbean 1880–1949. Rev Econ Hist 27:37–74
Mokyr J (1983) Why Ireland starved: a quantitative and analytical history of the Irish economy, 1800–1850. George Allen and Unwin, London
Myers RJ (1954) Accuracy of age reporting in the 1950 United States census. J Am Stat Assoc 49:826–831
Ó Gráda C (2006) Dublin Jewish demography a century ago. Econ Soc Rev 37:123–147
Reis J (2005) Economic growth, human capital formation and consumption in Western Europe before 1800. In: Allen RC, Bengtsson T, Dribe M (eds) Living standards in the past. Oxford University Press, New York, pp 195–227
Schofield RS (1973) Dimensions of illiteracy, 1750–1850. Explor Econ Hist 10:437–454
Stolz Y, Baten J (2012) Brain drain in the age of mass migration: does relative inequality explain migrant selectivity? Explor Econ Hist 49:205–220
Tollnek F, Baten J (2013) Farmers at the heart of the educational revolution: which occupational group developed human capital in the early modern era? Working paper, University of Tuebingen
Zelnik M (1961) Age heaping in the United States census: 1880–1950. Milbank Q 39:540–573

Church Book Registry: A Cliometric View

Jacob Weisdorf

Contents

Abstract

This chapter links economic history to demography, looking into the use of church book data to investigate topics in economic history. Using the Malthusian population model to cast light on scholarly debates about the Great Divergence and the wealth of nations, the chapter illustrates some of the main advantages (and drawbacks) to using church book registry in this context.

Keywords

Cliometrics · Demography · Development · Great Divergence · Malthusian model · Church book registers

Introduction

Church book registers provide data regarding three main life events: births, deaths, and marriages. Two claims central to the use of such statistics in economic history are that economics influences all of them and that all of them in turn influence economics. For example, the timing of a marriage or childbirth in the past depended on people's earning possibilities, and a death or a miscarriage often followed from

J. Weisdorf (✉)
University of Southern Denmark and CEPR, Odense M, Denmark
e-mail: jacobw@sam.sdu.dk

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_1

harvest failure and hunger. Conversely, births and deaths determine the size of the population, which impact on prices and wages through people's demand for goods and supply of labor.

These relationships can be formulated in a widely applicable, but visually simple framework: the Malthusian population model. Scholars commonly believed that historical societies were characterized by *Malthusian population dynamics*. These dynamics are easily illustrated in terms of Fig. 1, which captures the links between economics and demography outlined above. Malthus hypothesized that changes in wages exercised a dual effect on the growth of a population (Malthus 1798). On the one hand, lower wages causes fewer and later marriages, leading therefore to fewer births. This *preventive check* mechanism is captured by the upward-sloping birth schedule in Fig. 1. Lower wages simultaneously raise death rates, capturing the *positive check* mechanism illustrated by the downward-sloping death schedule in Fig. 1. The intersection point between the birth schedule and the death schedule determines the equilibrium wage rate, defined as the wage rate that keeps the population constant over time (a *steady state*).

The dynamics of the Malthusian population framework are completed by the addition of Ricardo's notion that population growth historically drove down the marginal product of labor (Ricardo 1817). This feature, in the Malthusian model, often builds on two key assumptions: a constant returns-to-scale production technology and a fixed factor of production (normally land). Hence, when wages are above the equilibrium wage rate, and births thus exceed deaths in Fig. 1, the population size grows. Diminishing returns to labor in production then puts downward pressure on the wage rate, leading to fewer births and more deaths, until the wage rate eventually returns to its equilibrium level and births equal deaths.

The Malthusian framework provides a powerful tool to help understand why some historical societies were rich and others poor. That is, a permanent deviation in the equilibrium wage rates between two societies must be grounded in different structural arrangement causing the positions of birth and death schedules to differ. Within this context, the underlying question often asked by scholars concerned with these topics are the following: what are the short- and long-term effects of shocks to the Malthusian system, and how might these shocks consolidate themselves in permanent shifts in the positions of the birth and death schedules? The answers

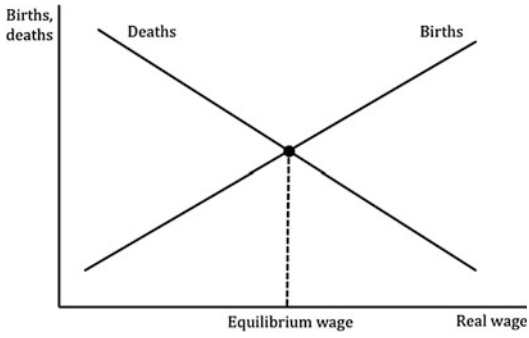


Fig. 1 The Malthusian model

can cast light on several unresolved debates in economic history, and church book registry is an excellent tool to achieve such hindsight.

One such debate concerns the Great Divergence in real wages within Europe. Why did England and the Low Countries pull away from the rest of Europe between 1500 and 1800? Bob Allen's European *welfare ratios*, which measure the number of (prefixed) consumption baskets that a family can afford (Allen 2001), are a great illustration of this (Fig. 2).

A temporary divergence in real wages can, of course, be viewed as a disequilibrium episode. Any shock that pushes the Malthusian system out of its steady state will ignite the dynamics described above and ensure a return to the system's steady state. That is, wages will fall back to their original level. The Black Death, a huge and sudden decline in Europe's population in the mid fourteenth century, is the perfect example of that. This disequilibrium approach, however, does not complete the picture. As the population level gradually restabilized in the centuries following the plague, most European wages fell back to their pre-Black-Death level. But in England and the Low Countries they did not. This leaves the question: did the Black Death do more than just push the system out of its equilibrium temporarily? Did it entail a structural transformation that shifting the birth schedule or the death schedule, escalating the English and Dutch equilibrium wage level?

There are multiple candidates for a structural change to the system. These have been extensively discussed by theorists. One example concerns the influence of the Black Death briefly mentioned above. It is well documented that the Black Death, by reducing the English population by almost 50%, entailed a "golden age of the English peasantry." That is, a dramatic increase in workers' remuneration as landowners struggled to recruit and retain laborers. De Moor and van Zanden (2010) and

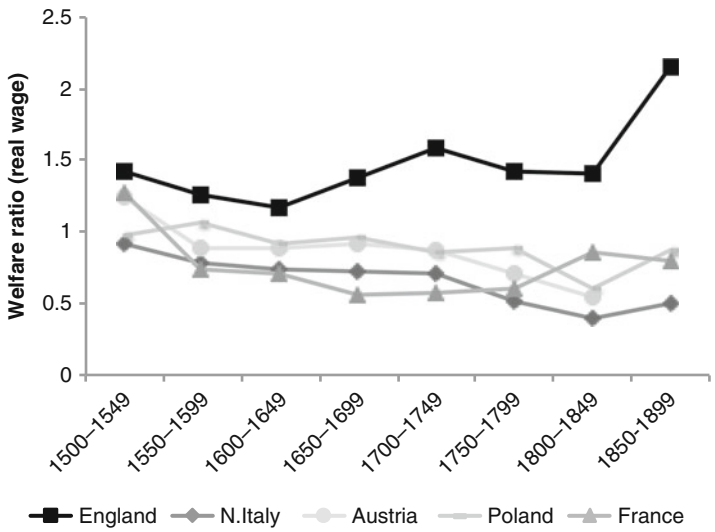


Fig. 2 The Great Divergence within Europe (Source: Allen (2001))

Voigtländer and Voth (2013) have linked the economic superiority of Northwestern European economies after 1500 to the demographic and economic legacy of the Black Death. The argument, which runs from the influence of women's remuneration on the timing of their marriage, is inspired by John Hajnal's hypothesis about the *European Marriage Pattern*. Hajnal (1965) noticed that for much of the medieval and early modern period, a line drawn from St. Petersburg to Trieste demarcated distinctive demographic regimes: in the east, women married young and almost everybody are married; in the west, brides were older and celibacy was higher. De Moor and van Zanden and Voigtländer and Voth have interpreted these distinct scenarios in terms of differences in the economic opportunities of women. Women's improved position in the post-plague labor market and especially the growth of opportunities as servants in husbandry linked to the relative expansion of "horn" (in which women had a comparative advantage) versus "grain" (in which they did not) allegedly pushed up female wages and labor force participation. Since women, unlike men, split their time between raising children and working, improved wages and labor opportunities for women place a premium on child birth, causing delayed marriage and increased celibacy, both resulting in reduced fertility. In terms of Fig. 1, the rising premium on childbirth would shift the Malthusian birth schedule downward inaugurating Hajnal's European Marriage Pattern. As an indirect consequence of the Black Death, Northwest Europeans therefore found themselves in a new steady state with greater female employment, later marriage, lower fertility, and higher per capita incomes (Voigtländer and Voth 2013).

Other hypotheses have been forwarded to try to motive Northwest Europe's economic superiority through shifts in the birth or death schedules in Fig. 1. De Vries's (2008) idea that a "consumer revolution" preceded the Industrial Revolution provides an alternative approach to understanding the rise of a premium on children and hence a downward shift in the position of the birth schedule in Fig. 1. The consumer revolution refers to the introduction of novel commodities (such as tea, coffee, sugar, books, and, clocks) between 1600 and 1750. The argument made, then, is that households have *love-of-variety* preferences and hence accommodate the novel commodities in their consumption basket by allocating *less* resources to the goods they already consume. Since children, or the goods that children consume, are among the goods already consumed, the demand for these goods, and hence for children, declines. Thus, the consumer revolution shifts the birth schedule downward, causing a new equilibrium with fewer births and deaths and a higher level of wages (Guzman and Weisdorf 2010).

Technical progress offers a third explanation for why the cost of children went up and pushed the birth schedule down. Galor (2011) hypothesizes that the complexity of new and more advanced jobs that industrialization entailed placed a premium on education. This incited parents to increase the investment in the human capital of their offspring, powered by a decline in their number of births. The increased cost per child would mean the birth schedule shifted downward, generating a new equilibrium of fewer births and fewer deaths but higher real wages.

Still others, such as Clark (2007), have used the Malthusian framework to highlight the "benign" effect of fatality on wages. Indeed, anything that pushes the

death schedule of Fig. 1 upward will cause the equilibrium wage rate to permanently rise. Voigtländer and Voth (2014) have use this point to draw a link between European wars and high rates of urbanization to explain Europe's economic prosperity vis-à-vis that of other world continents (i.e., the Great Divergence between Europe and the rest).

Common to all these theoretical accounts of what explains the Great Divergence – both within Europe and between Europe and other continents – is a lack empirical depth. This is where church book records can bring news to the table. The next section gives an illustration of the nature of the vital information concealed in church books, followed in the subsequent section by a series of examples of how previous studies have used the church recordings to cast light on some of the main research questions surrounding the debates about the Great Divergence and the wealth of nations. The chapter concludes by pointing toward some future research roads building on source material derived from church book registers.

The Nature of Church Book Registers

Early civil registration in Europe was usually done by the church on request from the crown. From the mid-nineteenth century on, civil registration was gradually taken over by secular institutions. Systematic census registrations, often first done every 5 or 10 years but later annually, were slowly replaced by a central authority gathering vital information as it appeared. The information recorded in the population censuses and later the central registers is obviously superior to that of church book records, partially because they include the entire population regardless of religious affiliation and partially because census data provides a spot image of the entire population and not just those reporting a vital event to the church in a given year.

The main advantage of the church book registry (also known as parish registry) is that these provide large-scale vital information before 1800, i.e., the period during which the Great Divergence began to take hold. Although the church book registry only captures people during three life events – birth, death, and marriage – it is nevertheless able to inform us about some key links between economics and demography, and vice versa, as we shall see below.

In the Old World, church registers become widespread in the late middle ages or early modern period. In the New World, notably in many of today's developing countries, vital registration began with the arrival of European missionaries after the mid-nineteenth century, a practice that has continued up until the present day. Although church book registers appeared much later in the New World than in the Old World, the fact that central registers emerged relatively late in many developing countries, often not until the 1960s or even later, makes church book registry of the New World a particularly interesting source of vital information during the late eighteenth and early nineteenth centuries.

Perhaps the most prominent dataset building on church book registration, and certainly the one most frequently subjected to scholarly scrutiny by economists and economic historians in the past several decades, is the so-called CAMPOP data

(Wrigley and Schofield 1981), collected by the *Cambridge Group for the History of Population and Social Structure*, which was founded in the 1960s by Peter Laslett and Tony Wrigley. The Group's work on collecting, transcribing, and analyzing English church book information, an effort spanning nearly five decades, has been used for three main purposes. The first is population *back projection*. By starting with the population level of the English census of 1831, and then counting the annual number of births and deaths recorded in 404 well-documented English parish registers, the Group was able to come up with an estimate of the size of the English population back to c. 1541 when church registration first began. The second purpose is *family reconstitution*. This is based on the idea, developed by French demographer Louis Henry, that vital events can be used to track the marriage date of a married couple, as well as the birth and death dates of their parents and their offspring, hence reconstructing the entire family based on church book statistics. The third and last purpose of the Group's work, which is still ongoing, is a reconstruction of the occupational structure of Britain based on the male occupations recorded in the church books.

One of the key advantages of church book registers is the information it provides *in addition* to the dates of the vital events. The details of the recorded statistics depend, of course, on the recording policies of the church in question. True of both the Protestant and Catholic churches, the recording of a birth (or baptism) would usually include the names of the parents as well as the time and place of the child's baptism. A marriage record would hold the names of the spouses, their civil status before the marriage, and the time and the place of marriage. This is occasionally supplemented by the names of fathers of the spouses, as well as those of (usually two) witnesses. Lastly, a death (or burial) would contain the name of the deceased person and the date and place of the burial.

An important notice concerning the dates is that church books normally record the dates of baptisms and burials rather than the dates of births and deaths. However, the time intervals between the ecclesiastical and the vital events were usually rather short. For obvious reasons, people were buried immediately after their death, typically in England within 3 days of death (Schofield 1970). Furthermore, English children were usually baptized within 1 month of birth (Midi Berry and Schofield 1971), although this could vary somewhat depending on local traditions and the distance from the family home to the church.

Also true of most Christian churches is that they would ask the parents or spouses (as well as fathers and witnesses) to certify and endorse the event in question by placing their signature in the church book. This practice has served as an important measure of literacy rates in past societies. When someone was unable to sign their name, the vicar would write their name instead, and the illiterate person was simply asked to leave a mark in its place to prove his or her consent. While it is obvious that people who are able to write down their name are not necessary literate, a signature has proven to be a reasonable proxy for this (Schofield 1973).

Sometimes the church registers hold even more profound information about human capital attainments than an indication of their literacy status. Some churches, during some time periods, also recorded the occupational title of the individuals

involved in the registration of the vital event. This is often (but not always) the case in Protestant church registers, a practice that, for Anglican Protestants especially, was made compulsory by the passing of Rose's Law in 1812, specifically asking the ministers to record the occupation title of parents, spouses, and fathers-in-law (and sometimes even the witnesses).

The recording of someone's profession provides a critical insight into the socio-economic conditions of that person, including his or her social status, working skills, and income potential. Occupational information would sometimes even include individual land holdings, providing further knowledge about social status and wealth of the person in question.

There are several ways in which occupational information can be coded and thus made subject to systematic studies of the links between demographic variables and socioeconomic conditions at the individual level. Starting with one of the broader systems for categorizing professions, the *Primary-Secondary-Tertiary* (PST) system, developed by Tony Wrigley of the Cambridge Group (Wrigley 2010), has been used to code the entire occupational dataset collected from British church books in order to study the occupational structure of Britain since medieval times. The great advantage of the system is its classification of all occupations depending on whether the work related to primary, secondary, or tertiary sector activities. A main downside to this system, however, which the Group is still struggling to solve (ibid.), is the problem of categorizing the occupation title "laborer," which was not only a very common occupational title but also one that does not reveal the nature (and sector) of the work conducted.

Another classification system, which is comparable as well as compatible to the PST system, is the *Historical International Standard Classification of Occupations* (HISCO), developed by Marco van Leeuwen, Ineke Mass, and Andrew Miles and documented in Van Leeuwen et al. (2002). This HISCO is an extension of ISCO (International Standard Classification of Occupations) for which the International Labour Organization (ILO) is responsible. The HISCO contains 1,675 historical job categories. The world coverage of the HISCO, along with its time range (spanning the sixteenth to twentieth centuries), allows a categorization of occupational titles from almost any historical population worldwide in which historical occupational records exist.

In a subsequent book, titled *HISCLASS*: *A historical international social class scheme*, labor historians have ranked all the occupations coded in HISCO based on an assessment of the working skills required for an average performance on the job (van Leeuwen and Maas 2011). The ranking of occupational titles builds on the principles of the *Dictionary of Occupational Titles* (DOT). The DOT was developed in the 1930s by the US Employment Service in response to a rising demand for standardized occupational information to assist job-placement activities (US Department of Labor 1939). In order to efficiently match jobs and workers, the public employment service system required that a uniform occupational language be used in all of its local job service offices. Through an extensive occupational research program, occupational analysts collected and provided data to job-market interviewers to help them match the specifications given in job openings to the

qualifications of job applicants. Based on the data collected by occupational analysts, the first edition of the DOT was published in 1939, containing some 17,500 job definitions, presented alphabetically, by title, with a coding arrangement for occupational classification.

The transformation in HISCLASS of occupational titles into working skills builds on two main scores used in the DOT: the *general educational development* score and the *specific vocational training* score. The score concerning the general educational development captures three key features regarding intellectual competencies necessary to fulfill the tasks and duties of an occupation: the incumbent's reasoning development, his or her ability to follow instructions, and the acquisition of language and mathematical skills needed to conduct the work. The score concerning specific vocational training captures the time investments needed in three main areas: that required by the worker to learn the techniques used on the job, that needed to acquire the relevant information to conduct the work, and that necessary to develop the competencies required for an average performance in a job-specific working situation.

Building on the expertise provided by Bouchard (1996) and a team of labor historians, van Leeuwen and Maas used the two DOT scores to code the occupational titles categorized in HISCO according to the skill content of the working titles contained in the HISCO, as part of a procedure to create a historical international social class scheme. In HISCLASS, occupational titles are grouped in four categories as either *unskilled*, *lower skilled*, *medium skilled*, or *higher skilled*. Ongoing work by van Leeuwen et al. (2014) is taking the skill categorization one step further, estimating the actual time investment needed to conduct the work that described the entire set of occupational titles contained in the HISCO system (van Leeuwen and Maas 2011). A further advantage of the HISCLASS scheme is its division of workers into blue-collar (manual) and white-collar (nonmanual) work.

Alan Armstrong's occupational classification scheme offers an alternative to using HISCLASS, splitting jobs into five class categories (Armstrong 1974): Professional, Intermediate Occupations, Skilled Occupations, Partly Skilled Occupations, and Unskilled Occupations. Both systems (HISCLASS and Armstrong's) are useful in their own rights depending on the question at hand. A further advantage of the HISCO scheme, however, is its extension system called HISCAM, a scheme for coding occupations according to the social status of the work linked to the job title offering a finer categorization of social status than the HISCLASS (Lambert et al. 2013). The SOCPO, a competing scheme to the HISCAM, provides a similar coding of occupational titles into social class (Van De Putte and Miles 2005).

Social status and working skills are, of course, both imperfect approximations of individual income or wealth. Greg Clark and Neil Cummins' work, which uses will records to link wealth to professional titles, provides a mapping of occupations into seven social groups based on the wealth recorded in the wills, as described in Clark and Cummins (2010). From the poorest to the richest, these social groups are laborers, husbandmen, craftsmen, traders, farmers, merchants, and gentry. This classification is particularly helpful for looking at links between income potential and fertility decisions (discussed below).

Last but not least, the church book data truly comes to life when combined with other database information. So far, very little work has been done in this regard. Klemp et al. (2013) offer a demonstration of this, linking the CAMPOP data to statistics regarding apprenticeship (see further below). Other possibilities include combinations with census data, will records, probate inventories, poor law information, and tax records. Much work needs to be done in this regard.

Church book records become especially helpful when they come in the form of reconstructed families. The huge advantage of family reconstitution data is the linkage of family members across family generations. This enables studies of intergenerational social mobility, marriage patterns, birth and death patterns, and much more. Although the work needed to reconstruct families based on the raw vital events can be quite laborious, the procedure is surprisingly simple. Start with a marriage. Then track the records back in time to find the birth date of the spouses (linking them to their parents) and possibly any previous marriage (indicated by the civil status at their current marriage). Go forward in time to find the death date of the spouses and, if the couple went to baptize (or bury) any children, to find the birth, marriage, and death dates of their offspring. It took several decades for the Cambridge Group to reconstitute the families within 26 English parishes (Wrigley et al. 1997). But this was before modern computer programming appeared that can aid this process significantly.

The work of reconstituting families based on church book data is further complicated by the fact that people do not always remain in their parish of origin or indeed in a parish where they were once observed. The flipside to that problem is that the lack of someone's birth or death indicates they moved into or out of the parish in question, conveying important information about patterns of migration in the past (Souden 1984). A head-on way of dealing with the issue of lifecycle migration is by tracking down individuals as they move from place to place (arguably an even more laborious task than sticking to the same location). The French TRA data provide such statistics, tracking individuals whose names begin with "Tra" (as in "Travers," a common French family name) across time and space. Comparable datasets exist for other European countries as well.

Some scholars have raised criticism against the transformation of church book data into family reconstitutions. Perhaps the most prominent critiques of the work done by the Cambridge Group come from Peter Razzell (2007) and Steven Ruggles (1999). Much of their criticism is focused against underregistration, linkage failure, selection bias, and the consequences thereof. These potential shortcomings are good to keep in mind when working with family reconstitutions.

How the Registers Have Been Used

There are numerous examples of how church registers have been used to analyze topics in economic history. This section focus on some recent studies connected to the debates regarding the Great Divergence and the wealth of nations, notably how and why the development path of rich countries parted from that of poor countries.

The Malthusian population framework described above has often served as a starting point for analyzing these questions. This scholarly work is split into two categories. One sets out to assess the relevance of the Malthusian model and its two main components, the *positive check* and the *preventive check*, for different countries and regions. The other uses the implications of the Malthusian model to understand various aspects of the Great Divergence and the wealth of nations. Church book data provide a key empirical foundation for analyzing both types of work.

Probably the most prominent statistics used for these purposes (the CAMPOP data discussed in the previous section) are based on British parish registers. These data are made available by the Cambridge Group (Wrigley and Schofield 1981; Wrigley et al. 1997). There are two main reasons for the large popularity of these data. One reason is that the British parish registers are of very high quality and cover three centuries of British population history, from the origins of parish registration in 1541 until the main census registrations started to appear in the early nineteenth century ending in 1871. The other reason for their high esteem is that England was the world's economic leader between 1500 and 1800 and that she was the first nation worldwide to experience an industrial revolution.

Tests of the relevance of the Malthusian population framework span from relatively uncomplicated empirical investigations, exploring the existence of short-term *preventive* or *positive check* mechanisms, to highly advanced econometric examinations of the short- and long-term dynamics and stability of the entire Malthusian framework.

Despite a strong belief in the relevance of the Malthusian framework and its widespread use to understand the process of economic development in preindustrial societies, there is surprisingly little evidence in support of the preventive check hypothesis (Kelly and O Grada 2012). This has faced scholars with a large challenge, because the idea that falling living standards entail a short-term reduction in birth or marriage rates seems particularly appealing to the English case.[1] There is general agreement among scholars, though, that the Malthusian population model is correct and thus that the failure to obtain supporting evidence is due to issues of data and mismeasurement.

A key suspect for the lack of empirical support for the preventive check in England is data aggregation. Ideally, one would explore the direct link between the living standard of a particular couple and the demographic decisions (marriage or birth) made by this couple. But the scholarly reality is that living standards (captured by wages and prices) as well as vital rates (captured by marriage and birth rates) are often measured at the national level. This inaccuracy can be eliminated by moving from the macro to the micro level.

Morgan Kelly and Cormac O Grada have taken a first step in this direction, looking for preventive checks at the *parish* level (Kelly and O Grada 2012). Instead

[1]Some scholars have even found evidence of the opposite, documenting a positive relationship between nuptiality and the price of wheat, a phenomenon they coined *permissive checks* (Sharp and Weisdorf 2009).

of using the aggregated data of the 404 parish registers included in the CAMPOP file, they look at the parish level response to changes in real wages. Even though the real wages are still at the national level, Kelly and O Grada are able to document the existence of preventive checks in many of the parishes included (but not all).

An even closer inspection of the preventive check mechanism requires access to even more detailed vital accounts than aggregations at the parish level. This is where the CAMPOP's family reconstitution data prove useful. While other works have relied entirely on the use of *crude* vital rates, meaning the number of birth, death, and marriages per 1,000 population, it is clear that these rates are only a rough approximation for family-level decision variables, such as the timing of a marriage and a birth.

Cinnirella et al. (2013) have used the CAMPOP's family reconstitution data to try to measure the effect of real wages and food prices on the timing of the marriage, the timing of the first birth, the timing of subsequent births, and the timing of the last birth. They find strong evidence of a *preventive check* mechanism operating in England in the three centuries leading up to England's fertility decline of the nineteenth century. Figure 3 illustrates how birth-spacing intervals expand when real wages are low, and vice versa. Although the wages and prices used to measure standards of living are still at the national level (and certainly never at the family level), the church book recordings of the occupational titles of the husbands help control for the exposure (or lack thereof) to economic pressure during economic downturns. The work similar to that of Cinnirella et al. has been done for Sweden (Bengtsson and Dribe 2006) and Germany (Dribe and Scalone 2010), also showing evidence of deliberate within-marriage birth-spacing behavior.

Among examples of more advanced (and holistic) approaches to testing the relevance of the Malthusian framework, Esteban Nicolini's original work, as well

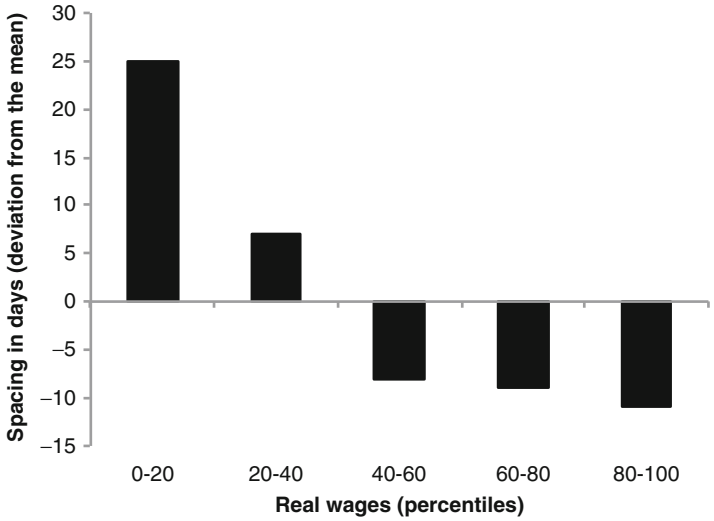


Fig. 3 Birth-spacing intervals by real wage percentiles (Source: Cinnirella et al. (2013))

as follow-up work by Marc Klemp, Niels Frameroze Moeller, and Paul Sharp (cited below), deserves a mention. Common to these studies is the use of crude vital rates for birth, deaths, and marriages (notably of the CAMPOP data), which they then attempt to link up with historical living standards measured by (national) real wages, usually provided by Clark (2005).

Nicolini's (2007) work does not fit some crucial assumptions of the Malthusian model. Using a vector autoregression for data on fertility, mortality, and real wages over the period 1541–1841 and applying a well-known identification strategy broadly used in macroeconomics, Nicolini's results show that endogenous adjustment of population to real wages functioned as Malthus assumed only until the seventeenth century: evidence of positive checks disappeared during the seventeenth century and evidence of preventive checks disappeared before 1740. This implies that the endogenous adjustment of population levels to changes in real wages – one of the cornerstones of the Malthusian model – did not apply during the period of the Industrial Revolution.

Moeller and Sharp (2014) reexamine the question using data identical to those of Nicolini but with a somewhat different economic specification. They formulate a post-Malthusian hypothesis that on the one hand involves co-integration between real wages and the birth and death rates. But on the other hand, it allows a negative Malthusian feedback effect from population on income (as implied by diminishing returns to labor) to be offset by a positive so-called *Boserupian-Smithian* scale effect of population on technology. This setup means they reach a different set of conclusions from Nicolini, namely, that, as early as two centuries preceding the Industrial Revolution, England had already escaped the pattern described by the standardized Malthusian model and instead had entered a post-Malthusian regime, where income per capita continued to spur population growth, but that the real wage was no longer stagnant. Tests of the relevance of the Malthusian (or post-Malthusian) framework are not confined to Britain. Klemp and Moeller (2013) have also experimented with church book data from Denmark, Norway, and Sweden, looking for evidence for the existence of a post-Malthusian phase in the transition from stagnation to growth in Scandinavia, and studies of other regions are currently in the making.

Gregory Clark and Gillian Hamilton (2006) provide an example of a crossroad study between assessing the validity of the Malthusian model and using its predictions. One of the key features of the Malthusian model is that there is a unique wage rate at which births equal deaths. But since the reality is that some earn more than others, the Malthusian model implies that the rich have more surviving offspring than the poor. Clark and Hamilton used information derived from will records to test this implication, investigating the relationship between the total value of the wealth left by male testators and the total number of offspring who inherited their wealth. Their results are replicated in Fig. 4.

The same exercise can be conducted using church book family reconstituted data. What the church book registry lacks in terms of wealth information, it compensates for by its vital statistics. Not only does it provide the total number of births by family, but it also permits a count of how many of these children actually made it through into their reproduction period (i.e., lived beyond the age of 15). By exploring the

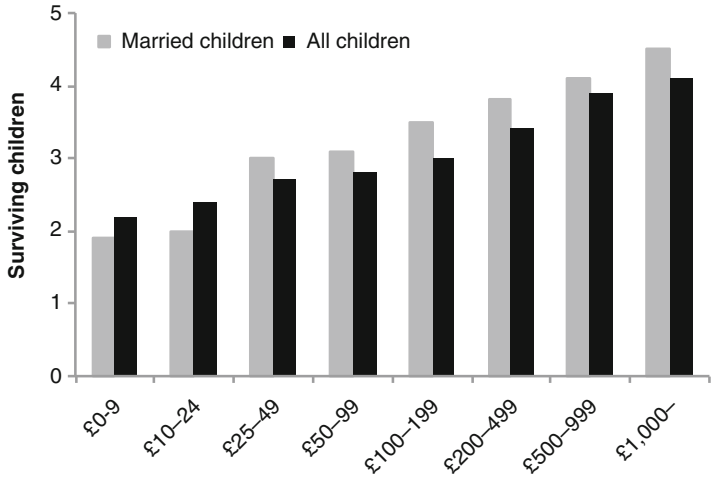


Fig. 4 Reproductive success by wealth (Source: Clark and Hamilton (2006))

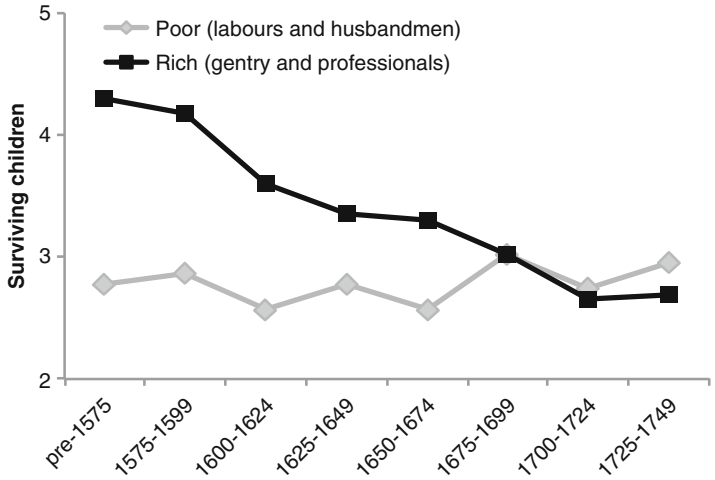


Fig. 5 Reproductive success of rich and poor (Source: Boberg-Fazlic et al. (2011))

CAMPOP statistics, Boberg-Fazlic et al. (2011) divided the male occupational titles found in the church registers across the seven income groups (see above) defined by Clark and Cummins (2010). Figure 5 illustrates how the preindustrial period confirms the inferences of the Malthusian model and also how this pattern dissolves around the time of the Industrial Revolution.

The CAMPOP family reconstitution was also used to look at Malthusian positive checks. While the magnitude of the *short-term* effects of hardship on mortality has received ample support (Galloway 1994), very little attention has been paid to the *long-term* effects: the influence of hardship on mortality later in life. Klemp and

Weisdorf (2012a) raised this question looking at the so-called "fetal origins" hypothesis. This is the idea that undernutrition in early life leads to a disproportionate growth in utero and in infancy, which in turn enhances the susceptibility to illness and hence increases the death risk later in life. Using survival analysis, they find that birth during the great English famine of the late 1720s entailed a largely increased death risk *throughout* life among those who survived the famine. The death risk at age 10 among the most exposed group – children born to English Midlands families of a lower socioeconomic rank – was up to 66% higher than that of the control group (children of similar background born in the 5 years following the famine). This corresponds to a loss of life expectancy of more than 12 years.

The Malthusian framework has been used repeatedly to understand the long-term economic development and the wealth (or lack hereof) of nations. One of the key arguments for why England enjoyed comparatively high living standards in the past is linked to the response of demography to economics. A central hypothesis concerns the parental trade-off between the quantity and quality of their offspring. The existence of a child quantity-quality trade-off is particularly relevant for the assessment of theories that explain the transition from millennia of economic stagnation to an era of sustained economic growth as well as the accompanying demographic transition (e.g., Galor and Moav 2002). Indeed, the leading theories explaining the origins of modern economic growth depend crucially on the presence of a trade-off between the number of children in a family and the attainment of human capital of the offspring. For instance, Galor and Weil (2000) have argued that the enhancement of technological progress during England's Industrial Revolution motivated parents to invest in the human capital of their offspring, leading to lower fertility and hence slower population growth, ultimately facilitating an increase in income per capita.

Census data has been a generous sponsor of the vital information needed to test the existence of a trade-off effect during early stages of industrialization. Basso (2012) has demonstrated the existence of a trade-off in Spain, Becker et al. (2010) in Prussia, Fernihough (2011) in Ireland, and Perrin (2013) in France.

Church book statistics, notably in the form of family reconstitutions, provide an alternative to using census data to test the relationship between the total number of family births and the human capital achievement of the offspring. The work by Marc Klemp and coauthors provide some examples. Using the CAMPOP data, Klemp and Weisdorf (2012b) show a negative link from parental reproductive capacity to the socioeconomic achievements of their offspring later in life. Using the time interval between the date of marriage and the first birth as a proxy for the couples' reproductive potential (i.e., their fecundity) and hence unplanned variation in family size, the authors establish that children of parents of low fecundity were more likely literate and employed in skilled and high-waged work than those of highly fecund parents. Along similar lines, Galor and Klemp (2013) have used Canadian church book data to show that a parental disposition toward having many children was not as conducive for long-run reproductive success as more moderate reproductive dispositions: subsequent generations of couples prone to restrained fertility turned out to be more successful in terms of reproduction than those of more fertile couples.

Finally, the church book data can be linked up with other databases. Klemp et al. (2013) provide an example of how the CAMPOP family reconstitution data can be linked with records of substantial educational achievements. Although the church books may provide someone's occupational title, and hence give a hint about the educational attainments of the person in question, they do not record any specific information about the schooling or actual occupational training. By use of a matching procedure, Klemp and coauthors were able to link up the vital statistics from the church books with information from nationwide Stamp Tax registers providing the names of apprentices and fees paid by apprentices to masters. The linkage of family data to individual apprenticeship training opens the possibility to explore a long line of questions regarding parental education decisions, such as whether parents followed customary tradition (such as birth order) or decided to educate children was based on their aptitude.

The availability of church book records is, of course, not limited to Britain and continental Europe. Wherever the European missionaries went, they left a trace revealing the local demography elsewhere. More than that, because the missionaries brought with them the recording methodologies used in Europe, church book registry elsewhere is often fully identical to the registry in Europe. This means church book data of the Americas, Asia, and Africa can help understand the economic development, or lack hereof, in the third world, particularly those areas that were previously colonized by Europeans.

One of the main topics in the context of understanding the Great Divergence between Europe and the third world is the influence of European colonial powers on the economic development of third-world regions. Figure 6 captures well the Great Divergence between England on the one hand and sub-Saharan Africa on the other (Frankema and van Waijenburg 2012).

Much of the data used to analyze the Great Divergence between Europe and the third world come from the colonizers themselves. For example, empirical

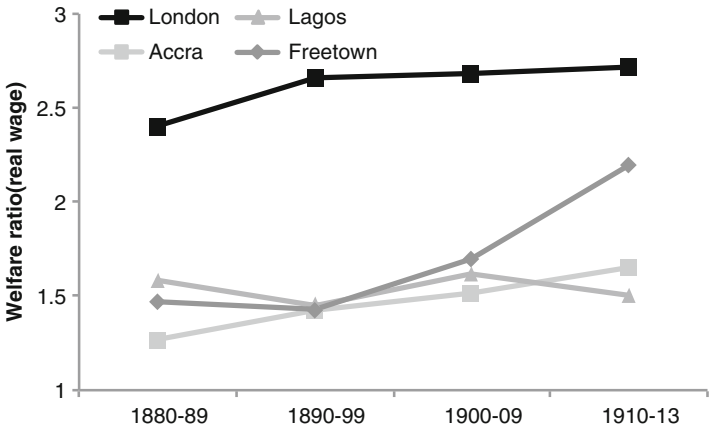


Fig. 6 The Great Divergence between Europe and Africa (Source: Allen (2001), Frankema and van Waijenburg (2012))

investigations into Africa's economic past are often limited to the study of national-level variables (production, export, taxes, etc.) recorded long ago by colonial agents who gave primacy to numbers concerning the colonizers' own activities. Church book records provide a source of information *independent* of those recorded by the colonizers. Christianity caught on rapidly in Africa, especially in sub-Saharan Africa, which today is predominantly Christian. Some Christian churches, such as the Anglican Church, recorded the occupational information of their affiliates. This means church book data represents a broader section of the population than those working for the colonial administration. Moreover, Christian missionaries often arrived ahead of the colonial agents, making it possible to explore the statistics of the church books to investigate not only the impact of colonial influences on Africans but also the results of African independence.

The work of Felix Meier zu Selhausen and coauthors demonstrates well the potential of Sub-Sahara African church book registers (Meier zu Selhausen 2014; Meier zu Selhausen and Weisdorf 2014; Meier zu Selhausen et al. 2014). Meier zu Selhausen and his collaborators have used marriage registers from one of the earliest and largest Protestant churches in sub-Saharan Africa, St. Paul's Cathedral in Kampala, Uganda, to study the long-term evolution in human capital formation and labor market participation among Protestants affiliates. British missionaries arrived in Uganda in the 1870s, shortly prior to the British colonizers who ruled Uganda until the 1960s. The chronology in the line of events makes it possible to study the demographic influence of the missionaries, followed by the colonizers, followed by the exit of the colonizers and subsequent independence of Uganda and up until the present day.

The consistent recordings of (especially) women's occupations since the arrival and spread of Protestant missionaries in Africa in the latter half of the nineteenth century made it possible to explore several aspects of gender (in)equalities and the influence hereon of both missionaries and colonial powers. One of the key indicators of female agency is the spousal age gap (Carmichael 2011): the older the husband is and the younger the wife, the more power the husband is assumed to hold and the less agency the wife has. This is also captured by the so-called *girl-power* index, measured as the female age at marriage minus the spousal age gap. By dividing women into two groups, depending on whether or not they engage in salaried work, Meier zu Selhausen (2014) finds that those women who worked for wages married significantly later than others and that the spousal age gap among these women was smaller, and the girl-power index higher, than among other women. Women rarely worked for the colonial administration, however. Their sole employer was the missionaries, who trained and used their expertise in mission schools and hospital work (as teachers, nurses, and midwives).

Other variables used to measure gender inequality include the literacy rates, the numeracy rate (ability to deal with numbers), the labor force participation rates, the wage rates, and the rates of skilled and nonmanual (high-status) workers. Meier zu Selhausen and Weisdorf (2014) found that males quickly acquired literacy, which helped provide access to formal-sector (salaried) jobs. Women took somewhat longer to obtain literacy and considerably longer to enter into salaried work. The authors observe a *gender Kuznets curve*: although inequality in literacy and access to

Fig. 7 The shares of Kampala men and women in salaried work (Source: Meier zu Selhausen and Weisdorf (2014))

salaried jobs grew substantially during the early colonial period, it gradually vanished during the postcolonial period. Today it is largely gone. Figure 7 shows the evolution in the share of men and women employed in salaried work in historic Kampala, indicating the gender gap therein.

The passing of Rose's Law in the early nineteenth century (mentioned above) meant that Anglican Protestant records are particularly useful for the purpose of studying social mobility. The reason is the recording of not only the spouses' occupations but also those of their fathers, setting the scene for a study of intergenerational social mobility at the family level. Meier zu Selhausen et al. (2014) found that social mobility in Uganda was very large during the colonial period, primarily because of the salaried labor market that arose following the creation of a colonial economy. After Uganda's independence from the British colonizers, economic development slowed down; fewer new jobs were created; and the prospects for social mobility declined.

Uganda is just one example of how church book records from third-world countries can be explored to shed light on the Great Divergence between Europe and the third world. The missionaries' recordings of vital events can be found practically everywhere the missionaries went, including most of today's developing regions covering the continents of Africa, Asia, and South America.

What Is Next?

There are two scholarly roads forward that can make even better use of church book data in future research. The first is to improve the use of existing data. Many independent datasets exist, but these are currently not directly comparable, making

it difficult to conduct cross-country or cross-regional comparison. Comparable work is important to reach an understanding of the influence of demography on the different economic performances of past economies. Much of the church book data, which are already transcribed, can be used for the purpose of family reconstitution. That will provide a much more profound understanding of family patterns and household decisions in past societies.

The second road forward concerns the collection of more data. We know practically nothing about the demographic history of Africa, Asia, and the Americas. By collecting this information, it can be used to shed light on developments and fertility, mortality, life expectancy, literacy rates, occupational structures, gender inequality, social mobility, and their connection to economic development, notably in third-world countries.

References

Allen RC (2001) The great divergence in European wages and prices from the middle ages to the first world war. Explor Econ Hist 38:411–447

Armstrong A (1974) Stability and change in an english country town. A social study of York 1801–1851. Cambridge University Press, Cambridge

Basso A (2012) Essays in comparative economic development and growth. PhD dissertation, University of Alicante

Becker S, Cinnirella F, Woessmann L (2010) The trade-off between fertility and education: evidence from before the demographic transition. J Econ Growth 15:177–204

Bengtsson T, Dribe M (2006) Deliberate control in a natural fertility population: southern Sweden, 1766–1864. Demography 43:727–746

Boberg-Fazlic N, Sharp P, Weisdorf J (2011) Survival of the richest? Patterns of fertility and social mobility in England. Eur Rev Econ Hist 15:365–392

Bouchard G (1996) Tous les métiers du monde. Le traitement des données professionnelles en histoire sociale. Les presses de l'université de Laval, Saint-Nicolas

Carmichael SG (2011) Marriage and power: age at first marriage and spousal age gap in lesser developed countries. Hist Fam 16:416–436

Cinnirella F, Klemp M, Weisdorf J (2013) Malthus in the bedroom: birth spacing as a preventive check mechanism in pre-modern England. University of Warwick working paper no 174–2013

Clark G (2005) The condition of the working class in England, 1209–2004. J Polit Econ 113:1307–1340

Clark G (2007) A farewell to alms: a brief economic history of the world. Princeton University Press, Princeton

Clark G, Cummins N (2010) Malthus to modernity: England's first fertility transition, 1760–1800. MPRA working paper no 25465

Clark G, Hamilton G (2006) Survival of the richest in pre-industrial England. J Econ Hist 66:707–736

De Moor T, van Zanden JL (2010) Girl power: the European marriage pattern and labour markets in the North Sea region in the late medieval and early modern period. Econ Hist Rev 63:1–33

de Vries J (2008) The industrious revolution: consumer behavior and the household economy, 1650 to the present. Cambridge University Press, New York

Dribe M, Scalone F (2010) Detecting deliberate fertility control in pre-transitional populations: evidence from six German villages, 1766–1863. Eur J Popul 26:411–434

Fernihough A (2011) Human capital and the quantity-quality trade-off during the demographic transition: new evidence from Ireland. Working papers 201113 School of Economics, University College Dublin
Frankema EHP, van Waijenburg M (2012) Structural impediments to African growth? New evidence from real wages in British Africa, 1880–1960. J Econ Hist 72:895–926
Galloway PR (1994) Secular changes in the short-term preventive, positive, and temperature checks to population growth in Europe, 1460 to 1909. Clim Chang 26:3–63
Galor O (2011) Unified growth theory. Princeton University Press, Princeton
Galor O, Klemp M (2013) Be fruitful and multiply? Moderate fecundity and long-run reproductive success. Brown University discussion paper no 2013–2010
Galor O, Moav O (2002) Natural selection and the origin of economic growth. Q J Econ 117:1133–1191
Galor O, Weil DN (2000) Population, technology, and growth: from malthusian stagnation to the demographic transition and beyond. Am Econ Rev 90:806–828
Guzman R, Weisdorf J (2010) Product variety and the demand for children. Econ Lett 107:74–77
Hajnal J (1965) European marriage pattern in historical perspective. In: Glass DV, Eversley DEC (eds) Population in history. Edward Arnold, London, pp 101–143
Kelly M, O Grade C (2012) The preventive check in medieval and preindustrial England. J Econ Hist 72:1015–1035
Klemp M, Moeller NF (2013) Post-malthusian dynamics in pre-industrial Scandinavia. Brown University working paper no 2013–2014
Klemp M, Weisdorf J (2012a) The lasting damage to mortality of early-life adversity: evidence from the English famine of the late 1720s. Eur Rev Econ Hist 16:233–246
Klemp M, Weisdorf J (2012b) Fecundity, fertility, and family reconstitution data: the child quantity-quality trade-off revisited. CEPR discussion paper no 9121
Klemp M, Minns C, Wallis P, Weisdorf J (2013) Picking winners? The effect of birth order and migration on parental human capital investments in pre-modern England. Eur Rev Econ Hist 17:210–232
Lambert PS, Zijdeman RL, Van Leeuwen MHD, Maas I, Prandy K (2013) The construction of HISCAM: a stratification scale based on social interactions for historical comparative research. Hist Methods 46:77–89
Malthus TR (1798) An essay on the principle of population. J. Johnson, London
Meier zu Selhausen F (2014) Missionaries, marriage and power: dynamics and determinants of women's empowerment in colonial Uganda, 1880–1950, Utrecht University Mimeo
Meier zu Selhausen F, Weisdorf J (2014) European influences and gender inequality in Uganda: evidence from protestant marriage registers, 1895–2011, Utrecht University Mimeo
Meier zu Selhausen F; van Leeuwen MHD, Weisdorf J (2014) From farmers to clerks: social mobility in Uganda, 1895-201!, Utrecht University Mimeo
Midi Berry B, Schofield RS (1971) Age at baptism in pre-industrial England. Popul Stud 25:453–463
Moeller NF, Sharp P (2014) Malthus in cointegration space: evidence of a post-Malthusian pre-industrial England. J Econ Growth 19:105–140
Nicolini E (2007) Was Malthus right? A VAR analysis of economic and demographic interactions in pre-industrial England. Eur Rev Econ Hist 11:99–121
Perrin F (2013) Gender equality and economic growth in the long-run. A Cliometric analysis. PhD dissertation, University of Strasbourg
Razzell P (2007) Population and disease: transforming english society, 1550–1850. Caliban Books, London
Ricardo D (1817) On the principles of political economy and taxation. Cambridge University Press, Cambridge
Ruggles S (1999) The limitations of English family reconstitution: English population history from family reconstitution 1580–1837. Contin Chang 14:105–130

Schofield RS (1970) Perinatal mortality in Hawkshead, Lancashire, 1581–1710. Local Popul Stud 4:11–16
Schofield RS (1973) Dimensions of illiteracy, 1750–1850. Explor Econ Hist 10:437–454
Sharp P, Weisdorf J (2009) From preventive to permissive checks: the changing nature of the malthusian relationship between nuptiality and the price of provisions in the nineteenth century. Cliometrica 3:55–70
Souden D (1984) Movers and stayers in family reconstitution populations. Local Popul Hist 33:11–28
US Department of Labor (1939) The dictionary of occupational titles, 2 vols. US Department of Labor, Washington, DC
Van De Putte B, Miles A (2005) A social classification scheme for historical occupational data. Hist Methods 38:61–94
Van Leeuwen MHD, Maas I (2011) HISCLASS. A historical international social class scheme. Leuven University Press, Leuven
Van Leeuwen MHD, Maas I, Miles A (2002) HISCO: historical international standard classification of occupations. Leuven University Press, Leuven
Van Leeuwen MHD, Maas I, Weisdorf J (2014) Human capital from occupations: quantifying educational attainments in the past, Utrecht University Mimeo
Voigtländer N, Voth HJ (2013) How the west 'invented' fertility restrictions. Am Econ Rev 103:2227–2264
Voigtländer N, Voth HJ (2014) The three horsemen of riches: plague, war and urbanization in early modern Europe. Rev Econ Stud (forthcoming)
Wrigley EA (2010) 'The PST system of classifying occupations, University of Cambridge Mimeo
Wrigley EA, Schofield RS (1981) The population history of England 1541–1871. Edward Arnold, Cambridge
Wrigley EA, Davies R, Oeppen J, Schofield RS (1997) English population history from family reconstitution. Cambridge University Press, Cambridge

Part III

Growth

Cliometrics of Growth

Claude Diebolt and Faustine Perrin

Contents

Abstract

This chapter lays the theoretical foundations of long-run economic growth. After providing an overview of the three fundamental regimes that have characterized the process of development over the course of human history on the basis of the seminal work of Galor and Weil (2000), we review existing theories offering explanations of the different stages of development. In particular, we examine the predictions and underlying mechanisms of the traditional theories of economic growth and the theories of demographic transitions. We then show the relevance of the Unified Growth Theory to explain and capture the underlying mechanisms of the development process. Finally, we highlight the importance of integrating a gendered perspective in the study of long-run economic growth.

C. Diebolt (✉) · F. Perrin
BETA/CNRS, University of Strasbourg Institute for Advanced Study, Strasbourg, France
e-mail: cdiebolt@unistra.fr; faustine.perrin@unistra.fr

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_3

Keywords
Economic History · Economic Development · Growth · Demographic Transition · Unified Growth Theory · Gender

Introduction

The movement of the production potential of the industrialized nations over long periods of time is at the center of the very latest economic debates. This preoccupation is far from new. The classical economists were already concerned about how to increase welfare by increasing growth. The subject remained controversial after World War II with the theoretical debate on the long-term stability of market economies. However, through Solow's (1956) economic-growth model, neoclassical thinking gradually exerted its power. Its reasoning is clear, and it also explains numerous aspects related to economic growth, which are summarized perfectly in Kaldor's (1963) six "stylized facts." At the same time – perhaps paradoxically – scientific interest in work on growth and economic fluctuations disappeared. There were two main reasons for this: First, the short sightedness of economists whose attention was centered almost exclusively on the study of short-term movements and second, the comparative weakness of theoretical models unable to solve the aspects that remain unexplained by the different theories of growth. This partially explains why the postwar neoclassical models are unsatisfactory. Indeed, in the long run, they only account for economic growth by involving exogenous factors (except for Ramsey's (1928) model that was rediscovered very recently). In addition, Solow's reference model does not provide any way of explaining the divergence in growth rates at the international level. The theory of long-run equilibrium suggests that all countries should progress at identical, exogenous rates of technical progress. Similarly, it should be noted that the hypothesis of the systematic existence of a negative correlation between income level and economic growth rate is not based on any satisfactory empirical verification. Finally, nothing really corroborates the convergence hypothesis, that is to say, the transfer of capital from the richest to the poorest countries.

However, the work of Lucas (1988) and Romer (1986, 1990) attracted attention, and the 1980s marked a renaissance of the neoclassical theory of growth. The prime objective was to go beyond the weakness of the old theoretical models. The aim was also to answer new questions: What are the determinants of sustainable economic growth? Can technical progress alone increase social welfare or can capital accumulation also lead to a permanent increase in per capita income? What are the factors of production that engender sustainable economic growth: physical capital, environmental capital, human capital, or technological knowledge? What are the mechanisms that guarantee growth over a long period for a market economy? And finally, what is/are the market structure/s within which economic growth can be achieved? Strengthened by its focus on these questions, the debate on the determinants of the economic growth

process has recently attracted renewed attention, both in the importance of its implications in terms of economic policy and in the number of theoretical and empirical analyses that it engendered.

In fact, during the past two centuries, the Western world witnessed dramatic economic, demographic, and cultural upheavals. This period marked a turning point in historical economic and demographic trends. Despite some variations in terms of timing and speed of changes (Galor 2012), Western countries exhibited similar patterns of economic and demographic transition. Before the Industrial Revolution, all societies were characterized by a very long period of stagnation in per capita income with high fertility rates and the dominance of physical capital over human capital (Clark 2005). Since this fateful period Western countries experienced a complete reversal with high and sustained income per capita and low fertility rates (Becker et al. 2012; Klemp 2012). Human capital became an important source of income.

The main objective of this chapter is to present the theoretical approaches attached to the understanding of the process of development and growth. Empirical regularities raise numerous questions about the potential interactions linking demographic developments and the economic transition and about the role they have played in the transition from the stagnation to sustained growth. What are the underlying behavioral forces behind this demographic transition? What are the endogenous interactions between population and production? What accounts for the unprecedented rise in income per capita? Why has the transition to a state of sustained economic growth occurred together with the demographic transition?

This chapter that lays the theoretical foundations of cliometric analyses that aim at providing a better understanding of the long-run economic growth is organized as follows. First, we provide an overview of the stylized facts of three fundamental regimes that have characterized the process of development over the course of human history on the basis of the seminal work of Galor and Weil (2000).[1] Second, we explore existing theories offering explanations of the different stages of the process of development. We briefly examine the predictions and underlying mechanisms of the traditional theories of economic growth and development and the theories of demographic transitions. Third, we highlight the relevance of the unified growth theory to explain and capture the underlying mechanisms of the development process, and we provide an example of the unified growth model, introducing a key concept of development: the level of gender equality.

[1]The seminal work of Galor and Weil was quickly followed by new contributions, including Jones (2001), Lucas (2002), Hansen and Prescott (2002), Galor and Moav (2002), Doepke (2004), Galor (2005), Cervellati and Sunde (2005), Strulik and Weisdorf (2008), among others.

The Stylized Facts of the Development Process

Evolution of Output and Population Growth in Western Europe

Demographic behaviors are a key underlying aspect of the process of development that occurred in Western countries over the past 200 years. In order to have a better comprehension of the evolution of economic growth, demographic trends must be studied coincidentally with economic developments.

Figure 1 presents a broad picture of the joint evolution of output growth and population growth in Western Europe over six periods between 1000 and 2008. The first two periods, 1000–1500 and 1500–1700, are highly similar with a population growth rate slightly larger than the output growth rate (respectively, around 0.16% and 0.12%). Both average annual growth rates start to increase slowly over the period 1700–1820 (respectively, 0.41% and 0.15%). The wealth generated was absorbed by the rise in population growth. This positive relationship between income and population continues over the period 1820–1870 but becomes progressively narrower.

The period 1820–1870 experiences a sharp rise in economic growth. The takeoff in growth rates of GDP per capita was associated with a rise in population growth as observed in all regions of the world (Galor 2011). However, the population growth remains relatively restrained in comparison to the output increase. More precisely, the average growth rate in GDP per capita in Western Europe between 1820 and

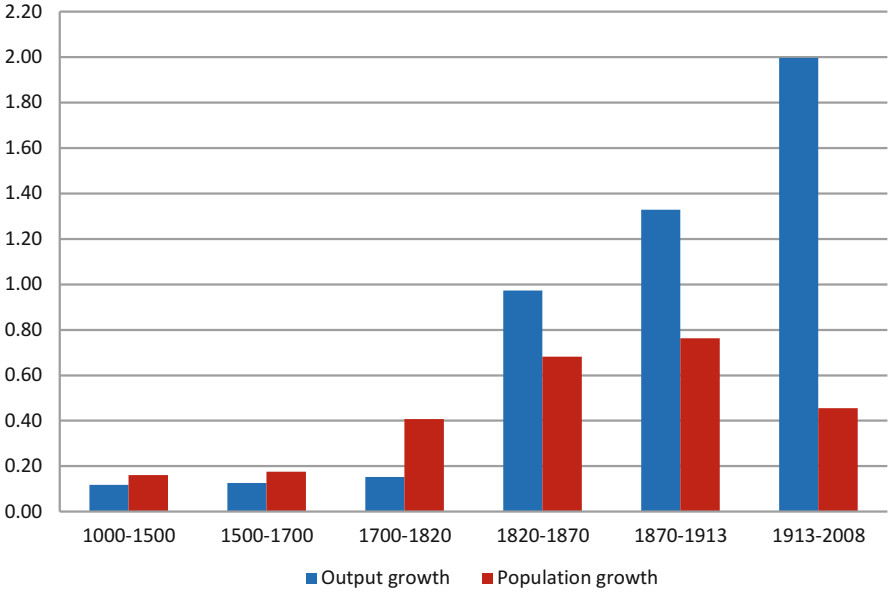


Fig. 1 GDP per capita and population growth rates in Western Europe (30 countries) (Source: Data from Maddison (2008))

1870 rose to an annual growth rate of 0.97% (from 0.15% during the period 1700–1820) while the average population growth rate increased to 0.68% (from 0.41% during the period 1700–1820). If we compare Western Europe with France, we note that population growth was significantly lower in France than in Western Europe, with an average annual growth rate of 0.42% over the same period. From 1870 to 1913 the pace of the population growth rate slowed down (0.42%) while that of the GDP per capita increased further to 1.11%. The last period, 1913–2008, is marked by an unprecedented reversal in the relationship between population and output growth. For the first time, the rate of population growth decreased while the growth rate of per capita GDP continued to rise. The rate of GDP per capita then grew by 2% per year while population growth rate declined to a yearly average of 0.45%. Ultimately, Western Europe experienced a demographic transition in parallel to the continuous increase in GDP per capita.

The Three Phases of the Development Process

Several important features stand out from Maddison's (2008) data. Human history can be divided into three fundamental regimes: the Malthusian Epoch, the Post-Malthusian Regime, and the Modern Growth Regime.

Stagnation: Malthusian Era

Maddison indicates that the average level of world per capita income fluctuated around $450 per year over the period 1–1000 and around $670 per year from then until the end of the eighteenth century. The monotonic increase in income per capita during the Malthusian era was associated with a uniform evolution of the average population growth rate (0.01% per year in the first millennium; 0.1% per year in the years 1000–1500; 0.27% per year over the period 1500–1820), keeping living standards fairly stable. The stagnation has characterized human history for thousands of years. At that stage, population growth was positively affected by the level of income per capita. The monotonic increase in income per capita during the Malthusian era was associated with uniform growth rate of the population, which did not result in variations in the standard of living (Galor 2011). The absence of significant changes in the level of technology trapped the income per capita around a subsistence level, and population size remained relatively stable.

Takeoff: Post-Malthusian Phase

At the beginning of the nineteenth century, Western countries experienced a takeoff from Malthusian stagnation. This shift took place with the increase in the pace of technological progress in association with the process of industrialization, presumably stimulated by the accumulation of human capital.[2] Based on Maddison (2008), we note that the world average growth rate of output per capita increased from 0.05%

[2]The demand for education increased from the end of the period.

per year for the period 1500–1820 to 0.54% per year during the period 1820–1870 and reached 1.3% per year in the years 1870–1913. Similarly, the average rate of population growth in the world increased from 0.27% per year in the period 1500–1820 to 0.4% per year in the years 1820–1870 and to 0.8% per year in the interval 1870–1913. Hence, we note that this period is still marked by a positive relation between income and population growth. The acceleration of technological progress resulted in a significant increase in the growth rate of output per capita, generating an unprecedented increase in population growth. The timing of the takeoff differs across regions. In less developed countries[3], the takeoff occurred progressively with a one-century delay, from the beginning of the twentieth century. The decline in population growth marked the end of the so-called Post-Malthusian Regime by the end of the nineteenth century in Western countries and by the second half of the century in less developed regions.

Sustained Growth: Modern Growth Regime

The acceleration of technological progress during the second phase of industrialization, its interaction with the human capital accumulation, and the reversal in the relation between income per capita and population growth marked the transition toward a state of sustained economic growth. The entrance in the Modern Growth Regime, associated with the phenomenon of demographic transition, has led to a great divergence in income per capita in Western countries over the past two centuries (Galor 2011).

Using Maddison's data, the reversal in the rate of population growth occurred by the end of the nineteenth century and the beginning of the twentieth century for particular regions of the world (Western Europe, Western Offshoots, and Eastern Europe). From an average of 0.77% per year in the period 1870–1913 in Western Europe, the population growth rate decreased to an average of 0.42% per year in the years 1913–1950, while it continued to grow in other parts of the world. At the same time, the world average growth rate of GDP per capita kept on increasing, reaching a peak of 2.82% per year between 1951 and 1973.

Although industrialization initiated the demographic transition in most Western countries by the late nineteenth century, the process started nearly a century earlier in France. Figure 2 makes a comparison of the ratios of output and population growth in France, the United Kingdom, and Western Europe over the period 1000–2008. After centuries of stability in the output-to-population growth rates,[4] there was a sudden and dramatic rise. France, UK, and Western Europe witnessed this unprecedented increase at the same time, namely, by the first decade of the nineteenth century.

[3]By less developed countries, we mean Latin America, Asia, and Africa.

[4]About 0.6 in France, 0.8 in UK, and 0.7 in Western Europe.

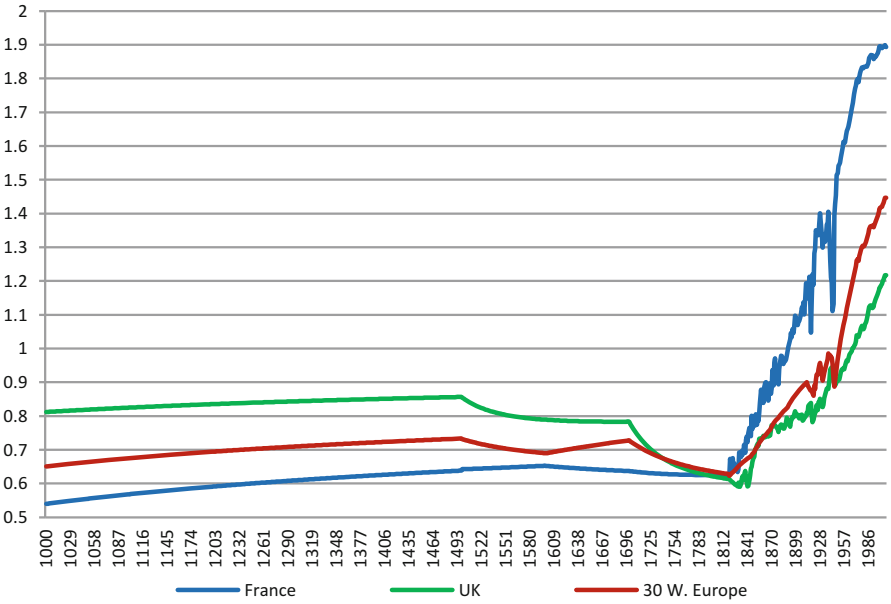


Fig. 2 Ratio of output-to-population growth rates in France, UK and Western Europe, 1000–2008 (Source: Using data from Maddison (2008))

However, while the ratio of population and output growth rates reached one in France in 1891, Western Europe reached this ratio in 1953 and the United Kingdom in 1968 only. The growth rate of GDP per capita relative to population growth has been much faster and intense in France than in the rest of Western Europe. Two main issues emerge from these findings. Why did population and output growth reverse at the same time in France and in other Western European countries? Why was the rise in the ratio between output and population growth so much faster in France than in the rest of Western Europe?

Main Challenges

As previously mentioned, the development process raises a number of questions and puzzles. This has piqued the interest of cliometricians specializing in the field of growth and development. Unprecedented upheavals occurred during this process. The demographic transition, the transition from stagnation to growth and the phenomenon of great divergence in income per capita, took place at different times across regions of the world. Many mysteries persist. Contemporary growth theorists, as well as cliometricians, need to improve their understanding of the development process and of the driving forces and underlying determinants that led to the escape from the Malthusian trap and allowed for the transition to sustained growth.

The main questions addressed (Galor 2005, 2011) are the following:

- What can explain the centuries of stagnation that characterized most of human history?
- What are the driving forces that account for the sudden increase in growth rates of GDP per capita and the persistent stagnation in others?
- What led to the Industrial Revolution? Why did this phenomenon occur first in Great Britain?
- What factors can account for the relationship between population and output growth? Why has the positive link between income and population growth reversed its course in some economies but not in others?
- What are the main forces that initiated the process of demographic transition? Why did this phenomenon occur first in France?
- What has caused the Great Divergence in income per capita across regions of the world over the last two centuries? Would this transition have been possible without the demographic transition?

In other words, what are the underlying behavioral and technological structures that could simultaneously account for these distinct phases of development? Additionally, what are their implications for the contemporary growth process of developed and underdeveloped countries?

Toward a Unified Theory of Growth: Theoretical Background

The fundamental challenge faced by cliometricians specializing in economic growth is to provide reliable answers to the previous set of questions using the contributions of economists, historians, and sociologists. The issue for growth theorists is to develop a unified theory of growth that can account for the main features of the three distinct phases that have characterized the process of development. This was first undertaken by Galor and Weil (1999, 2000), with the development of the unified growth theory.[5] This theory aims at giving a better understanding of the driving forces that triggered the escape from the Malthusian trap and the subsequent transition to a state of sustained growth.

Traditional Theories of Economic Growth

The theories and models of economic growth have evolved considerably over time. The theories of endogenous growth have emerged in response to the inability of

[5]The term was coined first by Galor (2005).

exogenous growth models to explain the origin of technological progress. These two types of modeling have themselves borrowed some basic elements from the classical theories of growth and stagnation.

The Malthusian Theory

The world of economic history has been dominated by the Malthusian stagnation. For a long time, theories aimed at explaining economic growth and development found their inspiration in Malthusian and neoclassical conceptions. In his *Essay on the Principle of Population*, Malthus (1798) defends a "pessimistic" vision of the impact of population growth on long-run economic development, coherent with the world economic history prior to the Industrial Revolution. Malthus's thinking can be summarized by the two following postulates: (i) Population growth is bounded by the means of subsistence and (ii) population increases with livelihoods in a geometric progression while production of food grows in arithmetic progression. The theory developed by Malthus matches the empirical evidence of the relation between income and population dynamics prior to the Industrial Revolution fairly well. According to this theory, the effect of population growth is counterbalanced by the expansion of resources, reflecting the fluctuations of the income per capita around a subsistence level. Malthus argued that two types of barriers contributed to reduce the size of the population at the subsistence level: the "positive checks" and the "preventive checks." The "positive checks" raise the death rate through hunger, disease, or war. The "preventive checks" affect birth rates through birth control, abortion, late age of marriage, or celibacy.

Without changes in the level of technology and resources, both the population size and the income per capita would remain stable. However, periods of technological progress and expansion of resources would lead to an increase in population growth, which ultimately triggered a decline in income per capita. Despite the capacity of the Malthusian theory to capture the characteristics of the epoch of stagnation, its predictions appear inconsistent with the features of the post-demographic transition era and the Modern Growth Regime. At the end of the nineteenth century, liberal economists such as Leroy-Beaulieu (1913) found that the theory was contradicted by the facts. He found that the movement of population was slowing down and output growth was accelerating. As a consequence, doctrines evolved toward the idea that population growth followed different rules than output growth. Boserup (1965, 1981) notably argued that the demographic pressure would lead to a reorganization of agricultural production. According to Boserup, the size of the population drives changes in the operating modes and not the subsistence level. Technological progress may then allow the subsistence level of production to consistently exceed population growth.

Classical economists such as Malthus (1798), Smith (1776), Ricardo (1817), and later Schumpeter (1934) have provided basic ingredients that appear in modern growth theories, such as the interplay between income per capita and the rate of population growth, the role of technological progress, and the accumulation of physical and human capital.

The Neoclassical Theory

Exogenous Growth Model

Contrary to the Malthusian theory that has investigated the relation between population and production prior to the demographic transition, neoclassical growth models focused largely on the growth process during the Modern Growth phase. Far from being limited to agricultural productivity, population growth is affected by complex socioeconomic-cultural phenomena related to the enrichment of society, culture, and choices of social organization that triggered families to limit their number of children. Growth models only gradually started to integrate these aspects.

In opposition to Malthus' approach, exogenous growth models, such as Solow (1956) and Swan (1956), deal with demographic growth as an exogenous variable and assume demographic behaviors to be independent of wages, incomes, and prices. Without technological progress, the income per capita converges toward a stable steady state independently of the size of the population. The Solow model is based on the assumption that the factors of production separately have diminishing returns. However, returns to scale are assumed to be constant, and factors of production are assumed to be used effectively by all countries. In an economy with more capital, the productivity of labor increases. As a consequence of diminishing returns to factors of production, economies will reach a point where any increase in production factors does not generate an increase in output per capita. In neoclassical growth models, the rate of long-run growth is determined by factors that remain unexplained (exogenous), such as the rate of technological progress in the Solow model.

New Home Economics

Parallel to the evolution of exogenous growth models, a branch of theoretical economic literature started to methodically analyze household decisions, such as consumption, savings, and labor supply. The lack of consideration of family behavior and its impact on economic models indeed led to the creation of a new stream of research, the so-called "New Home Economics."[6] The New Home Economics extended the domain of microeconomic analysis to a wide range of behaviors and human interaction, such as demographic behavior, investments in human capital, and intergenerational transfers. The (static) modeling of household production and time allocation was notably used to explain the sexual division of labor and the market behavior of household members. Among the first publications were Becker (1960) on fertility, Mincer (1962) on women's labor supply, and Becker (1965) on the allocation of time. Key assumptions of this literature are that institutions and cultures influence decisions in the home (Folbre 1994) and that these decisions are made by families as a unit. Manser and Brown (1980) have introduced household (two-sex) bargaining models, taking into account the separate interests of individual household

[6]Ironically, the etymology of "economics" is derived from the Greek *oîkos* (house, dwelling) and *nómos* (law, custom) and refers to the art of properly administrating one's home.

members. Their framework was then extended by authors such as Chiappori (1992) and Lundberg and Pollak (1993). A decade after the creation of the New Home Economics, Nerlove (1974), Razin and Ben-Zion (1975), and Srinivasan (1988) developed models linking demographic behaviors to macroeconomic evolutions in order to analyze their implications on the general equilibrium.[7]

The Endogenous Growth Theory

Endogenous growth models were developed in the 1980s as an extension to exogenous growth models in order to address the issue of the origin of technological progress – holding that economic growth results from endogenous (and not external) forces. The first endogenous growth model was published by Romer (1986) and was then extended by Lucas (1988), Romer (1990), and Barro (1990). These theories are constructed around the central idea that factor returns no longer decrease when it is accepted that components other than physical capital (such as human capital) exist and can display endogenous accumulation. Endogenous theorists identified four key factors of growth: returns to scale, research and innovation (Romer 1990; Grossman and Helpman 1991; Aghion and Howitt 1992), knowledge and human capital (Lucas 1988), and state intervention (Barro 1990). The structure of these models is identical. Endogenous growth becomes possible after the introduction of a new accumulation factor that compensates the decreasing returns of capital accumulation. According to Lucas, the source of economic growth lies in the unlimited accumulation of human capital. This boundless increase in human capital is based on major hypotheses of nondecreasing returns of technology and training and the existence of externalities. In the models in line with Romer (1990), economic growth is a function of research and development that depends on the share of human capital allocated to the research sector. The accumulation of knowledge (innovations) is the engine of growth. Other models achieve self-maintained growth through similar mechanisms by means of hypotheses concerning the nondecreasing returns of the new factors of accumulation.

The AK Model. The simplest version of endogenous growth models is the AK model. This formalization eliminates all the fixed factors that are not reproducible and therefore cannot be accumulated, thus making it possible to achieve endogenous growth in spite of the absence of increasing returns to scale or externalities. The essence of endogenous growth resides in the use of reproducible factors that can be accumulated. This central hypothesis makes it possible to affirm that capital returns are constant. The production function is then summarized by the following expression: $Y = AK$, where A is an exogenous scale parameter indicating the level of technology and K describes capital, including human capital, the stock of knowledge, and financial capital. Human capital is subject to accumulation and substitutes for the labor factor – which is by nature not reproducible. Capital is therefore a

[7]Within the framework of the neoclassical growth model with endogenous fertility, the authors attempt to determine the optimal population growth rate.

composite component incorporating all the accumulation factors. The nondecreasing returns allow self-maintained growth.

Family-Based Endogenous Growth Models. Inspired by the New Home Economics literature and by endogenous growth models, growth models with explicit microeconomic foundations of family have developed progressively (Barro and Becker 1989; Becker et al. 1990; Ehrlich and Lui 1991; Galor and Weil 1996; Dahan and Tsiddon 1998; Iyigun 2000). Growth theorists, exploring mechanisms by which fertility and growth are related, focused primarily on the modern era (Barro and Becker 1989; Barro and Sala-i-Martin 1997; Becker et al. 1990; Moav 2005; Tamura 1994, 1996). The so-called endogenous growth theory, taking into account family behavior (as a single decision-maker), is able to explain the empirical regularities that characterized the growth process of developed countries over the last 100 years. The pursued objective of these models is to provide a theoretical growth model with microeconomic foundations consistent with the stylized facts of the demographic transition.

The Theories of Demographic Transition

The demographic transition is identified as having played a key role in the process of development. From a theoretical point of view, different factors have been put forward to explain the process of demographic transition. Becker (1960) argued notably that the rise in per capita income had an effect on both households' income and opportunity cost of raising children. However, this explanation does not seem sufficient to fully explain the empirical regularities described previously. Why did demographic transitions occur simultaneously across countries that significantly differ in income per capita? Why did France experience its demographic transition prior to other countries?

The gradual rise in the demand for human capital along the process of industrialization has been seen by some researchers as a prime force leading to the onset of the demographic transition, specifically during the second phase of the Industrial Revolution. Taking family as a single decision-maker, Becker's models manage to generate the demographic transition but do not differentiate between the behaviors of males and females. Becker et al. (1990) model the relationship between human capital, fertility, and economic growth. In this "one-sex" model with altruistic parents, higher productivity leads to higher wages and favors human capital accumulation, which in turn raises the opportunity cost of children. This feature highlights the existence of two locally stable steady states: a Malthusian steady state with many children and little human capital and a steady state with few children and high human capital.[8] In the interpretation of the model, they consider changes in female labor force as implicit. Galor and Weil (1999, 2000) developed the idea that the acceleration in the rate of technological progress would gradually increase the

[8]Tamura (1994) finds the same result.

demand for human capital, inducing parents to invest in the quality of their offspring rather than in the quantity. The existence of a negative correlation between education and fertility has been demonstrated by Becker et al. (2012) with county-level evidence for Prussia in 1816. Ultimately, the process of human capital accumulation would induce a reduction in fertility rates as the growth rate of technological progress increases.

The decline in the gender gap is also considered a reinforcing mechanism impacting fertility rates. Galor and Weil (1996) investigate the relationship between fertility, gender gap in wages, and economic growth by explicitly assuming that men and women have different abilities and do different kinds of work. The authors postulate that technological progress and capital accumulation positively impact the relative wages of women along the process of industrialization, which increases the opportunity cost of raising children and ultimately leads to a reduction in fertility. Hence, economic growth would contribute to the closing of the gender gap in earnings, which would further lower fertility and reinforce economic growth. In a dynamic model with endogenous fertility, Iyigun and Walsh (2007) investigate how the evolution of spousal bargaining power within the couples' decision-making problem may trigger the decline in fertility.[9] Doepke and Tertilt (2009) study the opposite direction of causation. Based on a model with a quantity-quality trade-off on children, they investigate what economic forces may be at the origin of the progressive rise in women's rights throughout the process of industrialization. For Falcão and Soares (2008), it is the demographic transition that increases the supply of female labor and decreases the female-to-male wage gap. They show that gains in adult longevity increase the returns to human capital and reduce fertility. The subsequent decline in demand for household production (initially the specialization of women) increases the fraction of time spent by women in the labor market and reduces the gender earning gap. De La Croix and Vander Donckt (2010) employ the notion of intra-household bargaining power (called "welfare weight") and analyze how its variations may affect demographic and economic outcomes.

The progress of neoclassical growth models with endogenous fertility provides plausible explanations of the modern experience of economic growth in developed economies. Nonetheless, they do not provide a global understanding of the development process. They are unable to explain some of the most fundamental features of the process of development. They capture neither the recent negative relationship between population growth and income per capita nor the positive effect of income per capita on population growth and the economic factors that triggered the demographic transition. This left the door opened to a new generation of growth theorists (Galor and Weil 2000; Jones 2001; Galor and Moav 2002; Hansen and Prescott 2002; Doepke 2004; Strulik and Weisdorf 2008) to face the challenge of developing a theory consistent with the entire process of development.

[9]In this chapter, the authors do not focus on economic development and leave aside the question of how changes in gender heterogeneity may affect long-run growth.

The Unified Growth Theory

Unified growth theories are endogenous growth theories consistent with the whole process of development – accounting for empirical evidence that has characterized the growth process over longer time horizons in developed and less developed economies.

The Building Blocks of the Theory

Advanced first by Galor and Weil (1999, 2000) and developed by Galor (2005, 2010), the unified growth theory intends to capture, in a single framework, the main characteristics of the transition from the Malthusian era to the modern era, as well as the associated phenomenon of the Great Divergence and Demographic Transition.

The unified growth theory integrates the main features of the Malthusian economy in a context where the sizes of population and technology are linked. First, the increase in technological progress and the capital accumulation counterbalance the negative effect of population growth on income per capita highlighted by the Malthusian theory. As proposed by Galor and Weil (2000):

> . . .during the Malthusian epoch, the dynamical system would have to be characterized by a stable Malthusian steady-state equilibrium, but ultimately due to the evolution of latent state variables in this epoch, the Malthusian steady-state equilibrium would vanish endogenously leaving the arena to the gravitational forces of the emerging Modern Growth Regime.

Galor and Weil (1999, 2000) develop the idea that the acceleration in the rate of technological progress gradually increases the demand for human capital, inducing parents to invest in the quality of their offspring rather than in the quantity. Ultimately, the process of human capital accumulation induces a reduction in fertility rates in response to the increasing growth rate of technological progress. This leads to a demographic transition and sustained growth. The model, therefore, generates a transition from the Malthusian stagnation to the Modern Growth Regime. Later on, models incorporating new mechanisms emerge. Galor and Moav (2002) and Lagerlöf (2003) share similar intuitions by suggesting the existence of innate/inherited preferences in terms of children quality. Based on a unitary approach of the family, Lagerlöf (2003) explains how high-quality preferences may have spread over time and generate higher prosperity and lower fertility – considering changes in gender discrimination in education exogenous. In Cervellati and Sunde (2005), the authors introduce complementary mechanisms/channels based on the relations linking life expectancy, human capital, and technological progress. In a simple model, Strulik and Weisdorf (2008) provide a unified theory that captures the interplay between technological progress, mortality, fertility, and economic growth. Using a two-sector framework with agriculture and industry, the authors demonstrate how fertility responds differently to productivity and income growth between both sectors. Agricultural productivity and income growth make food, goods, and

therefore children, relatively less expensive, while industrial productivity and income growth, on the other hand, makes them relatively more expensive. Common to all these models (and to our model) is the central role played by the quantity-quality substitution in the phase transition. Empirically, the existence of a negative correlation between education and fertility has notably been demonstrated by Becker et al. (Becker et al. 2012) with county-level evidence from 1816 Prussia.

The unified growth theory generates the endogenous driving forces allowing the economy to experience a demographic transition that ultimately led to a takeoff from the era of stagnation toward a state of sustained economic growth. As highlighted in section "Introduction," Western countries experienced similar patterns of economic and demographic transition. This theory, which seems to be consistent with empirical regularities, is based on the interaction between four key elements: the building aspects of the Malthusian theory, the engines of technological progress, the origin of human capital accumulation, and the triggering forces of the demographic transition. The theory suggests that the acceleration in the pace of technological progress increased the importance of human capital. The rise in the demand for human capital and its impact on the accumulation of human capital led to a decline in fertility and to a rise in living standards.

However, one paradox persists. The French-English paradox (Chesnais 1992) raises a central question: why demographic development came so late in England and so early in France, while economic development was early in England and comparatively late in France. One underlying aspect of the development process may be missing.

Complementary Factors: The Role of Female Empowerment

Other central determinants of the development process have been left out of the first attempts at modeling a unified theory of growth. This left the door open to cliometricians and growth theorists to bring to light and explore additional and complementary mechanisms of the transition from stagnation to sustained growth. One such example is the issue of gender.

Gender-related issues have become central to the field of labor economics[10] and economic history (Goldin 2006). Empirical literature on the link between gender equality and economic development is rather abundant (Schultz 1995; Dollar and Gatti 1999; Klasen 2002; Knowles et al. 2002, among many others). However, the contributions remain rare in the field of economic growth. Few growth models explicitly consider the role played by gender on economic development: Galor and Weil (1996), based on the assumptions of different gender abilities; Lagerlöf (2003), taking gender differences as exogenous variables; or more recently De La Croix and Vander Donckt (2010), focusing especially on the pathways by which improvement

[10]Notably the pioneering work of Jacob Mincer (1962) that contributed to the development of economic analysis of the household.

in gender equality may affect fertility are among the few growth theorists who have integrated gender differentiation into their models.

Galor and Weil (1996) have engaged a first step toward a better integration of gender in growth theory by addressing the issue of the relationship between fertility, gender gap in wages, and economic growth with an inter-temporal dimension. Nevertheless, the model focuses on the modern era of economic growth and does not aim at providing a global framework of analysis for the evolution of economies over the entire course of human history. Lagerlöf (2003) sets up a model capturing gender stereotypes in which increasing gender equality can account for the important changes in growth rates of income per capita and population, in a unitary approach of the family. However, the model does not capture the notion of gender decisional empowerment, as noted by De La Croix and Vander Donckt (2010).

The role played by the rise in gender equality has been examined by Diebolt and Perrin (2013b). They argue that female empowerment has been at the origin of the demographic transition and engaged the takeoff to modern economic growth. More specifically, they develop a unified cliometric growth model capturing the interplay between fertility, technology, and income per capita in the transition from stagnation to sustained growth. The model suggests that gender empowerment is a crucial factor of both demographic and economic transition. In particular, the theory points out that the acceleration of skill-biased technological progress generates a positive externality on the level of gender equality. Both wages and gender equality are key variables in the education decision process of individuals. More specifically, higher gender equality reinforces individuals' incentives to acquire skilled human capital. In turn, female choices in terms of time and quality of educational investments increase their endowment in human capital and impact positively the fraction of the subsequent generation of individuals acquiring skilled education. In other words, improvements in technological progress, gender equality, and skilled human capital reinforce each other. Ultimately, the presence of a sufficiently high fraction of skilled individuals in the population yields to sustained economic growth. In the early stage of development, the low rate of technological progress does not provide any incentive to invest in skilled education. Therefore, the fraction of skilled individuals is low and the economy remains trapped in the Malthusian steady-state equilibrium, with low education, low standard of living, and low gender equality. Technological progress is assumed to increase monotonically from generation to generation. Therefore, as technological progress grows, we observe a qualitative change, and the subsequent income effect triggers (temporarily) higher fertility rates. After sufficiently many generations, increases in the returns from investments in skilled education (productivity growth) – driven by the rise in technological progress – makes investing in skilled education more profitable so that gender equality improves. The dynamic system of skilled human capital and gender equality is therefore characterized by multiple steady-state equilibria. Since gender equality becomes high enough, a substantially larger fraction of individuals acquire skilled human capital, which triggers rapid developments and reinforces gender equality. Due to larger educational investments (in terms of time units), the opportunity cost of having children increases and average fertility declines: The demographic transition occurs along

with the process of human capital accumulation. Ultimately, in later stages of development, gender equality and the fraction of skilled individuals converge toward their maximum. Thus, the economy is characterized by the Modern Growth steady-state equilibrium, where living standards are high, gender equality is high, and fertility is low.

Conclusion

The unified theory of growth has been developed as an alternative theory of exogenous and endogenous models that can capture the main characteristics of the process of development in a single framework. The unified growth theory sheds light on the driving forces that enable countries in a state of Malthusian stagnation to take off toward a state of sustained economic growth. In the Malthusian Regime, the economy remains trapped around a substantial level of output. During the Post-Malthusian Regime, the pace of technological progress accelerated under the effect of the increase in the population size and allowed economies to generate a takeoff. In the Modern Growth Regime, the output per capita increases along with the rate of population growth and human-capital accumulation (Galor and Weil 2000). Rapid technological progress, resulting from human capital accumulation, triggers a demographic transition with a constant decrease in fertility rates. The unified growth theory suggests that the transition from stagnation to sustained growth is an "inevitable by-product" (Galor 2011) of the process of development.

The purpose of future cliometric research in the growth theories area is to close the gap between *Geisteswissenschaften* and *Naturwissenschaften*, i.e., to move from the historical *verstehen*, or understanding, side to the economic *erklären*, or explaining, side. Even better, mixing both approaches, facts and stylized facts, may increase knowledge of the past, present and future economic and social development of developed and developing economies (Diebolt 2012; Diebolt and Perrin 2013a).

References

Aghion P, Howitt P (1992) A model of growth through creative destruction. Econometrica 60:323–351

Barro RJ (1990) Economic growth in a cross section of countries. Q J Econ 106(2):407–443

Barro RJ, Becker GS (1989) Fertility choice in a model of economic growth. Econometrica 57:481–501

Barro RJ, Sala-i-Martin Barro X (1997) Technological diffusion, convergence, and growth. J Econ Growth 2(1):1–26

Becker GS (1960) An economic analysis of fertility. In: Becker GS (ed) Demographic and economic change in developed countries. Princeton University Press, Princeton, pp 209–240

Becker GS (1965) A theory of the allocation of time. Econ J 75:493–517

Becker GS, Murphy KM, Tamura R (1990) Human capital, fertility, and economic growth. J Polit Econ 98:12–37

Becker SO, Cinnirella F, Woessmann L (2012) The effect of investment in children's education on fertility in 1816 Prussia. Cliometrica 6:29–44
Boserup E (1965) The conditions of economic growth. Aldine, Chicago
Boserup E (1981) Population and technological change. University of Chicago Press, Chicago
Cervellati M, Sunde U (2005) Human capital formation, life expectancy and the process of development. Am Econ Rev 95:1653–1672
Chesnais JC (1992) The demographic transition: stages, patterns, and economic implications. Clarendon, Oxford
Chiappori PA (1992) Collective labor supply and welfare. J Polit Econ 100:437–467
Clark G (2005) Human capital, fertility and industrial revolution. J Eur Econ Assoc 3(2–3):505–515
Dahan M, Tsiddon D (1998) Demographic transition, income distribution, and economic growth. J Econ Growth 3:29–52
De La Croix D, Vander Donckt M (2010) Would empowering women initiate the demographic transition in least-developed countries? J Hum Cap 4:85–129
Diebolt C (2012) The cliometric voice. Hist Econ Ideas 20(3):51–61
Diebolt C, Perrin F (2013a) From stagnation to sustained growth: the role of female empowerment. Am Econ Rev Pap Proc 103(3):545–549
Diebolt C, Perrin F (2013b) From stagnation to sustained growth: the role of female empowerment. AFC working paper, WP2013-4
Doepke M (2004) Accounting for fertility decline during the transition to growth. J Econ Growth 9:347–383
Doepke M, Tertilt M (2009) Women's liberation: what's in it for men? Q J Econ 124(4):1541–1591
Dollar D, Gatti R (1999) Gender inequality, income and growth: are good times good for women? Policy research report on gender and development working paper series, n 1. The World Bank, Washington, DC
Ehrlich I, Lui FT (1991) Inter-generational trade, longevity, and economic growth. J Polit Econ 99:1059–1129
Falcão BL, Soares RR (2008) The demographic transition and the sexual division of labor. J Polit Econ 116(6):1058–1104
Folbre N (1994) Children as public goods. Am Econ Rev 84(2):86–90
Galor O (2005) From stagnation to growth: unified growth theory. In: Aghion P, Durlauf SN (eds) Handbook of economic growth, vol 1A. North Holland, Amsterdam, pp 171–293
Galor O (2011) Unified growth theory. Princeton University Press, Princeton
Galor O (2012) The demographic transition: causes and consequences. Cliometrica 6:494–504
Galor O, Moav O (2002) Natural selection and the origin of economic growth. Q J Econ 117:1133–1191
Galor O, Weil DN (1996) The gender gap, fertility, and growth. Am Econ Rev 86:374–387
Galor O, Weil DN (1999) From Malthusian stagnation to modern growth. Am Econ Rev 89:150–154
Galor O, Weil DN (2000) Population, technology, and growth: from Malthusian stagnation to the demographic transition and beyond. Am Econ Rev 90:806–828
Goldin C (2006) The quiet revolution that transformed women's employment, education, and family. National Bureau of Economic Research, working paper no 11953
Grossman G, Helpman E (1991) Trade, knowledge spillovers, and growth. Eur Econ Rev 35(2):517–526
Hansen GD, Prescott EC (2002) Malthus to Solow. Am Econ Rev 92:1205–1217
Iyigun MF (2000) Timing of childbearing and economic growth. J Dev Econ 61:255–269
Iyigun MF, Walsh RP (2007) Endogenous gender power, household labor supply and the demographic transition. J Dev Econ 82:138–155

Jones CI (2001) Was an industrial revolution inevitable? Economic growth over the very long run. Adv Macroecon 1:1–43
Kaldor N (1963) Capital accumulation and economic growth. In: Lutz FA, Hague DC (eds) Proceedings of a conference held by the international economics association. Macmillan, London
Klasen S (2002) Low schooling for girls, slower growth for all? Cross-country evidence on the effect of gender equality in education on economic development. World Bank Econ Rev 16:345–373
Klemp M (2012) Price, wages and fertility in pre-industrial England. Cliometrica 6:63–78
Knowles S, Lorgelly PK, Owen PD (2002) Are education gender gaps a brake on economic development? Some cross-country empirical evidence. Oxford Econ Pap 54(1):118–149
Lagerlöf NP (2003) Gender equality and long-run growth. J Econ Growth 8:403–426
Leroy-Beaulieu P (1913) La question de la population. F. Alcan, Paris
Lucas RE (1988) On the mechanics of economic development. J Monet Econ 22:3–42
Lucas RE (2002) Lectures on economic growth. Harvard University Press, Cambridge, MA
Lundberg S, Pollak RA (1993) Separate spheres bargaining and the marriage market. J Polit Econ 101(6):988–1010
Maddison A (2008) Statistics on world population, GDP and per capita GDP, 1-2008 AD. http://www.ggdc.net/maddison/Maddison.htm
Malthus TR (First published 1798, this edition 1992) Essai sur le principe de population, 2 Vols., GF-Flammarion, Paris
Manser M, Brown M (1980) Marriage and household decision-making: a bargaining analysis. Int Econ Rev 21(1):31–44
Mincer J (1962) Labor force participation of married women: a study of labor supply. In: Lewis HG (ed) Aspects of labor economics. Princeton University Press, Princeton, pp 63–97
Moav O (2005) Cheap children and the persistence of poverty. Econ J 115(500):88–110
Nerlove M (1974) Toward a new theory of population and economic growth. J Polit Econ 84:200–216
Ramsey FP (1928) A mathematical theory of saving. Econ J 38(152):543–559
Razin A, Ben-Zion U (1975) An intergenerational model of population growth. Am Econ Rev 65:923–933
Ricardo D (First published 1817, English edition of 1821, this edition 1992) Des principes de l'économie politique et de l'impôt, GF-Flammarion, Paris
Romer P (1986) Increasing returns and long-run growth. J Polit Econ 94:1002–1037
Romer P (1990) Endogenous technological change. J Polit Econ 98:S71–S102
Schultz TP (1995) Investments in schooling and health of women and men: quantities and returns. In: Schultz TP (ed) Investment in women's human capital. University of Chicago Press, Chicago
Schumpeter JA (1934) The theory of economic development. Harvard University Press, Cambridge
Smith A (First published 1776, this edition 1991) Recherches sur la nature et les causes de la richesse des nations, 2 Vols., GF-Flammarion, Paris
Solow RM (1956) A contribution to the theory of economic growth. Q J Econ 70:65–94
Srinivasan TN (1988) Population growth and economic development. J Policy Model 10:7–28
Strulik H, Weisdorf J (2008) Population, food, and knowledge: a simple unified growth theory. J Econ Growth 13:195–216
Swan TW (1956) Economic growth and capital accumulation. Econ Rec 32(2):334–361
Tamura R (1994) Fertility, human capital and the wealth of families. Econ Theory 4:593–603
Tamura R (1996) From decay to growth: a demographic transition to economic growth. J Econ Dyn Control 20:1237–1261

Preindustrial Economic Growth, ca. 1270–1820

Alexandra M. de Pleijt and Jan Luiten van Zanden

Contents

Abstract

This chapter presents an overview of the current state of the art regarding the trends and causes of preindustrial economic growth in Western Europe between ca. 1270 and 1820. In doing so, the chapter introduces measures of living standards – i.e., real wages and per capita GDP. The Low Countries and England showed more or less stable real wages after the increase in wages in the fourteenth century following the Black Death, whereas real wages elsewhere in Europe declined in the long run. This "Little Divergence" between the North Sea area and the rest of the continent is also evident from estimates on per capita GDP. On the continent, per capita GDP stagnated or declined, whereas Holland and England showed a lot of progress – they were substantially richer in 1750 than in 1500. There were two key

A. M. de Pleijt
University of Oxford, Oxford, UK
e-mail: alexandra.depleijt@economics.ox.ac.uk

J. L. van Zanden (✉)
Department of History and Art History – Economic and Social History, Utrecht University, Utrecht, The Netherlands
e-mail: J.L.vanZanden@uu.nl

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_63

developments in this process. The first is the Black Death, which increased living standards in the countries bordering the North Sea regions due to well-functioning labor and capital markets. The second is the shift to premodern economic growth following the Black Death. The chapter summarizes the causes of the vast increases in living standards in the North Sea region between 1348 and 1820, including human capital formation and international trade.

Keywords
Preindustrial growth · Little divergence · Industrial revolution · North-Western Europe

Introduction

The British Industrial Revolution is arguably the most important break in world history. There is general agreement among economic historians on some aspects of the Industrial Revolution. For instance, the break that it represented was the appearance for the first time of continuous technological progress. However, its timing, location, and causes are a matter of debate. Can we explain the Industrial Revolution by looking at growth in the eighteenth century only, or is a much longer time-span required? Is growth a subnational/regional phenomenon, is the "nation"-state the appropriate unit of analysis, or is growth not limited by "national" borders? Were "efficient" institutions the main cause of the Industrial Revolution, or are there other important determinants of growth that need to be considered? These are highly relevant questions when studying the transition from an economy dominated by Malthusian forces to one characterized by "modern economic growth," as it supposedly occurred in Britain in the eighteenth century.

This chapter presents an overview of the current state of the art regarding the trends and causes of preindustrial economic growth in Western Europe between ca. 1270 and 1820. The first question that is dealt with concerns the location of preindustrial economic growth, for which the concept of the "Little Divergence" is relevant. A substantial body of evidence – starting with the real wage estimates of Allen (2001), and including the new generation of per capita GDP series (Bolt and van Zanden 2014; Fouquet and Broadberry 2015) – points to the fact that there was a divergence in levels of economic performance within Europe between 1500 and 1800. North-Western Europe, notably the Low Countries and England, showed more or less stable real wages after the increase in real wages in the fourteenth century following the Black Death, whereas real wages in Eastern and Southern Europe declined in the long run. In terms of per capita GDP, there is a similar divergence between the North Sea region and the rest of the continent: in the latter, per capita GDP stagnated or declined, whereas Holland and England showed a lot of economic progress: they were substantially richer in 1750 than in 1500 (Broadberry et al. 2015; van Zanden and van Leeuwen 2012).

The "Little Divergence" is also evident from other indicators of economic progress. The latest estimates on urbanization show that cities grew in the North Sea region, whereas city growth slowed down in the rest of Europe (Bosker et al.

2013). Similarly, the development of political institutions (measured by the activity of European Parliaments) declined almost everywhere in Europe between 1500 and 1800, except in the Netherlands and England (van Zanden et al. 2012b). Finally, estimates on per capita book production and rates of literacy and numeracy illustrate that the growth in human capital formation was faster in the North Sea region than elsewhere in Europe (Baten and van Zanden 2008; A'Hearn et al. 2009).

Economic growth between 1300 and 1800 was an international phenomenon originating in the region bordering the North Sea. Of this region, the Low Countries and England formed the core, but parts of Belgium, Northern France, North-West Germany, and Scotland were probably also experiencing growth. The North Sea region was strongly integrated. In the Late middle ages, England, for example, supplied the wool for the textile industry of the Low Countries, which was by then the main source of employment of the large Flemish cities. In the fourteenth–fifteenth centuries, Flanders formed the urban core of this economic system – and England its "periphery." In the sixteenth century, Antwerp took over the role of being the core. After 1585, the urban center moved to Holland, a switch that resulted in the Dutch "Golden Age" of the seventeenth century. After 1650 London gradually replaced Amsterdam as the central hub in the commercial network of North-Western Europe, and the urban core switched to England, as a result of which the Netherlands in the eighteenth century started to specialize in livestock products for the English market, thereby confirming its role as "new" periphery.

The second issue, concerning the beginning and periodization of the process of economic growth in the North Sea region, is related to the above. On the basis of the latest estimates of per capita GDP, a division into three periods can be suggested: before the Black Death ("classic Malthusian economy"), between 1347 and 1820 (slow but consistent growth that was limited to the regions bordering the North Sea region), and after ca. 1820 (rapid "modern" growth that started in England and from there spread to the rest of the continent). In explaining the transition from a Malthusian economy to modern growth, it is necessary to focus on two transformations: the Late Middle Ages, which prompted the "Little Divergence," and the early nineteenth century – the classic Industrial Revolution, which started in Britain and from there spread to the rest of Europe.

The third debate addressed in this chapter concerns the causes of preindustrial economic growth. There are numerous explanations. It has been characterized as "Smithian," as driven by the growth of trade and the increased division of labor made possible by it. The debate about its deep causes, however, focuses on the relevance of institutions – at the level of the state and/or at the level of the household – and the role of human capital formation in explaining the process.

Stylized Facts About Preindustrial Growth in Europe

There are two ways to measure economic performance in Europe in the long run. The first concerns the development of real wages. The classic paper is by Allen (2001), who estimated the real earnings of men in nine leading European cities from the Middle Ages up to the World War I. The second way to estimate levels of

economic development in the long run is per capita gross domestic product (henceforth GDP). This section briefly looks at the real wage series by Allen and the estimates of per capita GDP that have become available recently. Both indicators of preindustrial economic growth show that there was a "Little Divergence" in Europe between 1347 and 1800. Growth in this period was largely concentrated in the North Sea region, notably in Holland and England. The North Sea area was already dynamic during the Middle Ages, when Flanders became the urban center of North-West Europe. It continued to perform well in the one and a half centuries after the shock of the Black Death. In the sixteenth century, economic growth was likely concentrated in the southern parts of the Netherlands, and in the seventeenth century, the center moved to the province of Holland. In the second half of the seventeenth century/early seventeenth century, Britain took the lead. At the same time, living standards declined (e.g., Italy) or stagnated (e.g., Spain) elsewhere in Europe.

Real Wages

The divergence between the North Sea region and the rest of Western Europe is first of all clear from evidence on real wages. Allen (2001) developed a method to estimate the real wages of unskilled and skilled craftsmen (building workers) for nine leading European cities between 1400 and 1914. Figure 1 shows his results for six European countries. In the fifteenth century, real wages were relatively high across Europe, which can be attributed to the Black Death creating labor shortages.

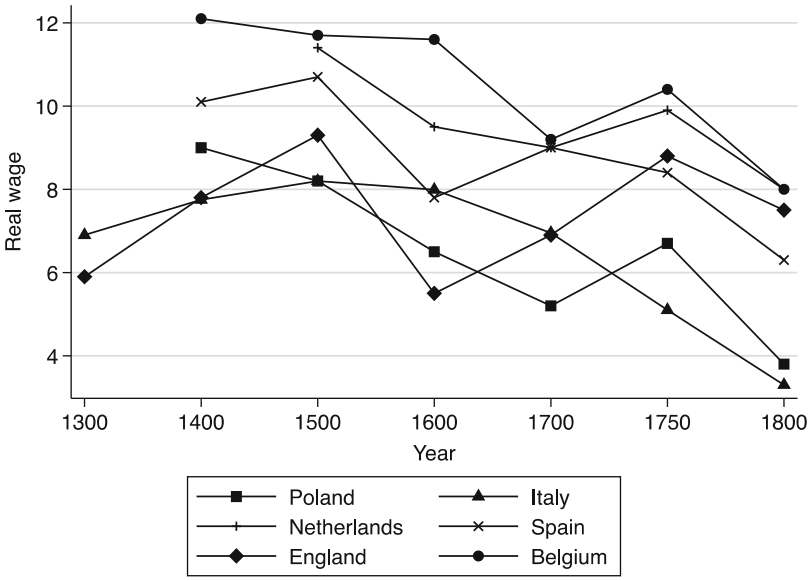


Fig. 1 Real wages in Western Europe, 1300–1800. (Source: Allen 2003)

Thereafter, however, there is long-term stabilization of living standards in the North Sea region (Belgium, the Netherlands, and England) in the centuries following the Black Death. Elsewhere in Europe, a long-term decline began and lasted into the late eighteenth century. According to Allen (2001), the Little Divergence started in the seventeenth century when the series for England (i.e., London) starts to show an upward trend. Pamuk (2006) and van Zanden (2009) date it earlier. The difference between the wages in the North Sea area and the rest of Europe already began to emerge during the second half of the fifteenth century.

The wage series by Allen (2001) has been challenged on several grounds. The real wages have been calculated by taking day wages and multiplying them by an assumed number of days worked per year (250) to determine annual income. This is then compared to the budget needed for a family of four people (man, woman, and two children) to derive the purchasing power of those wages over time and places. There are two versions of this basket. Initially, Allen reconstructed a respectability basket, close to what he assumed to be the actual budget in early modern Europe. When this basket was used in comparison with other regions of the world – such as Japan and China – it proved to be too generous, and a more austere barebones basket was constructed, which contained the basic essentials to survive (Allen et al. 2011). The problem with any standard of living is of course that it does not really fit the historical reality of all societies concerned – and perhaps, because it is some kind of average, no society in particular. This is inevitable in any attempt to make large international comparisons. It has been argued that the number of days worked per year (250) is too high (Hatcher 2011; Stephenson 2018b), or that the wage data used are not representative (Malanima 2011; Stephenson 2018a). Moreover, the wages used only capture the male population, and less is known about how much females worked and earned in the past. In a pioneering study, Humphries and Weisdorf (2015) have estimated the real earnings of women in England, and currently, attempts are being made to extend this research to Western Europe as a whole (De Pleijt and van Zanden 2018).

Per Capita GDP

Another way to measure economic performance in the past concerns per capita GDP. Economic historians since the rise of "New Economic History" have worked on a grand project to reconstruct the long-term evolution of the various parts of the world economy in the past millennium. Simon Kuznets in the 1950s and 1960s (building on work by, among others, Colin Clark) created the first international comparative studies to chart the development of GDP and its components in the long run (Clark 1940; Kuznets 1966). This work was continued by many economic historians. Angus Maddison became the central "hub" of this research, publishing various syntheses of the work in historical national accounting that was undertaken globally (Maddison 2001). This work is now carried out by the Maddison Project – a collaboration of the specialists in the field, coordinated by scholars from the University of Groningen (Maddison's home base). It synthesizes the research by many

economic historians who have combined data on population, employment structure, the rate of urbanization, nominal wages, prices, real wages, output levels in different industries, and many other sources to estimate GDP and its components for historical societies. Sometimes this is based on very rich historical sources – near-complete censuses such as the *Domesday Book* for post-Conquest England, the Catasto for Tuscany in 1427, or the Informacie for Holland in 1514. In other cases, indirect proxies have to be used to estimate the evolution of real income via, for example, real wages, which determine the demand for foodstuffs. Because different definitions of GDP via the output, demand, and income approach have to result in identical levels of real output and income, a lot of cross-checking of the results is possible, making the estimates more robust.

The Maddison project has resulted in a systematic comparison of estimates of GDP per capita, often going back to the (European) Middle Ages, which allows us to shed new light on the transition to modern economic growth. For England, the work by Broadberry et al. (2015) has produced annual estimates of GDP and its components from the 1270s up to the 1870s, as well as one point estimate for the year of the *Domesday Book* (1086). Likewise, van Zanden and van Leeuwen (2012) have collected comparable data for Holland between 1347 and 1807. Similar projects have been carried out for Italy by Malanima (2011), for Spain by Alvarez-Nogal and Prados de la Escosura (2013), for France by Ridolfi (2016), and for Sweden by Krantz (2017) and Schön and Krantz (2015). What is new about this research is that it has produced annual estimates of GDP per capita from the fourteenth century onward, making it possible to analyze the underlying trends more systematically than the "older" research that usually resulted in a number of benchmark estimates. We can now confidently observe breaks in growth and long-term performance trends.

As Fig. 2 shows, the "Little Divergence" is also very clear from the evidence on GDP per capita. The growth in real GDP per capita between 1300 and 1800 was largely restricted to the countries bordering the North Sea: Flanders, Holland, and England. In 1750, just before the start of the Industrial Revolution, the level of GDP per capita in Holland and England had increased to 2355 and 1666 (international) dollars of 1990, respectively, compared to 876 and 919 dollar in 1347 (just before the arrival of the Black Death) and 1454 and 1134 in 1500. Contrary to what happened in the North Sea region, there was no economic growth in Southern and Central Europe before the nineteenth century (Bolt and van Zanden 2014; Fouquet and Broadberry 2015).

The origins of the Little Divergence are of special interest because after the exogenous shock of the Black Death, real incomes in Britain and the Low Countries did not return to pre-plague levels, but remained much higher than before (ca. 20–40%), and, especially in the case of Holland, began to show a consistent rate of growth, which resulted in a doubling of GDP per capita over the next 250 years (not taking into account the effect of the Black Death). As Fig. 2 shows, the English growth experience was somewhat different. Between 1400 and 1600, real incomes fluctuated around a plateau of about 1000 dollars (against 700–800 dollars before 1347), and only in the second half of the seventeenth century did

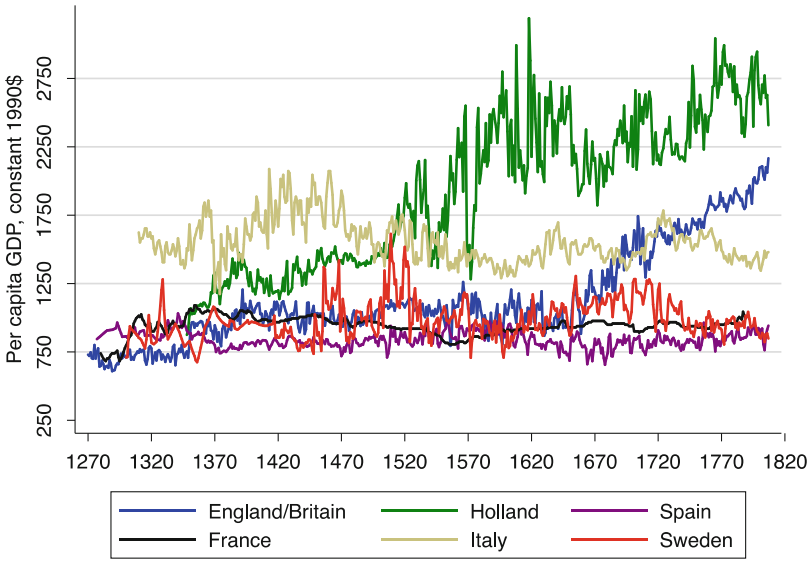


Fig. 2 Per capita GDP in Western Europe, 1270–1807. (Sources: Broadberry et al. 2015; Van Zanden and Van Leeuwen 2012; Malanima 2011; Alvarez-Nogal and Prados de la Escosura 2013; Ridolfi 2016; Krantz 2017; Schön and Krantz 2015)

growth really take off. Less is known about growth in Belgium, but the few point estimates that have been made by Buyst (2011) suggest that it followed a third pattern. Its level was fairly high at the start of the sixteenth century, there was some per capita income growth until the 1560s–1570s, but the late sixteenth and seventeenth centuries saw a decline of GDP per capita. In other parts of Europe (e.g., Spain), the collapse of population did not have similar positive effects on real incomes, which, as will be discussed below, tells us something about the quality of institutions in the North Sea region.

The per capita GDP series clearly demonstrate that the Black Death led to positive income growth in Holland and England/Britain. However, as Fig. 2 shows, a related phenomenon concerns the dynamics within the North Sea region itself. The late sixteenth and early seventeenth centuries saw a decline of GDP per capita in Belgium due to the loss of industrial and tertiary activities to the North of the Netherlands. The decline of per capita GDP in Belgium and the rise of the Netherlands were part of the same process, in which the urban core of the North Sea region moved from Antwerp to Amsterdam. The next shift in the urban system, from Amsterdam to London, started after about 1670 and had similar consequences – i.e., England started to grow, whereas Holland stagnated at a fairly high level.

Therefore, each country knew its separate growth cycles, characterized by a period of rapid growth followed by (and/or preceded by) long periods of stagnation. Flanders boomed in the Late Middle Ages, but declined after about 1560; Holland had its most spectacular growth spurt during its Golden Age (1585–1670), but slowed down in the eighteenth century; and, finally, English real incomes rose

rapidly after 1670. Sweden can be interpreted as perhaps demonstrating what happened to a region that (partly) belonged to the North Sea area but did not at any moment become the industrial and commercial core of the region.

Explanations for Preindustrial Economic Growth

Why did the Black Death led to an increase in per capita GDP in the North Sea region but not elsewhere in Europe? And what accounts for the phase of pre-industrial economic growth in the Low Countries and England between 1347 and 1800? This section gives a brief overview of the current state of the art.

The Black Death

To consider the effect of the Black Death, Fig. 3 gives a brief look at the "Malthusian" economy in the period before 1347. This series of GDP per capita reconstructed by Broadberry et al. (2015) shows no growth before 1348. The trend is slightly negative, mainly due to serious crises in the 1280s and the 1310s, and population growth seems to matter a lot. The population increases from 4.2 million in 1250 to 4.8 million in 1290, when GDP per capita registers a declining trend; after the crisis of the 1280s, the population starts to fall (to about 4.6 million in 1340), which results in a stabilization of real income, albeit with huge fluctuations, such as the "Great Famine" of the 1320s. In short, this economy shows distinct Malthusian

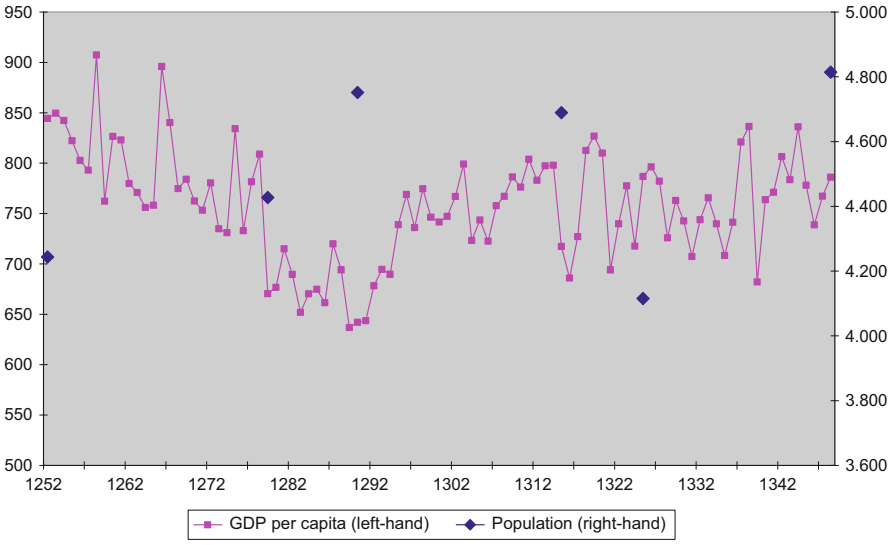


Fig. 3 English per capita GDP and population growth, 1252–1347. (Sources: Broadberry et al. 2015)

features, with a decline of real GDP during the (last stage of the) great Medieval boom (1080–1280), during which the population increased rapidly, and stability after the 1280s because population growth came to a halt. The level of GDP per capita is, however, clearly above subsistence levels, especially at the start of this century. During the 1250s, the English GDP per head of 800–850 dollars is between at least twice and almost thrice subsistence level (of 300–350 dollars) and quite high by international standards – only the more wealthy ancient economies reached similar levels, if one believes the recent estimates of pre-1000 GDP (Bolt and Van Zanden 2014).

The shock of the Black Death changed everything. As shown in the previous section, the average level of per capita GDP in Holland and England was significantly higher after the Black Death than in the period before. In England, it had increased from an average of 714 dollars before the arrival of the Black Death to 993 dollars in the period 1348–1500. Something similar happened in Holland. Van Zanden and van Leeuwen (2012) only have one point-estimate for 1347, but this suggests that it had increased from 876 dollars before the Black Death to 1301 dollars thereafter. In Spain, on the contrary, per capita GDP hardly responded to the decline in population: it fell from 900 dollars between 1270 and 1347 to 816 dollars between 1347 and 1500. The Black Death increased real incomes in Italy, but there, it gradually fell back to its pre-plague level over the course of the fifteenth century.

Why did the North Sea area respond so strongly to the population decline due to the Black Death, resulting in an immediate and sizable increase in GDP per capita? The population of Europe declined by about one-third to one-half, which led to major changes in relative prices. Wages went up, capital and land became relatively cheap, which, standard economic theory predicts, impacted the relative mix of inputs used. Labor productivity increased due to the larger availability of land and capital, and GDP per capita grew proportionally. This is arguably how a fully developed market economy would respond to such a crisis. That the North Sea area reacted in this way, therefore, meant that it behaved as a market economy. In other words, factor markets (markets for labor, land, and capital) and product markets were apparently developed to such an extent that the Black Death had such an impact on GDP (van Bavel and van Zanden 2004).

The point that the North Sea region in the Late Middle Ages was already characterized by highly developed factor markets has been made in the literature. A large part of the population – perhaps as much as 40–50% – was dependent on wage labor, both in the countryside and in the cities (Dyer 1989). Perhaps the rural labor market was even more developed than those in the cities (Van Bavel 2006). Both men and women were earning a wage income, and the gender wage gap was relatively low (women may have earned about 80% of the wages of men), stimulating the labor market participation of women (Van Zanden 2011; Humphries and Weisdorf 2015). Capital markets were also well developed, in particular in the Low Countries, with (after 1348 in particular) low interest rates (5–6% annually) and large-scale participation again by both men and women (Van Zanden et al. 2012a). Finally, on land markets, the leasing of land became increasingly popular, further adding to the commercialization of agriculture. Summing up, the fact that the

economies of the North Sea area responded to the shock of the Black Death in the way that it did appears to be rooted in an institutional structure with "thick" markets, which were used intensively.

The real issue is, arguably, why the other European countries did not respond in a similar way. The simple answer is that there were no comparable market economies. Alvarez-Nogal and Prados de la Escosura (2013) have addressed this issue for the case of Spain and conclude that the decline of GDP per capita after 1348 was due to the "frontier" character of the Spanish economy. The marketing system that had developed before 1347 could not be maintained anymore by the sharply reduced population. As a result, markets contracted, commercialization declined, and GDP per capita decreased. For Sweden, Krantz (2017) suggested a similar explanation of the absence of a Black Death bonus.

However, the marginalization of markets in a region with low population densities cannot be the explanation for the response of the Italian economy. According to the estimates by Malanima (2011), GDP per capita of Northern Italy (which is actually Tuscany in this period, as the data are from this region) initially increased by about 20–30% in the two decades following 1348 and then suddenly returned to the pre-1348 level for the rest of the fifteenth century. Larry Epstein (1991, 2001) has argued that institutional constraints related to the different relationship between city and countryside help to explain the relative stagnation of the fifteenth-century Tuscan economy. He argued that Florence was exploiting the Contado via taxation and other institutions that limited market participation in the countryside. Recent research seems to corroborate this view: the income gap between town and countryside was extremely high in Tuscany, and labor and land markets were relatively underdeveloped (Van Zanden and Felice 2017). In the countryside, the system of mezzadria, which spread rapidly in the fifteenth century, reflected those constraints, but it is equally striking that the rural labor market was extremely thin.

Summing up, the different responses to the Black Death were the consequence of different institutional settings and economic systems. Other studies that compared the impact of the Black Death on Late Medieval Egypt and England (Borsch 2005) or Central Europe and Western Europe/England (Brenner 1989) arrive at similar conclusions. The economic reaction of the North Sea area, where the well-developed market economy made possible a transition to a much higher level of GDP per capita, was probably exceptional. The deep roots of this peculiar development are probably not found in its specific political institutions, as these did not differ much between the North Sea area and the rest of Western Europe. The whole of Western Europe knew "feudal" sociopolitical relations, in which the power of kings, princes, and dukes was constrained by independent cities (strong in Northern Italy but also in the Low Countries) and by the integrative power of the Church (strong in all parts of the Latin West). The civil society of guilds, brotherhoods, and other religious and nonreligious institutions was not fundamentally different in the North from the South. An institutional "Little Divergence" between the North Sea region and the rest of Western Europe would only emerge in the sixteenth or even seventeenth century, when Parliaments acquired a much stronger position in the North-Western parts of Europe than in the rest of Europe, where absolutism arose (van Zanden et al.

2012b). Although political institutions may have enhanced the phases of growth after the sixteenth century, they do not seem to have made the difference in the centuries following the Black Death (de Pleijt and van Zanden 2016).

Explanations for the "Little Divergence"

Even more important than the response to the exogenous shock of the Black Death was the fact that it set in motion a process of preindustrial economic growth in Holland and England that persisted until the end of the eighteenth century. It is, as far as we know, the first time in world history that a region went through such an enduring process of sustained growth. Why were the Low Countries and England, already long before 1800, able to break through Malthusian constraints and generate a process of almost continuous economic growth? Various hypotheses have been suggested: institutional change (two versions: sociopolitical institutions such as Parliaments, demographic institutions such as the European Marriage Pattern); the impact of the growth of overseas – in particular – transatlantic trade (Allen 2003; Acemoglu et al. 2005); and the effect of human capital formation (Baten and Van Zanden 2008). Within the context of this paper, we can only briefly review some of the literature.

Premodern economic growth has been characterized as Smithian, after Adam Smith, who was the first to systematically analyze the process. He stressed the role played by the growth of international trade, urbanization, and specialization as the main drivers of the process. By contrast, growth after (and during) the Industrial Revolution is in this view primarily driven by technological change, the result of new ideas, and the application of science to production processes. In the Smithian approach, processes of structural change clearly played a large role (Broadberry et al. 2013). The North Sea region as a whole managed to capture an increasingly large share of international services and manufacturing output with high levels of value added. By 1300, only Flanders was highly urbanized – thanks to the combination of a highly specialized woollen textile industry exporting to large parts of Europe and international services mainly supplied by the commercial hub of Bruges. The ups and downs within the North Sea area in terms of GDP per capita are mainly linked to the spatial changes in the concentration of these high value-added activities – from Flanders to Antwerp to Amsterdam to London, respectively. At the same time, the relative size of these activities was growing, as the international markets for which they were catering expanded, until, in the seventeenth and eighteenth century, large parts of the world economy were linked to the central hubs in Amsterdam and London. The fact that other regions of Europe experienced GDP per capita decline – in particular, Northern Italy, but also Poland (Malinowski and Van Zanden 2017) – can be linked to the concentration of these activities in North-Western Europe, at the expense of Tuscany and Venice.

Another influential body of literature argues that it is the specific political economy of Western Europe, and in particular the balance of power between sovereigns and societal interests represented in Parliaments, that created the right

institutional conditions for Europe's specific growth pattern. Two versions of this hypothesis can be distinguished. The first one stresses the Glorious Revolution as the watershed between "absolutism" and some form of "parliamentary" government and sees this event as the main cause of the Industrial Revolution of the eighteenth century (North and Weingast 1989; Acemoglu and Robinson 2012). The other one argues that these institutions that resurfaced in 1688 had a much longer history and that forms of power sharing between the Prince and his (organized) subjects go back to the Middle Ages and are rooted in the feudal power structures of that period (Van Zanden et al. 2012b). The general idea shared by this literature is that the sovereign had to be constrained in order to protect the property rights of citizens. In republican systems with a strong Parliament, property rights were more secure than in states ruled by absolutist kings. This translated into, for example, lower interest rates at the capital market (Hoffman and Norberg 1994). Clark (1996) contests this view, arguing that there is little evidence of significant insecurity among private owners before 1688.

Mokyr (2002, 2009) does not refute the importance of efficient institutions for economic growth. However, he argues that the presence of efficient institutions alone was not enough for an economic takeoff to take hold: the Industrial Revolution was the result of an interaction between favorable institutions and the arrival of a new set of "ideas and beliefs." The Scientific Revolution of the seventeenth century produced "useful knowledge," such as knowledge about mathematics, that laid the foundation for "Industrial Enlightenment." This Industrial Enlightenment, which he defines as "the application of scientific and experimental methods to the study of technology" (Mokyr 2009, p. 29), connected the Scientific Revolution to the waves of technological innovations after 1760. In other words, the takeoff of England depended on what people knew and believed, which in turn affected their economic behavior.

An equally influential body of literature suggests that the root causes of "modern economic growth" should be found in an interplay of demographic and economic changes, affecting the "quality-quantity" trade-off (Becker 1981; Galor 2011) and resulting, on the one hand, in limitations on fertility and population growth and increased human capital formation on the other. The emergence of the European Marriage Pattern (EMP) in the North Sea area in the Late Middle Ages has been hypothesized as the crucial demographic change, which also resulted in increased investment in education of the fewer children from smaller families (De Moor and Van Zanden 2010; Voigtländer and Voth 2013). An important part of the mechanism was the increase in the average age of marriage of women (and men), which both limited fertility and increased opportunities for human capital formation.

The explanation focusing on the role of the EMP has, however, not been accepted generally. The data collected by Humphries and Weisdorf (2015) shows that wages for unmarried servants were developing less favorably than those of married women who worked for day wages in the century after the Black Death. From this they have concluded that it was not rational for women to postpone marriage. Dennison and Ogilvie (2014) were even more outspoken in their criticism of the European

Marriage Pattern, as they did not find a link between marriage patterns and economic performance in Early Modern Europe. Subsequently, however, Dennison and Ogilvie's results have been challenged. Carmichael et al. (2016) have argued that they did not conceptualize the European Marriage Pattern correctly. The focus of Dennison and Ogilvie was on the share of singles, the age of marriage of females, and the share of nuclear families, whereas attention should be on the broader context of how marriage responds to economic circumstances (Carmichael et al. 2016; Dennison and Ogilvie 2016).

Unified Growth Theory and related literature points to the importance of human capital formation for economic growth in general and the transition from Malthusian stagnation to Kuznetsian "modern economic growth" in particular (Galor 2011). From the Late Middle Ages onward, Western Europe saw a strong rise of literacy, numeracy, and schooling in general, which has also been attributed to social and religious changes, such as the Modern Devotion in the Low Countries (Akcomak et al. 2016) and the Reformation (Becker and Woessmann 2009). For England, a "first educational revolution" has been documented, which began in the fifteenth century and continued until the middle of the seventeenth century (Hoeppner Moran 1985; Orme 2006). But industrialization as such was not clearly related to rising literacy, as the English case demonstrates. There is increased evidence that in regions where the first Industrial Revolution was concentrated (in North-Western England), levels of literacy declined during the first stages of the process (Stephens 1987). In England, the eighteenth century was more generally a period in which levels of schooling stagnated, albeit at a relatively high level by international standards. This suggests that human capital formation may have played a role in creating the preconditions for industrialization but that this process itself was not driven by growing literacy and numeracy. Mokyr (2005) and Meisenzahl and Mokyr (2012) suggest that in this stage of industrialization, it was not the general level of education that mattered, but the skills of a small elite of educated and highly trained engineers and craftsmen, which played a key role in the wave of innovations in the eighteenth century.

Conclusion

This chapter reviewed the process of preindustrial growth, which occurred in the North Sea area between the Black Death and the onset of industrialization in eighteenth-century England. The latter process resulted in an acceleration of economic growth at about 1820, implying that the slow but sustained growth of GDP per capita that was characteristic of preindustrial growth continued between 1347 and 1820. The debate about the exact causes of the process is still going on, but institutional explanations, related to the organization of the state and the household, play a large role in it. Human capital formation, useful knowledge, the Enlightenment, and, last but not least, international trade also figure prominently in this debate about the "preconditions" of the Industrial Revolution.

Cross-References

- ▶ Economic-Demographic Interactions in the European Long Run Growth
- ▶ Human Capital
- ▶ The Industrial Revolution: A Cliometric Perspective

References

A'Hearn B, Crayen D, Baten J (2009) Quantifying quantitative literacy: age heaping and the history of human capital. J Econ Hist 68(3):783–808

Acemoglu A, Robinson J (2012) Why nations fail? The origins of power, prosperity and poverty. Crown, New York

Acemoglu D, Robinson JA, Johnson S (2005) The rise of Europe: Atlantic trade, institutional change, and economic growth. Am Econ Rev 95(3):546–579

Akcomak IS, Webbink D, ter Weel B (2016) Why did the Netherlands develop so early? The legacy of the brethren of the common life. Econ J 126:821–860

Allen RC (2001) The great divergence in European wages and prices from the middle ages to the first world war. Explor Econ Hist 38(4):411–447

Allen RC (2003) Progress and poverty in early modern Europe. Econ Hist Rev LVI(3):403–443

Allen RC, Bassino JP, Ma D, Moll-Murata C, van Zanden JL (2011) Wages, prices, and living standards in China, 1738–1925: in comparison with Europe, Japan, and India. Econ Hist Rev 64:8–38

Alvarez-Nogal C, Prados de la Escosura L (2013) The rise and fall of Spain (1270–1850). Econ Hist Rev 66(1):1–37

Baten J, van Zanden JL (2008) Book production and the onset of modern economic growth. J Econ Growth 13(3):217–235

Becker G (1981) A Treatise on the Family. Harvard University Press. Cambridge, MA

Becker SO, Woessmann L (2009) Was weber wrong? A human capital theory of Protestant economic history. Q J Econ 124:531–596

Bolt J, van Zanden JL (2014) The Maddison project: collaborative research on historical national accounts. Econ Hist Rev 67(3):627–651

Borsch SJ (2005) The black death in Egypt and England. University of Texas Press, Austin

Bosker M, Buringh E, van Zanden JL (2013) From Baghdad to London: unraveling urban development in Europe, the Middle East, and North Africa, 800–1800. Rev Econ Stat 95(4):1418–1437

Brenner R (1989) Economic backwardness in Eastern Europe in light of developments in the west. In: Chirot D (ed) The origins of backwardness in Eastern Europe. University of California Press, Berkeley, pp 15–53

Broadberry SN, Campbell B, van Leeuwen B (2013) When did Britain industrialise? The sectoral distribution of the labour force and labour productivity in Britain, 1381–1851. Explor Econ Hist 50:16–27

Broadberry SN, Campbell B, Klein A, Overton M, van Leeuwen B (2015) British economic growth, 1270–1870. Cambridge University Press, Cambridge, UK

Buyst E (2011) Towards estimates of long term growth in the Southern low countries, ca.1500–1846. Results presented at the Conference on Quantifying Long Run Economic Development, Venice, 22–24 March 2011

Carmichael SG, de Pleijt AM, van Zanden JL, de Moor T (2016) The European marriage pattern and its measurement. J Econ Hist 76(1):196–204

Clark C (1940) Conditions of economic progress. Macmillan and Co, London

Clark G (1996) The political foundations of modern economic growth: England, 1540–1800. J Interdiscip Hist 26:563–588

De Moor T, van Zanden JL (2010) Girl power: the European marriage pattern and labour markets in the North Sea region in the late medieval and early modern period. Econ Hist Rev 63(1):1–33
de Pleijt AM, van Zanden JL (2016) Accounting for the little divergence: what drove economic growth in preindustrial Europe, 1300–1800? Eur Rev Econ Hist 20(4):387–409
de Pleijt AM, van Zanden JL (2018) Two worlds of female labour: gender wage inequality in Western Europe, 1300–1800. EHES Working Papers in Economic History, no. 138
Dennison T, Ogilvie S (2014) Does the European marriage pattern explain economic growth? J Econ Hist 74(3):651–693
Dennison TK, Ogilvie S (2016) Institutions, demography, and economic growth. J Econ Hist 76(1):205–217
Dyer C (1989) Standards of living in the late middle ages; social change in England c. 1200–1520. Cambridge University Press, Cambridge, UK
Epstein SR (1991) Cities, regions and the late medieval crisis: Sicily and Tuscany compared. Past Present 130(1):3–50
Epstein SR (2001) Freedom and growth: the rise of states and Markets in Europe 1300–1750. Routledge, London
Fouquet R, Broadberry SN (2015) Seven centuries of European economic growth and decline. J Econ Perspect 29(4):227–244
Galor O (2011) Unified growth theory. Princeton University Press, Princeton
Hatcher J (2011) Unreal wages: problems with long-run standards of living and the 'golden age' of the fifteenth century. Unpublished manuscript, presented at the annual meeting of the Economic History Society
Hoeppner Moran JA (1985) The growth of English schooling, 1340–1548: learning, literacy and laicization in Pre-Reformation York Diocese. Princeton University Press, Princeton
Hoffman PT, Norberg K (1994) Conclusion. In: Hoffman PT, Norberg K (eds) Fiscal crises, liberty, and representative government 1450–1789. Stanford University Press, Stanford, pp 299–310
Humphries J, Weisdorf JL (2015) The wages of women in England, 1260–1850. J Econ Hist 72(2):405–447
Krantz O (2017) Swedish GDP 1300–1560 a tentative estimate. Lund Papers in Economic History: General Issues, No. 152
Kuznets S (1966) Modern economic growth: rate, structure, and spread. Yale University Press, New Haven
Maddison A (2001) The world economy: a millennial perspective. OECD Publishing, Paris
Malanima P (2011) The long decline of a leading economy. GDP in North Italy 1300–1911. Eur Rev Econ Hist 15:169–219
Malinowski M, van Zanden JL (2017) Income and its distribution in preindustrial Poland. Cliometrica 11:1–30, (forthcoming)
Meisenzahl R, Mokyr J (2012) The rate and direction of invention in the British industrial revolution: incentives and institutions. In: Lerner J, Stern S (eds) The rate and direction of inventive activity revisited. University of Chicago Press, Chicago, pp 443–479
Mokyr J (2002) The gifts of Athena: historical origins of the knowledge economy. Princeton University Press, Princeton
Mokyr J (2005) Long-term economic growth and the history of technology. In: Handbook of economic growth, vol 1. Elsevier, Amsterdam, pp 1113–1180
Mokyr J (2009) The enlightened economy: an economic history of Britain, 1700–1850. Yale University Press, New Haven
North DC, Weingast B (1989) Constitutions and commitment: evolution of institutions governing public choice in seventeenth century England. J Econ Hist 49(4):803–832
Orme N (2006) Medieval schools: from Roman Britain to renaissance England. Yale University Press, New Haven
Pamuk S (2006) The black death and the origins of the 'Great Divergence' across Europe, 1300–1600. Eur Rev Econ Hist 11:289–317
Ridolfi L (2016) The French economy in the longue durée. A study on real wages, working days and economic performance from Louis IX to the Revolution (1250–1789). Dissertation IMT

School for Advanced Studies, Lucca, available at http://e-theses.imtlucca.it/211/1/Ridolfi_phdthesis.pdf
Schön L, Krantz O (2015) The Swedish economy in the early modern period: constructing historical national accounts. Eur Rev Econ Hist 16:529–549
Stephens WB (1987) Education, literacy and society, 1830–70: the geography of diversity in provincial England. Manchester University Press, Manchester
Stephenson JZ (2018a) 'Real' wages? Contractors, workers, and pay in London building trades, 1650–1800. Econ Hist Rev 71(1):106–132
Stephenson JZ (2018b) Looking for work? Or looking for workers? Days and hours of work in London construction in the eighteenth century. University of Oxford Discussion Papers in Economic and Social History, no. 162
Van Bavel BJP (2006) Rural wage labour in the 16th-century low countries: an assessment of the importance and nature of wage labour in the countryside of Holland, Guelders and Flanders. Contin Chang 21:37–72
Van Bavel BJP, van Zanden JL (2004) The jump-start of the Holland economy during the late-medieval crisis, c. 1350 – c. 1500. Econ Hist Rev 57(3):503–532
Van Zanden JL (2009) The long road to the industrial revolution: the European economy in a global perspective, 1000–1800. Global economic history series, vol 1. Brill, Leiden
Van Zanden JL (2011) The Malthusian intermezzo: Women's wages and human capital formation between the late middle ages and the demographic transition of the 19th century. Hist Fam 16:331–342
Van Zanden JL, Felice E (2017) Benchmarking the middle ages. XV century Tuscany in European perspective. CGEH Working Papers, No. 81
Van Zanden JL, van Leeuwen B (2012) Persistent but not consistent: the growth of national income in Holland, 1347–1807. Explor Econ Hist 49(2):119–130
Van Zanden JL, de Moor T, Zuijderduijn J (2012a) Small is beautiful. On the efficiency of credit markets in late Medieval Holland. Eur Rev Econ Hist 16:3–22
Van Zanden JL, Buringh E, Bosker M (2012b) The rise and decline of European parliaments, 1188–1789. Econ Hist Rev 65(3):835–861
Voigtländer N, Voth H-J (2013) How the west "Invented" fertility restriction. Am Econ Rev 103(6):2227–2264

The Industrial Revolution: A Cliometric Perspective

Gregory Clark

Contents

Abstract

The Industrial Revolution in England represented most importantly a change in the growth rate of the efficiency of the economy from close to zero in the years before 1800 to rates typical of those for modern England or the USA by 1860. This chapter details the overall change in productivity growth rates and shows also how this created an even greater increase in income per capita from induced capital accumulation. It also details the sectoral sources of this growth. Lastly, the chapter considers how this fundamental economic transformation might be explained as a function of institutions, ideas, demography, and human capital investments.

Keywords

Economic Growth · Industrialization · Industrial Revolution

G. Clark (✉)
University of California, Davis, CA, USA
e-mail: gclark@ucdavis.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_10

Introduction

Much is known of the story of the Industrial Revolution: the innovations in industry, the enclosure of the common fields, the turnpike trusts, the growth of cities, the spread of railways, the people, and the personalities. This essay, however, concerns the quantitative underpinning of the Industrial Revolution and what can be learned of its nature from such a quantitative analysis.

The first issue is the overall rate of growth of the macroeconomic aggregates for this period: output per person, income per person, the capital stock per person, the average wage, returns to capital, and land rents. The traditional approach to estimating output has been through estimating the sectoral outputs of the economy: agriculture, industry, transport, services, and government.[1] Table 1 shows these aggregate estimates for benchmark years. However, such estimates are still quite fragile in some areas.[2] Thus, Table 1 shows that the output-based estimates of GDP show a faster rise of output in the Industrial Revolution era than do estimates based on the payments to the factors of production: labor, land, and capital. The factor payments approach is based on an assumption of a constant number of hours worked per worker over the years 1700–1870. If hours increased in the Industrial Revolution era, then growth could have been somewhat faster than shown in Table 1. However, there is no strong evidence of any substantial increase in hours in these years. Already in the late eighteenth century, a day of work for building workers is assumed in accounts to be 10 h.[3]

Whatever series more accurately reflects the growth of output, the data is very clear that in the eighteenth century, economic growth was slow at a rate of less than 0.3% per capita per year (and likely only 0.17%) and only moved upward toward modern rates at the beginning of the nineteenth century. But even in the later years of the classic Industrial Revolution era, output per person was growing at only about a third the typical rate of growth in the modern economy.

The slow growth rates in the early years of the Industrial Revolution explains why none of the first generations of political economists had any idea of the momentous transformation of economic possibilities that was occurring all around them. None of Adam Smith's *The Wealth of Nations* (1776), Thomas Robert Malthus' *An Essay on the Principle of Population* (1798), David Ricardo's *On the Principles of Political Economy and Taxation* (1821), or James Mills' *Elements of Political Economy* (1821) contains any hint of the growth possibilities unleashed by the Industrial Revolution. Indeed, the very term *Industrial Revolution* did not enter currency until the 1880s.

[1]The contributors here included Deane and Cole (1962), Crafts (1985), Crafts and Harley (1992), and Broadberry et al. (2014).

[2]These estimates tend to assume large increases in agricultural output so that implied efficiency growth in agriculture is at a faster rate for the economy as a whole during the Industrial Revolution. Evidence from prices, wages, land rents, and capital returns in agriculture does not support such an optimistic assessment.

[3]Clark (2005).

Table 1 Estimates of growth in the industrial revolution era

Decade	N[a]	Real GDP[a]	Real GDP/N[a]	Net national income[b]	NNI/N[b]
1700s	100	100	100	100	100
1760s	122	144	117	133	110
1800s	176	234	133	195	118
1860s	381	807	212	610	170

Sources: Broadberry et al. (2014), Clark (2010)
Notes: All values set to 100 in the 1700–1709
N total population, *NNI* net national income
[a]Britain
[b]England

At the aggregate level, the transformation the Industrial Revolution represents is very simple. The growth of output per person in modern economies has two major proximate sources: more capital per worker and more efficiency in translating input into output. At the proximate level, all growth since the Industrial Revolution can be decomposed as

$$g_y = ag_k + g_A \tag{1}$$

where g_y is the growth of output per worker hour, a is the share of capital in national incomes, g_k is the growth of the capital stock per worker, and g_A is the growth rate of efficiency. Since the onset of the Industrial Revolution, the capital stock has grown roughly as rapidly as output. Also, the share of capital in all earnings has remained about a quarter. Figure 1, for example, shows the earning shares of labor, capital, and land in England 1750–2000. Thus, only about a quarter of all modern growth in income per person comes directly from physical capital. The rest is a steady rise in the measured efficiency of the economy.

The Industrial Revolution fits squarely into this modern growth pattern. Indeed, as Table 2 shows, efficiency was more heavily responsible for growth during the Industrial Revolution than at any time since.[4] Increased investments in physical capital per worker in England 1760–1860 were relatively unimportant in explaining the overall growth of output per worker. Capital per worker rose no faster than output per worker so that from the onset of modern growth, efficiency growth dominated.

While Eq. 1 suggests that efficiency growth and physical capital accumulation are independent sources of growth, in practice in market economies there has been a strong correlation between the two proximate sources of growth. Economies with substantial efficiency growth are also those with substantial growth rates of physical capital. Something links these two sources of growth.

Some economists, most notably Paul Romer, have theorized that this correlation derives from physical capital accumulation creating substantial external benefits not

[4]This is because a significant drag on the growth of output per person in the Industrial Revolution era was the decline in farmland per person. Since 1870, the landshare in all incomes became so modest that this drag became unimportant.

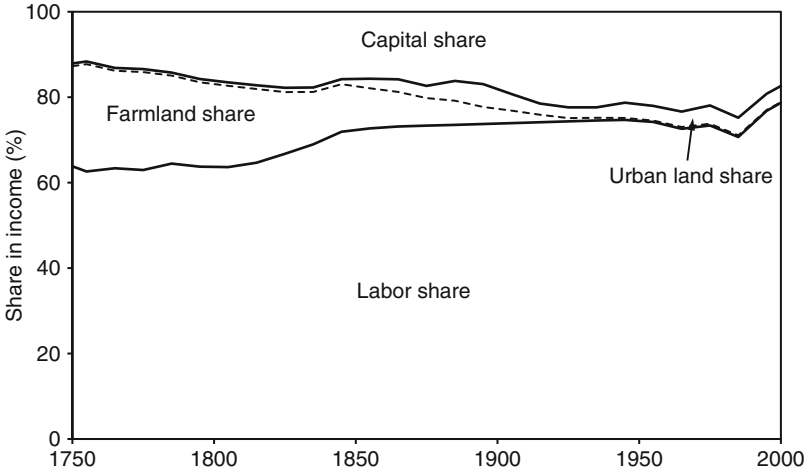


Fig. 1 Factor shares 1750–2000, England (Source: Clark 2007a, Fig. 14.4)

Table 2 Growth rates in England/Britain 1700 and later

Period	Real GDP/N[a] (% per year)	NNI/N[b] (% per year)	Efficiency (% per year)
1700s–1760s	0.26	0.16	0.11
1760s–1800s	0.32	0.18	0.37
1800s–1860s	0.78	0.60	0.58
1860s–1900s	2.14	–	–
1900s–1950s	1.68	–	–
1950s–2000s	2.60	–	–

Sources: Broadberry et al. (2014), Clark (2010)
Notes: All values set to 100 in the 1700s
N total population, *NNI* net national income
[a]Britain
[b]England

captured by the investors in capital (Romer 1986). However, for this explanation to work, there would have to be $3 of external benefits accruing to physical capital investments for every $1 of privately captured benefit. Most of the modern physical capital stock, however, is still such mundane stuff as houses, shops, warehouses, factories, roads, bridges, and water and sewer systems. These types of investment we would not expect to generate substantial external benefits. So, if productivity advance is systematically associated with the growth of the stock of such physical capital, there must be another mechanism.

The most plausible one is that the association of physical capital accumulation with efficiency advance stems just from the effects of efficiency advance on increasing the marginal product of capital. In a world of relatively constant real interest rates since the Industrial Revolution, such a rising marginal product will induce more investment. And indeed, if the economy is roughly Cobb-Douglas in its production

structure, efficiency advances will induce a growth of the physical capital stock per person at a rate equal to the growth of output per person so that the capital–output ratio is constant. This is to a first order what we observe since the Industrial Revolution.

Thus, at a deeper level, all modern growth seemingly stems from this unexplained rise in economic efficiency as a product of a rise in knowledge about production processes. Somehow, after 1780, investment in such knowledge increased, or enquiry became much more effective in creating innovation.

Before the Industrial Revolution, we find no sign of any equivalent efficiency advances. This is true globally all the way from 10,000 BC to 1800, where we can measure the implied rate of productivity advance just from the rate of growth of population. In this long interval, average estimated rates of efficiency advance are 0.01% per year or less. We know this because we can assume before the Industrial Revolution, because of the Malthusian trap, that output per person and capital per person was in the long run constant. In that case, any gains in efficiency will be absorbed by population growth according to the formula[5]

$$g_A = cg_N \tag{2}$$

where c is the share of land in national income and g_N the rate of population growth.

We can thus approximate efficiency growth rates from population growth rates if we look at sufficiently long intervals. Table 3 shows these calculations at a world level. Implied rates of technological advance are always extremely slow even in the 250 years leading up to the Industrial Revolution.

But it is also true that implied rates of technological advance are also slow for those economies where we can measure actual efficiency levels before 1800 through measurements of the real payments to factors. Figure 2 shows the implied efficiency in England from 1250 to 2000 calculated from the formula

Table 3 Population and technological advance at the world level, 130,000 B.C. to 1800

Year	Population (millions)	Population growth rate (%)	Technology growth rate (%)
130,000 BC	0.1	–	–
10,000 BC	7	0.004	0.001
1 AD	300	0.038	0.009
1000 AD	310	0.003	0.001
1250 AD	400	0.102	0.025
1500 AD	490	0.081	0.020
1750 AD	770	0.181	0.045

Source: Clark (2007a, Table 7.1)

[5]For a more detailed explanation, see Clark (2007a, pp. 379–382).

Fig. 2 Estimated efficiency of the english economy, 1250–2000 (Source: Clark 2010)

$$A = \frac{r^a p_k^a w^b s^c}{p} \tag{3}$$

where A indexes economic efficiency, p is an index of output prices, r is the real return on capital, p_k is an index of capital goods prices, w is an index of real wages, and s is an index of land rents. a, b, and c are the shares of each input type in national incomes. As can be seen, there is, surprisingly, in England no sign of any significant improvement in the efficiency of the economy all the way from 1250 to 1800. Only around 1800 does the modern age of steady efficiency advance appear. Before that, the measured efficiency of the economy fluctuated, peaking around 1450 but with almost no upward trend.

The Industrial Revolution thus seems to represent a singularity, a unique break in world history but also an event where we know clearly what we have to explain. Why did the rate of expansion of knowledge about production efficiency increase so dramatically in England around 1800? Figure 3 shows that the upturn in productivity growth rates can be located to the 1780s/1790s. That upturn is preceded by seven decades in which the average annual productivity growth rate was a mere 0.14% per year, fast by the standards of the preindustrial world but glacially slow in modern terms. Overall, productivity growth rates from 1780–1789 to 1860–1869 averaged 0.58% per year, more than half way to fully modern levels.

We also know what sectors contributed most of the productivity advance during this period. For each sector, we can calculate an efficiency growth rate from the formula (3), in growth rate terms,

$$g_{A_j} = -g_{p_j} + a_j g_{r_j} + a_j p_{kj} + b_j g_{w_j} + c_j g_{s_j} \tag{4}$$

where j indicates each sector. To implement the formula by sector in this way, however, where we use output prices as p_j, would require being able to measure the capital, labor, and land embodied in all the inputs purchased by each sector. For coal mining, we would need to know the shares of capital, labor, and land embodied

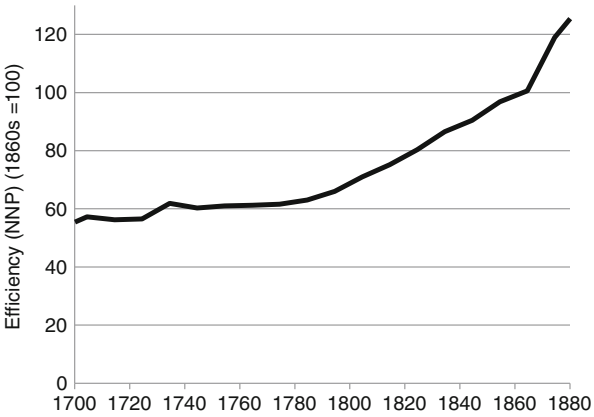


Fig. 3 Efficiency levels, England, 1700–1880 (Source: Clark 2010)

in horse feed, bricks, pit timbers, and steam engines. A more feasible procedure is one where we measure efficiency where p_j measures just the value added in the industry, the difference between the value per unit of output and the cost of purchased inputs.

In this case, national productivity growth will be related to productivity advances in individual sectors through the equation

$$g_A = \sum \theta_j g_{Aj} \tag{5}$$

where g_{Aj} is the growth rate of productivity by sector and θ_j is the share of j in total value added in the economy.[6] These results are shown in Table 4.

Textiles contributed nearly half, 43%, of all measured productivity advance. Improvements in transport, mainly the introduction of the railway, was the next biggest source of advance, contributing 20%. Agriculture, ironically, also contributed almost 20%. Coal, iron, and steel were in themselves minor contributions despite the fame of these sectors and their innovations in this period and despite the huge growth in coal and iron production. Productivity growth in the half of the economy not covered in Table 4 was modest, less than 0.20% per year.

[6]Another procedure to measure productivity growth at the sectoral level is to treat the major inputs in the same way as capital, labor, and land and measure efficiency growth as

$$g_{A_j} = -g_{p_j} + a_j g_{r_j} + a_j p_{kj} + b_j g_{w_j} + c_j g_{s_j} + d g_{m_j}$$

where d is the share of purchased inputs in all costs and m indexes the price of such inputs. In this case,

$$g_A = \sum \varphi_j g_{Aj}$$

where φ is the ratio of sales in each industry to national income. $\sum \varphi_j > 1$ since some output is used as input into other industries and not for consumption or investment. This approach to measuring Industrial Revolution productivity advance was pioneered by McCloskey (1981).

Table 4 Sources of industrial revolution efficiency advance, 1780s–1860s

Sector	Efficiency growth rate (%)	Share of value added	Contribution to national efficiency growth rate (% per year)
All textiles	2.3	0.11	0.25
Iron and steel	1.8	0.01	0.02
Coal mining	0.2	0.02	0.00
Transport	1.5	0.08	0.12
Agriculture	0.4	0.30	0.11
Identified advance	–	0.51	0.49
Whole economy	–	1.00	0.58

Source: Clark (2007a, Table 12.1)

The decomposition in Table 4 establishes some things already. The Industrial Revolution has been thought of by some as essentially consisting of the arrival of the first of what have been called *general-purpose technologies*, the steam engine. *General-purpose technologies (GPTs)*, a rather nebulous concept, have been variously defined. They can be loosely thought of as innovations that have pervasive application throughout the economy, that go through a prolonged period of improvement, and that spawn further innovation in the sectors in which they are employed.[7] Various GPTs have been identified such as steam power in the Industrial Revolution, the introduction of electricity, and the recent information revolution.

Steam power in England certainly permeated a number of areas in the Industrial Revolution. It was important in coal mining, on the railroads, and in powering the new textile factories. The steam engine itself underwent a long process of improvement in thermal efficiency and in the ratio of power to weight from its first introduction by Thomas Newcomen in 1707–1712 to the 1880s. The earliest engines had a thermal efficiency as low as 0.5%, while those of the 1880s could achieve thermal efficiencies of 25%. The steam engine was associated also with the widespread use of fossil energy in the economy to replace wind, water, and animal power sources in transport, home heating, and manufacturing. Output of the coal mining industry rose from Table 4 suggests, however, that whatever role steam power played in economy-wide productivity advance after the 1860s, its role up to then in the new productivity advance of the Industrial Revolution was minor. Coal mining and iron and steel production contributed very little to Industrial Revolution productivity advance, and most of the productivity advance in these industries did not stem from the introduction of steam power.[8] By the end of the eighteenth century, most coal mines used steam engines to do the winding of the coal and to pump water out from the workings. But horses were technically a viable alternative power source. Thus,

[7]Bresnahan and Trajtenberg (1996).

[8]Clark and Jacks (2007).

Table 5 Cost increase from absence of steam in mining, by epoch

Period	Share of costs coal for winding, pumping (%)	hph/ton	Cost increase (d./ton)	Cost increase (%)
1720–1759	6.0	1.6	0.4	1
1770–1799	4.4	2.0	5.3	14
1800–1839	4.9	4.8	12.2	20
1840–1869	3.0	2.9	6.2	10

Source: Clark and Jacks (2007, Table 4)

in the Walker colliery in the northeast coalfield in 1765, the deepest coal mine in England at that point at 600 ft, the coal was still lifted from the mine by a gin powered by eight horses.[9]

Table 5 calculates how much the absence of coal would have raised costs of production in each epoch. The method here is to calculate how many pounds of the equivalent of best coal were used at the colliery per ton of coal raised using the share of mining costs reported as coal consumed in winding or pumping. These pounds of coal were then translated into horsepower-hours per ton of coal, shown in column 3 of the table. The extra cost of supplying this energy as horsepower as opposed to steam power is given in column 4, and the percentage increase in production costs this would imply appears in the last column. The implication is that production costs in the nineteenth century would have risen by 10–20% absent the introduction and development of steam power in collieries. The absence of the new steam technology would not have crippled the industry even late into the Industrial Revolution.

Even in transport, a substantial part of the productivity advance is attributable to the improvement of the traditional road transport system, the introduction of canals, and improvements in sailing ships. The textile factories of the Industrial Revolution could, if necessary, have still been powered by waterwheels even as late as the 1860s. As in coal mining, power costs would have been higher in this case, but power costs were also a small share of total costs in textile mills. Advances in textiles and agriculture explain the majority of the Industrial Revolution.

Recent accounts of the Industrial Revolution, most noticeably in the work of E. A. Wrigley and Kenneth Pomeranz, would still make coal the key actor despite the absence of much sign of productivity growth in coal mining in Table 4.[10] Both argue that the switch from a self-sustaining organic economy to a mineral resource–dependent inorganic economy was central to the Industrial Revolution. Indeed, Pomeranz's account of the Industrial Revolution was dubbed "Coal and Colonies" by one reviewer.[11] Pomeranz argues that Britain, in contrast to China, had accessible deposits of coal near population centers. That, rather than differences in innovative potential, explains British success and Chinese failure. While the absence

[9]By 1828, three-quarters of mines in the Newcastle area were still less than 600 f. deep. Clark and Jacks (2007, Table 2).

[10]Wrigley (1988), Pomeranz (2000).

[11]Vries (2001).

of steam power would not have impeded growth much, would the absence of the coal deposits altogether have prevented the growth of the Industrial Revolution?

Coal output expanded enormously in the Industrial Revolution era. By the 1860s, it was supplying power for domestic purposes that was equivalent to the annual energy production of 25 million acres of woodland. This would have required nearly the entire farmland area in England in these years. Thus, if England had to depend only on its own supplies of energy, costs would soon have soared and the economy taken a very different path. There was, however, in the Baltic region alone a lot of wood available to the English economy throughout the Industrial Revolution era. By the nineteenth century, the Baltic was a major supplier of timber to England and the Netherlands. The regions bordering the Baltic produced enough energy in the form of wood to completely replace the energy supplied by coal for domestic purposes even as late as the 1860s. That energy would be more expensive, but the value of coal at the pithead in the 1860s in England was only around 2% of national income. But declines in shipping costs between the seventeenth and nineteenth centuries meant that the transport costs for this wood fuel in the 1860s would not have been much greater than the cost of domestic coal supplied to consumers in places like London. So total energy costs to the economy would likely have risen only modestly, and the gains in efficiency from textiles, farming, and transport would have been largely preserved.[12]

The diverse nature of productivity advance in this era makes the Industrial Revolution all the more puzzling. The revolution in textiles came through mechanical innovations that can be traced to a number of heroic individual innovators: John Kay, Richard Arkwright, James Hargreaves, Samuel Crompton, Edmund Cartwright, and Richard Roberts. But the improvements in agriculture stem from the advances of thousands of anonymous farmers in improving yields, mainly involving nonmechanical changes. Such celebrated figures of narrative accounts of the agricultural revolution such as Jethro Tull, "Turnip" Townsend, and Arthur Young on examination played no important role. Tull had no idea of what the sources of plant growth actually were.[13] Turnips were introduced into Norfolk rotations in the 1660s, long before Townsend, born in 1674, began farming.[14] Further, there is little sign that the new rotations promoted by Townsend and Young were themselves an important element in improved yields. Mark Overton looking at grain yields in probate inventories in the seventeenth century finds these to be no higher on farms which had introduced the new rotations.[15] Young made vigorous claims for the productivity benefits of such institutional reforms as the privatization of common fields and the ending of tithe payments. But the efficiency gains from enclosing common lands were miniscule. Even though a quarter of English farmland was

[12]For more details, see Clark and Jacks (2007).

[13]Wicker (1957).

[14]Overton (1985, Table 1).

[15]Overton (1991, pp. 309–310).

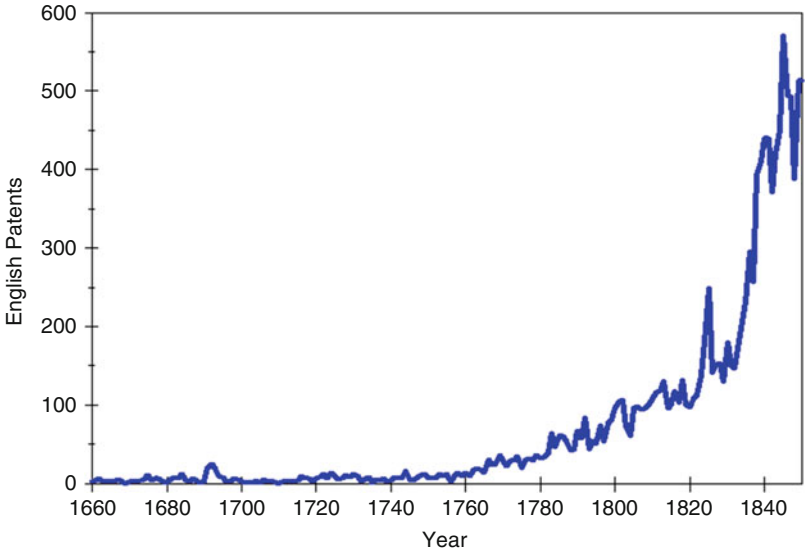


Fig. 4 Patents per year, England, 1660–1851 (Source: Mitchell 1988)

enclosed 1750–1830, the gain in farming efficiency would be less than 1%. And the tithe reforms of the 1830s had even smaller measured benefits.[16]

What is clear, however, is that at a fundamental level, the Industrial Revolution was driven by an upturn in the rate of technical innovation within the English economy.[17] This upturn, at least at interest in innovating, shows up clearly in English patent statistics, as summarized in Fig. 4. Clearly in the 1760s, before there was any general perception of an upturn in the rate of economic change and before the aggregate productivity data shows any signs of faster efficiency growth, more innovators were finding their way to London to file patent applications. What triggered this process?

The Problem of the Netherlands

When it comes to explaining why England was the first nation to experience modern productivity growth rates, the puzzle is deepened by the example of one earlier economy, the Netherlands, in the years 1581–1795. After a revolt that began in 1568, in 1581 the northern provinces of Flanders successfully attained independence from the Spanish Crown (in the form of the Hapsburg Empire). Despite a continuing military struggle against the Hapsburgs that ended only in 1648, the break was

[16]Clark (1998), Clark and Jamelske (2005).

[17]This is also the analysis of Mokyr (2003, 2012).

associated with a period of growth and prosperity known as the Dutch Golden Age, which spanned the seventeenth century. As one English commentator stated in 1669,

> Scarce any Subject occurs more in the learned discourse of ingenious man than that of the marvelous progress of this little state . . . which has grown to a height infinitely transcending all the ancient Republicks of Greece but not much inferior in some respects even to the greatest Monarchies of these latter Ages.[18]

The Netherlands economy was characterized by largely free markets internally for labor, land, capital, and commodities.[19] The Netherlands was also a very open economy, conducting extensive trade with the Baltic region (for grains and raw materials), the rest of Europe, and Asia. The political structure guaranteed property rights, contract enforcement, and freedom of movement of labor. As the Hapsburg forces recaptured Ghent, Bruges, and Antwerp, the Protestants of these cities, which included many skilled craftsmen and merchants, largely migrated north to Dutch territories. The Netherlands also welcomed Jewish refugees from the Iberian peninsula, both those continuing to practice Judaism and New Christians still treated as second-class citizens in their home countries.

By 1595, soon after gaining independence, Dutch merchants began sending ships to engage in the spice trade of the East, heretofore a Portuguese monopoly. Despite armed clashes with the Portuguese, the Dutch were able to force their way into the trade. After the formation in 1600 of the English East India Company with monopoly privileges in trade in the East, the Dutch set up their own East India Company in 1602, the *Vereenigde Oostindische Compagnie* (VOC). This proved an enormously profitable enterprise all through the seventeenth century. As a reflection of its scale, between 1602 and 1800 the VOC recruited almost a million men for work in Asia as traders, sailors, and soldiers. At this time, the population of the Netherlands was only around two million. In pursuit of spice trade profits, the VOC colonized the main islands of Indonesia for the Dutch. Its operations and profits were significantly greater than those of its main rival, the English East India Company.

This influx of talent made the Netherlands the leading economy of Europe in terms of living standards, science, intellectual life, and the arts by 1600. Real wages in the western Netherlands were the highest in Europe and maintained this position until well into the Industrial Revolution. Thus, de Vries and van der Woude estimate that in 1660, Dutch GDP per capita exceeded that of England by more than 30%.[20] By the mid-seventeenth century, trade and industry made up the bulk of the economy, with less than 40% of the labor force employed in farming. Along with its eastern trade empire, the Dutch developed, despite their high labor costs, a major shipbuilding industry with many technical innovations in ship design and construction. Its merchant shipping fleet was the largest in Europe in the seventeenth century.

[18] Aglionby (1669, pp. 3–4).

[19] The discussion below is largely based on De Vries and Van der Woude (1997), though see also Freist (2012) and de Vries (2000).

[20] De Vries and Van der Woude (1997, p. 710).

Again despite high labor costs, it had a major textile industry. It also had a number of industries based on the exploitation of cheap peat fuel such as ceramics (bricks, tiles, pottery, and clay pipes), brewing, and sugar refining.

Capital was unusually cheap in the Dutch republic as witnessed by the low rates of return on government debt, land, and housing.[21] This, along with the flat topography, allowed the Dutch to develop an extensive canal system linking all major cities. Canal boats would travel hourly between the major cities in much the way airlines now shuffle passengers between major US and European destinations.[22] The canal system also allowed for the cheap supply of peat fuel to industry and urban areas.

With the developments in trade, industry, and agriculture came innovation in finance. The Amsterdam Stock Exchange, established in 1602, is the oldest in the world. The Bank of Amsterdam, established in 1609, was a precursor to the central banks of the modern world.

The riches of the economy and the openness to immigration of talent from across Europe made the seventeenth century Netherlands a center for both the arts and scientific enquiry. By the seventeenth century, average levels of education in the Netherlands were as high, or higher, than those in Industrial Revolution England. One measure of this is the share of grooms and brides signing marriage registers. Table 6 shows the rates for the Netherlands circa 1620, 1700, and 1800 compared to the rates in England and elsewhere in Europe circa 1800. In the seventeenth century, implied literacy rates are, at worst, as good in the Netherlands as for Industrial Revolution England. And literacy rates in northern France and northern Germany exceed those of Industrial Revolution England.

The extensive developments above led Jan de Vries and Ad Van der Woude to title their 1997 book summarizing this history *The First Modern Economy*. However, while the Netherlands was modern in all the ways described above, in another respect it remained firmly a preindustrial and pre-Industrial Revolution economy. For while all the changes outlined above were significant and associated with economic growth and prosperity, the Dutch of the Golden Age never achieved any breakthrough in terms of productivity growth. This is illustrated in Fig. 5, which shows real wages in the western Netherlands from 1500 to 1799. The growth of the Golden Age was not accompanied by any rise in real wages. This implies that the overall productivity of the Netherlands economy expanded little in the years 1550–1650. The level of productivity is the weighted average of real wages, real land rents, and real returns on capital, with the weights equal to the share of each of these factors in incomes received.[23]

Given the failure of real wages to increase, given that wages would be half or more of all incomes, and given the declining returns on capital, even if real land rents

[21]By 1665 the State of Holland was able to reduce rates on its long term debt to 4%.

[22]De Vries (1978).

[23]See Eq. 3 above.

Table 6 Literacy rates 1800 and earlier

Place	Year	Men	Women
Netherlands	1620	60	37
Netherlands	1700	65	48
Netherlands	1800	75	60
England	1800	60	40
N. France	1800	71	44
N. Germany	1800	85	44
Belgium	1800	60	37

Sources: De Vries and van der Woude (1997, pp. 170–171, 314) (Netherlands). Reis (2005, p. 202)

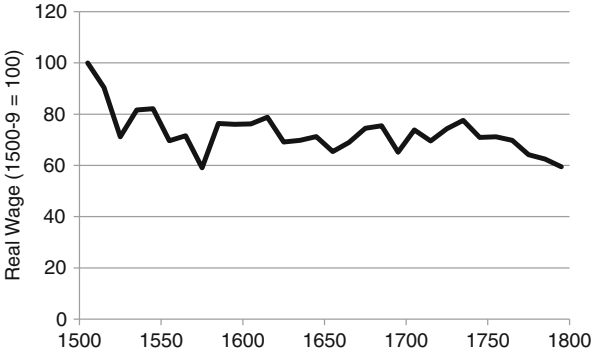


Fig. 5 Real wages, Western Netherlands, 1500–1799 (Sources: Based on the wage series in De Vries and Van der Woude (1997, pp. 610–611), with budget weights and cost of living as described in van Zanden (2008))

rose in this period, the overall gains in efficiency would be very modest. The Dutch had a Golden Age, but they did not have an Industrial Revolution.

The failure of Golden Age Netherlands to experience an Industrial Revolution suggests that most proposed explanations of the English Industrial Revolution are misguided. Robert Allen, for example, has recently proposed that the Industrial Revolution in England was driven by high labor and low energy costs in eighteenth-century England, leading to a replacement of hand labor by machines driven by steam engines.[24] Yet, 200 years earlier, we see in the Netherlands even higher wages than for England in 1780 and again low energy costs but no Industrial Revolution.

The example of the Netherlands is particularly a problem for the idea that the Industrial Revolution is the product of institutional innovation. A powerful modern school within economics is *institutionalism*, which asserts that institutions, formal or informal, explain most differences in economic outcomes and that systematically early societies had institutions that discouraged economic growth (see, e.g., Acemoglu et al. 2001, 2002, 2005; Acemoglu and Robinson 2012; DeLong and

[24] Allen (2009).

Shleifer 1993; Greif 2006; North 1981, 1994; North and Thomas 1973; North and Weingast 1989; North et al. 2012; Rosenthal 1992).[25]

The common feature that Douglass North and other such *institutionalists* point to in early societies is that political power did not derive from popular elections. In preindustrial societies, as a generalization, the rulers ultimately rested their political position on threats of violence. Indeed, there is a good empirical association between democracy and economic growth. By the time England achieved its Industrial Revolution, it was a constitutional democracy where the king was merely a figurehead. The USA, the major country with the highest GDP per person since the 1850s or earlier, has always been a democracy.[26]

Economic efficiency in any society requires that property rules be chosen to create the maximum value of economic output. In such a case, a disjuncture can arise between the property rules in the society that will maximize the total value of output and the property rules that will maximize the output going to the ruling elite. Indeed, North and others have to argue that such a disjuncture systematically arises in all societies before the Industrial Revolution. This idea has been restated recently as the replacement of extractive economic institutions designed just to secure income for a ruling clique with inclusive economic institutions designed to maximize the output of societies as a whole (Acemoglu and Robinson 2012).

One subset of such theories that has shown amazing persistence despite its inability to account for the most basic facts of the Industrial Revolution is that which links the Industrial Revolution to the earlier *Glorious Revolution* of 1688–1689. Thus, the recent widely read book by Acemoglu and Robinson, *Why Nations Fail*, has a chapter titled "How a political revolution in 1688 changed institutions in England and led to the Industrial Revolution" (Acemoglu and Robinson 2012).

The *Glorious Revolution* established the modern political system of the UK, a system that has been continuously modified but not fundamentally changed since then. The new political system made Parliament, the representative of the propertied classes in England in 1689, the effective source of power in what is nominally a monarchy.

A basic problem with placing political developments at the heart of the Industrial Revolution is that the *Glorious Revolution* of 1688–1689 had no discernible impact on economic efficiency before 1770, nearly three generations after the institutional change, as Fig. 6 shows. The figure shows the level of net national product per person across these years. It is also clear in the figure that even the earlier political and military disruptions of the Civil War of 1642–1649, when Parliament and the King were at war, and the Interregnum of 1649–1660 were not associated with any decline in the efficiency of operation of the economy in the seventeenth century.

[25]Clark (1996, 2007a, b) criticizes this approach.

[26]The recent rise of China is, however, an exception to the general association of growth and democracy.

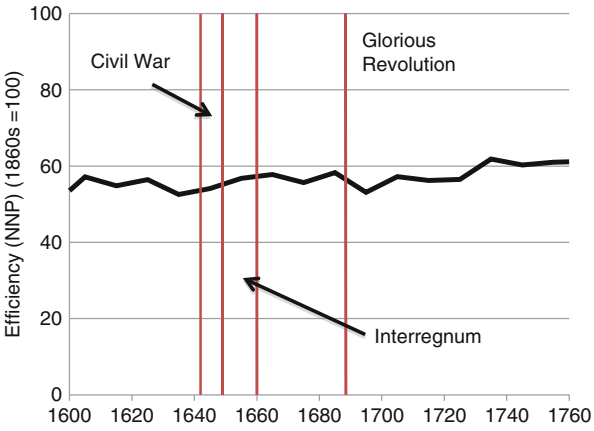


Fig. 6 Economic efficiency and political changes, England, 1600–1770 (Source: Clark 2010)

Further, there is no sign that private investors in England perceived a greater security of property even as a result of the Glorious Revolution. The return to private capital in the economy did not deviate from trend after 1689 (see Clark 1996). Private investors seem to have looked at the political changes with indifference. The return to government debt did eventually decline significantly after 1689 and had fallen to modern levels by the 1750s. This decline was no doubt driven in part by the enhanced taxing power of the government after 1689. But almost all of the money raised from those taxes went to finance the British Navy in the long struggle with France that ended only with the final defeat of Napoleon at Waterloo in 1815. Almost none of the tax revenues went into subsidizing innovation, investment, or education.

And we do see, long before the Glorious Revolution or the Industrial Revolution, societies that had stable representative political systems, the inclusive institutions of Acemoglu and Robinson, but little or no productivity advance. The Dutch Republic of 1588–1795 was, as discussed above, one such regime.[27] While the political system of the Dutch Republic had its tensions and contradictions and periodic instabilities, the leadership was responsive to the needs of the citizenry at city and national levels. The citizenry itself was mainly property-owning burghers. In most Dutch cities, you could become a citizen by inheritance of citizen status, by marrying the daughter of a citizen, by receiving the status as a reward, or by purchase. It is estimated that about half of adult males in cities would be citizens, but the large rural population was largely excluded from citizenship.[28] Even where town councils were not formally elected by citizens as in Holland, they were drawn from this milieu and subject to pressure through other organizations of burghers such as the militia. So the Netherlands was, par excellence, a property-owning democracy whose leaders had regard for the economic interests of the merchant and manufacturing classes.

[27]The Dutch Act of Abjuration of 1581 has been claimed by some to be the precursor of the Declaration of Independence of the USA of 1776.

[28]See Prak (1997), van Zanden and Prak (2006, pp. 121–122).

The Netherlands of the seventeenth century was just one of many earlier European societies that had property holder franchises. From 1223 to 1797, Venice was a Republic, with the government under the control of a mix of popular and patrician representatives. Policy was geared toward the needs of a trading and commercial empire. Venice developed an important trading empire in the Eastern Mediterranean with colonies and dependencies such as Crete, Cyprus, and Dalmatia. It also saw the growth of important manufacturing activities such as glass. But again, none of this institutional framework was reflected in the kind of sustained productivity advance seen in the Industrial Revolution.

Similarly, the free cities of the Hanseatic League were from the Middle Ages dominated by a politics that emphasized the needs of trade and commerce. Lübeck, for example, became a free city in 1226 and remained a city-state until 1937. After gaining its freedom in 1226, Lübeck developed a system of rule and government called Lübeck Law that spread to many other Baltic cities of the Hanseatic League in the Middle Ages such as Hamburg, Kiel, Danzig, Rostock, and Memel. Under Lübeck Law, the city was governed by a council of 20 that appointed its own members from the merchant guilds and other town notables. It was thus government by the leaders of the commercial interests of the cities. Though not democracy, this was government by interests that should have fostered commerce and manufacturing. Under such rule, the Hansa cities became rich and powerful, engaging in substantial manufacturing enterprises such as shipbuilding and cloth production as well as trade. But again, this was not associated with sustained technological advance.[29]

Property in Knowledge

If it is not insecure property rights in general that can explain the long delay in the arrival of the Industrial Revolution and its location finally in England, what about a more specific deficiency? Could it be that the problem was just that all earlier societies lacked the institution of allowing knowledge to be a kind of property?

In both ancient Rome and Greece, the concept that you could own property in ideas or innovations was missing. Thus, in both the Roman and Greek worlds, when an author published a book, there was no legal or practical way to stop the pirating of the text. Copies could be freely made by anyone who acquired a version of the manuscript (on papyrus rolls), and the copier could amend and alter the text at will. Texts might thus be reissued under the name of a new "author."[30] It was common to condemn such pirating of works or ideas as immoral. But writings and inventions were just not viewed as *commodities* with a market value.[31] It is frequently asserted

[29]There have been institutionalist arguments, however, about why Hansa institutions still deviated from those necessary for modern growth. See Lindberg (2009).

[30]This problem continued into at least the seventeenth century in England, where publishers quite freely pirated the works of authors.

[31]See Long (1991, pp. 853–857).

that the concept of intellectual property was alien to cultural norms in imperial China, with the first copyright law only being introduced in the late Qing era in 1910 under Western pressure. There were, however, some limited protections for publishers of block-printed books after the introduction of printing in China in the ninth century. But these seem to have been local and special protections within a legal environment where the idea of intellectual property rights was alien.[32]

While the European ancients and the Chinese may have lacked them, there were systems of intellectual property rights in place, however, long before the Industrial Revolution. The rudiments of a modern patent system were found already in the thirteenth century in Venice. By the fifteenth century in Venice, true patents in the modern sense were awarded regularly. Thus, in 1416, the Venetian Council gave a 50 year patent to Franciscus Petri from Rhodes, a foreigner, for a new type of fulling mill. By 1474, Venetian patent law had been codified. There is also evidence of patent awards in Florence in the fifteenth century. The Venetian innovation granting property rights in knowledge, which was very important to the famous Venetian glass industry, spread to Belgium, the Netherlands, England, Germany, France, and Austria in the sixteenth century as a consequence of the movement of Italian glass workers to these other countries. These workers demanded protection for their trade knowledge as an inducement to set up production in these other countries. Thus, by the sixteenth century, all the major European countries, at least on an ad hoc basis, granted property rights in knowledge to innovators. They did this in order to attract skilled craftsmen with superior techniques to their lands. The spread of formal patent systems thus predates the Industrial Revolution by at least 350 years.

The Netherlands in particular had a fully functioning patent system in place by 1590 and was issuing patents at a much higher rate than England in the early seventeenth century despite having a much smaller population. In the Dutch Golden Age, patent activity was much greater than in the eighteenth century, when the flow of patents slowed to a trickle just as the English patent issues were rising sharply.[33]

The claims of North and his associates for the superiority of the property rights protections afforded by the patent system in eighteenth-century England thus stem from the way in which the system operated after the Glorious Revolution of 1688–1689 established the supremacy of Parliament over the King. Under the patent system introduced in the reign of Elizabeth I, 1568–1603, the system was supervised by government ministers. Political interference led to the creation of spurious monopolies for techniques already developed or the denial of legitimate claims. After the Glorious Revolution, Parliament sought to avoid this by devolving the supervision of patents to the courts. Generally, the courts would allow any patent to be registered as long as no other party objected. No other major European country had a formal patent system as in England before 1791. But the system was in fact notoriously costly in terms of money and time, especially if the patentee wanted protection in Scotland and Ireland as well as England. The application had to pass

[32]Ganea and Pattloch (2005, pp. 205–206).

[33]De Vries and van der Woude (1997, pp. 345–348).

Fig. 7 Cotton spinning and weaving productivity, 1770–1869. *Note*: The *squares* show the decadal average productivities. The years 1862–1865 were omitted because of the disruption of the cotton famine (Sources: Cotton cloth prices, Harley (1998). Labor costs, return on capital, Clark (2010))

through seven court or government offices in London and a further five each in Scotland and Ireland.[34] Applications could be rejected on such technical grounds as the innovator having sold even one of the devices prior to the application. Also, as Fig. 4 shows, while the Glorious Revolution produced a brief increase in patent rates, there was no sustained increase in patenting rates until the 1760s, 75 years after the Glorious Revolution.

Another implausibility in the knowledge appropriability argument is the weak evidence for any increase in returns to innovators in England in the 1760s and later. The textile industry, for example, was in the vanguard of technological change in the Industrial Revolution period. Figure 7 shows efficiency in the production of cotton cloth measured in terms of value added to the cotton input per unit of land, labor, and capital. From 1770 to 1869, efficiency rose about 22 fold. Yet, the gains of the textile innovators were modest in the extreme. The value of the cotton textile innovations alone by the 1860s, for example, was about £115 million in extra output per year in England. But a trivially small share of this value of extra output flowed to the innovators. Table 7, for example, shows the major innovators in cotton textiles and the gains accruing to the innovators through the patent system or other means. Patents mostly provided poor protection, the major gains to innovators coming through appeals post hoc to public beneficence through Parliament. Also, the patent system shows none of the alleged separation from political interference. The reason for this is that Parliament could, on grounds of the public good, extend patents

[34]Khan (2008).

Table 7 The gains from innovation in textiles in the industrial revolution

Innovator	Device	Result
John Kay	Flying Shuttle, 1733	Impoverished by litigation to enforce patent. House destroyed by machine breakers 1753. Died in poverty in France
James Hargreaves	Spinning Jenny, 1769	Patent denied. Forced to flee by machine breakers in 1768. Died in workhouse in 1777
Richard Arkwright	Water Frame, 1769	Worth £0.5 m at death in 1792. By 1781 other manufacturers refused to honor patents. Made most of money after 1781
Samuel Crompton	Mule, 1779	No attempt to patent. Grant of £500 from manufacturers in the 1790s. Granted £5,000 by Parliament in 1811
Reverend Edmund Cartwright	Power Loom, 1785	Patent worthless. Factory destroyed by machine breakers. Granted £10,000 by Parliament in 1809
Richard Roberts	Self-Acting Mule, 1830	Patent revenues barely covered development costs. Died in poverty in 1864

Source: Clark (2007a, Table 12.2)

beyond the statutory 14 years to adequately reward those who made significant innovations. James Watt was the beneficiary of such a grant. But obtaining such Parliamentary grants depended on political patronage just as much as in the old days.

Productivity growth in cotton textiles in England from 1770 to 1870 far exceeded that in any other industry. But the competitive nature of the industry and the inability of the patent system to protect most technological advances kept profits low. Cotton goods were homogeneous. Yarn and cloth were sold in wholesale markets where quality differences were readily perceptible to buyers. The efficient scale of cotton spinning and weaving mills was always small relative to the market. New entrants abounded. By 1900, Britain had about 2,000 firms in the industry. Firms learned improved techniques from innovating firms by hiring away their skilled workers. The machine designers learned improved techniques from the operating firms. Thus, over time, the entire industry – the capital goods makers and the product producers–clustered more and more tightly in the Manchester area. By 1900, 40% of the entire world output of cotton goods was produced within 30 miles of Manchester. The main beneficiaries of this technological advance thus ended up being two parties: consumers of textiles all across the world and the owners of land in the cluster of textile towns, which went from being largely worthless farmland to valuable building sites.

The profit rates of major firms in the industry also provide good evidence that most of the innovation in the textile industry was quickly leaking from the innovators to other producers with no rewards to the innovators. Knick Harley has reconstructed the profit rates being made by some of the more successful cotton spinning and weaving firms in the early Industrial Revolution period (Harley 1998, 2010). The cotton spinners *Samuel Greg and Partners* earned an average profit from 1796 to 1819 of 11.7% per year, just the normal commercial return for a risky venture such as manufacturing. Given the rapid improvements in cotton spinning

productivity going on in the industry in these years, it suggests that whatever innovations were being introduced were spreading from one firm to another very quickly. Otherwise, leading firms such as *Samuel Greg* would have made large profits compared to their competitors.

Similarly, the firm of *William Grey and Partners* made less than 2% per year from 1801 to 1810, a negative economic profit rate. The innovations in the cotton spinning industry seem to have mainly caused prices to fall, leaving little excess profits for the firms that were innovating. From 1777 to 1809, *Richard Hornby and Partners* was in the handloom weaving sector of the industry, which had not yet been transformed by any technological advance. Yet, its average profit rate was 11.4%, as high as *Samuel Greg* in the innovating part of the industry.

The conclusion is that the host of innovations in cotton textiles do not seem to have particularly rewarded the innovators. Only a few, such as Arkwright and the Peels, became noticeably wealthy. Of the 379 people probated in 1860–1869 in Britain who left estates of £0.5 million or more, only 17 were in the textile industry even though, as noted, from 1760–1769 to 1860–1869, this one sector generated nearly half the productivity growth in the economy (Rubinstein 1981). The Industrial Revolution economy was spectacularly bad at rewarding innovation. This is why Britain has few foundations to rival the great private philanthropies and universities of the USA Its innovators captured little of the rewards.

A similar tale can be told for the other great nexus of innovation in Industrial Revolution England: coal mining, iron and steel, and railroads. Coal output, for example, exploded in England in the Industrial Revolution era. This coal heated homes, made ore into iron, and powered railway locomotives. Yet, there were no equivalents of the great fortunes made in oil, railways, and steel in America's late-nineteenth-century industrialization.

Though the first great innovations of the Industrial Revolution era did not offer much in the way of supernormal profits because of the competitive nature of the industry, the second, railroads, seemed to offer more possibilities. Railways have inherent economies of scale. At a minimum, one line has to be built between two cities. Once it is built, a competitor must enter with a minimum of one complete second line. Since most city pairs could not profitably support multiple links, exclusion and hence profits thus seemed possible.

The success of the Liverpool–Manchester line in 1830 – by the 1840s, shares on this line were selling for twice their par value – inspired a long wave of investment in railways. Figure 8 shows the rapid growth of the railway network in England from 1825 to 1869, by which time more than 12,000 miles of track had been laid across the tiny area of England. This investment and construction was so frenetic that so-called *railway manias* commenced in 1839 and 1846.

But again, the rush to enter quickly drove down profit rates to very modest levels, as Table 8 shows. By the 1860s, real returns, the return on the capital actually invested, were no greater than for very safe investments such as farmland or government debt. While new railway lines initially often had local monopolies, they ended up in constant competition with each other as additional links between nodes in the network were added.

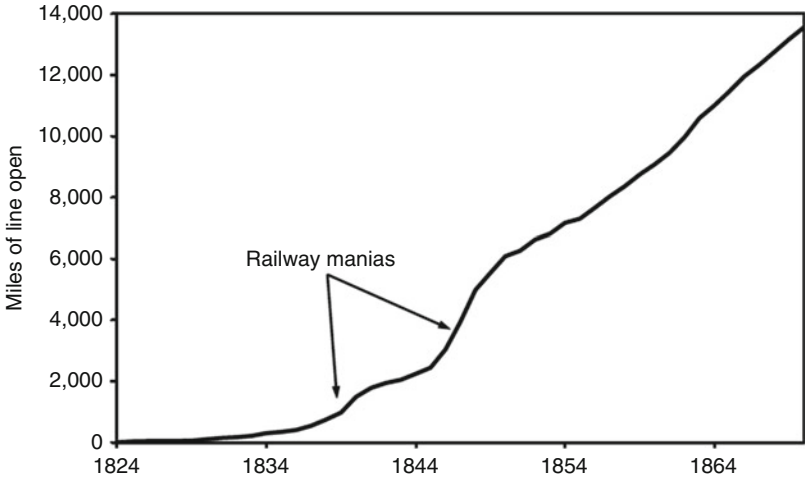


Fig. 8 English railroad construction, 1825–1869 (Source: Mitchell and Deane 1971, p. 225)

Table 8 Profit rates on the capital invested in British owned railways, 1860–1912

Period	Rate of return, UK (%)	Rate of return, British empire (%)	Rate of return. Foreign lines (%)
1860–1869	3.8	–	4.7
1870–1879	3.2	–	8.0
1880–1889	3.3	1.4	7.7
1890–1899	3.0	2.5	4.9
1900–1909	2.6	1.6	4.4
1910–1913	2.6	3.1	6.6

Source: Clark (2007a, Table 14.7)

Thus, while, for example, the Great Western may have controlled the direct line from London to Manchester, freight and passengers could cross over through other companies to link up with the East Coast route to London. Again, profits inspired imitation, which could not be excluded, and any supernormal profits were soon eliminated. Consumers were again the main beneficiaries. It is for this reason that in Britain, unlike in the USA, there are very few universities and major charities funded by private donors.[35] The new industrial priesthood, the engineers who developed the English coalfields, railways, and canals, made prosperous but typically moderate livings. Though their names survive to history – Richard Trevithick, George and Robert Stevenson, Humphrey Davy – they again captured very little of the social rewards their enterprise wrought. Richard Trevithick, the pioneer of locomotives, died a pauper in 1833. George Stevenson, whose famous locomotive *The Rocket* ran loaded at 15 miles an hour (an unheard-of speed for land travel in this era) in a trial in

[35]The industrialization of the United States created much greater private and family fortunes.

1829, did much better. But his country house in Chesterfield was, however, a pittance compared to his substantial contributions to railway engineering. But other locomotives competed in the famous trial, and soon, a swarm of locomotive builders were supplying the railway network. Humphry Davy died rich and celebrated. But he never patented his safety lamp for coal mining, giving the innovation to the industry for free.

Innovation in the Industrial Revolution era typically benefited mainly consumers in the form of lower prices. As coal output exploded, real prices to consumers steadily declined: the real price in the 1700s was 60% greater than in the 1860s. Coal, iron and steel, and rail carriages all remained highly competitive in England in the Industrial Revolution era. The patent system offered little protection to most of the innovations in these sectors, and innovations quickly leaked from one producer to another.

Textiles, the industry with the most dramatic productivity declines, became substantial exporters of products throughout the world, with about half of the production being exported by the end of the Industrial Revolution era. Thus, a large share of the benefits of the Industrial Revolution flowed abroad. This meant that real living standards in England in the Industrial Revolution period grew more slowly than did GDP since the price index for national expenditures grew more rapidly than the price index for national output.

One thing that is striking about the institutionalist explanations discussed above in general is the absence of any agreed metric for institutional quality. There is a belief in the physical sciences that a basic element in any scientific analysis of any phenomenon is to have a defined, objective, and shared system of measurement. There is no agreed metric for institutional quality. Institutionalists on this standard are still in the prescience world of phlogiston and other early theories.

Ideas and the Industrial Revolution

In search of what made England in 1800 different from the Netherlands in 1600 or France in 1800, some scholars have turned to culture. In particular, they have promoted a central role for ideas in germinating technological advance. Margaret Jacob has championed the thesis that the underpinning of technological advance in Industrial Revolution England lay in the grounding of English industrialists, mechanics, and engineers in Newtonian science.

> To be sure some makers of jennies and spindles were semi-literate, more visual than verbal, but by and large, the creators, installers and users of steam and hydraulic presses, the planners and builders of canals - the key players in the British Industrial Revolution - were mechanically literate and in possession of a distinctive cultural persona.[36]

[36]Jacob (2013, p. 8).

Similarly, Joel Mokyr explains the dynamism of the English circa 1800 versus the stasis of the Dutch circa 1650 as

> the advanced technology that helped propel the Dutch economy into unprecedented and even "embarrassing" riches in the seventeenth and eighteenth centuries was still mostly the traditional, pragmatic knowledge at the level of artisans or applied engineers: mechanically clever, well-designed techniques, but without much of an epistemic base in the deeper natural phenomena that made them work. As a consequence, technological progress ran into diminishing returns.[37]

Such propositions about the role of ideas and cultural forms are inherently difficult to test. It is easy, for example, to measure the level of education in different societies at the time of the Industrial Revolution, but measuring the extent to which elites were trained in Newtonian scientific ideas or had grounding in the principles of mechanics is intrinsically much more difficult.

Jacob, for example, has conducted detailed studies of the activities of Industrial Revolution entrepreneurs and innovators, showing their detailed knowledge of mechanical principles and careful attention to detail in introducing such things as steam power. Humphry Davy is one such paradigmatic figure. He grew up in Cornwall, far from major centers of learning, and had very little formal instruction in science, never completing grammar school or attending a university. But he was able to instruct himself utilizing the apparatus and libraries of a circle of local amateurs and professionals who took him under their wing. Thus, early in his career, he made the acquaintance of James Watt and of the Wedgwoods. Although he had no direct experience with the coal industry, when asked to solve the problem of underground explosions of firedamp sparked by the naked flames of lamps, he was able to utilize his scientific experience to rapidly create the Davy lamp in 1815.[38]

But at the same time, in response to the same disastrous underground explosion, George Stephenson devised a safety lamp on other principles. Stephenson had much more humble origins and less impressive scientific credentials. He did not learn to read until age 18. He learned his technical skills on the job as engineman at a colliery. But he became a successful locomotive designer, then railway projector, the father of the railway age.

Which was a more representative figure for the English Industrial Revolution: the scientifically immersed and inspired Davy or the unschooled practitioner Stephenson? Proponents of a culture of Newtonian science or of a British-style Enlightenment can certainly show that some of the leading innovators of the age were immersed in mechanical science and ideas of progress through rationality. But even for these individuals who may acquire more prominence than is their due because of their tendency to join scientific societies, to publish, and to leave private records, it is nigh impossible to demonstrate in most cases that their industrial achievements were the direct result of their scientific interests. Thus, the idea that

[37] Mokyr (1999).

[38] Jacob (2014, pp. 82–84). See also Jacob (1997).

the Industrial Revolution was the product of a particular intellectual culture in England in the eighteenth century is to a large degree untestable given current sources.

How Sudden was the Industrial Revolution? Revolution or Evolution?

One of the things that makes the Industrial Revolution so hard to explain is the apparent suddenness of the arrival of persistent efficiency advance in economies circa 1800. All other elements of the economy were seemingly evolving in a very slow manner in this era – the underlying institutional, political, and social variables were changing slowly if at all in England in the years 1700–1800 – so how could they produce the relatively abrupt change of the Industrial Revolution? See, for example, Fig. 9 showing literacy levels in England, 1580–1920. The Industrial Revolution did see a modest rise in male literacy rates and a more substantial rise for women. But the late nineteenth century was the period of much greater and more dramatic increases in literacy, long after the Industrial Revolution commenced. And for men, there is not much sign of major increases in literacy rates all the way from 1650 to 1800. What is true of literacy is true of many of the underlying variables in the economy in this period: wages, rates of return on capital, land rents, transport costs, population, life expectancy, and so on.

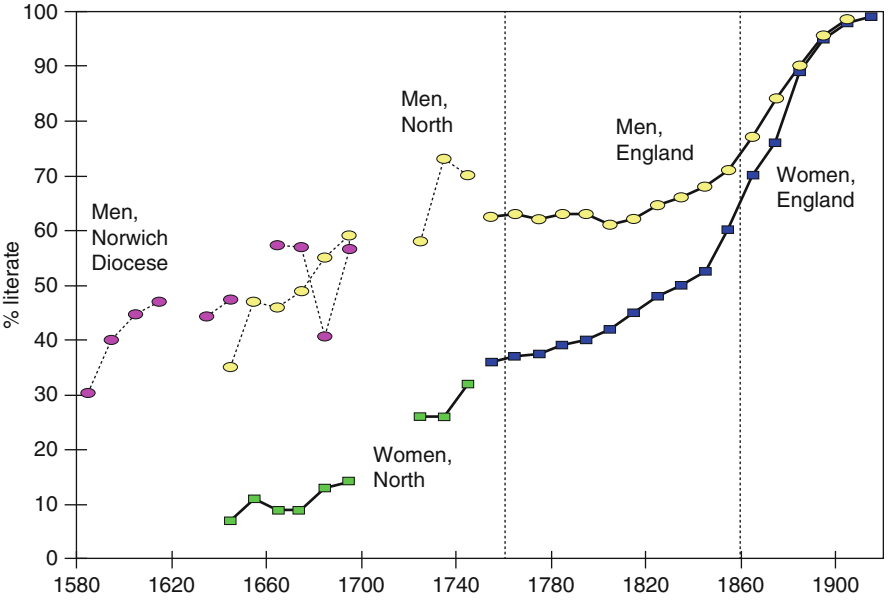


Fig. 9 Literacy in England, 1580–1920 (Source: Clark 2007a, Fig. 9.3, p. 179)

Viewed from the aggregate productivity level of the economy, the conclusion that the transition to modern growth was rapid seems to be at odds with the general historical picture of England from 1200 to 1780. England during this period was a society that was advancing in education, in scientific knowledge, in technical abilities in navigation and warfare, as well as in music, painting, sculpture, and architecture. England in 1780 was a very different place from England in 1250 even if the standard of living of the average consumer measured mainly in terms of their consumption of food, clothing, housing, heat, and lighting had changed little.

The reason for this mismatch is that as noted above in Eq. 4, national productivity growth will be related to productivity advance in individual sectors through

$$g_A = \sum \theta_j g_{Aj} \tag{6}$$

where g_{Aj} is the growth rate of productivity by sector and θ_j is the share of j in total value added in the economy. National efficiency advance is measured by weighting gains by sector with the value of output in that sector. The effects of innovation on national productivity measures is thus crucially dependent on the pattern of consumption.

Much of the technological advance of the period 1250–1780 had minimal impact on measured productivity at the national level because the share of expenditure on these goods was so small in the preindustrial economy. The printing press, for example, led to about a 25-fold increase in the productivity of written material between 1450 and 1600 in England. This was as great an increase in productivity as seen in cotton cloth production 1770–1870. But since the share of income spent on printed materials in the seventeenth century was only about 0.0005, the productivity gains from this innovation at the national level were miniscule (Clark and Levin 2001).

We can see in Fig. 10 that the production of such manufactured items as iron nails and glassware also saw significant productivity advances before 1780. But this efficiency advance would be a negligible contribution to national productivity advance because of the small share of total production value these goods represented in a preindustrial England. Iron nails had limited use, while glasswares were enjoyed only by the richest groups.

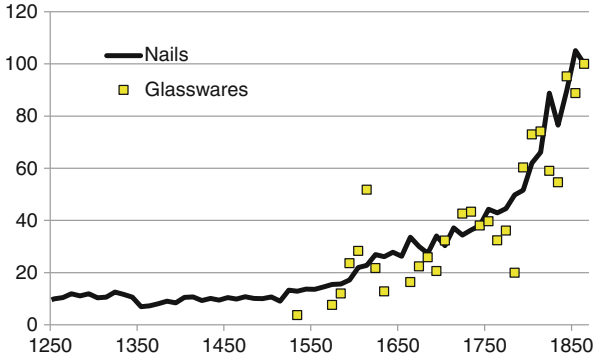


Fig. 10 Efficiency of production of nails and glassware, by decade, 1250–1869 (Source: Clark 2010)

Further, for many goods whose production was becoming more efficient through technological advances, no consistent series of prices can be calculated. There was, for example, a great advance in military technologies in European countries such as England over the years 1250–1780. The infantry of 1780 or a naval ship of that period would have swept the equivalent medieval force from the field. English troops of 1780 would have quickly overwhelmed the fortifications of 1250, but the fortifications of 1780 would have been impregnable even against medieval armies of major size.

For example, the evolution from the medieval crossbow to the arquebus in the late fifteenth century to the musket and then to the rifle in the nineteenth century saw a substantial increase in the firing rate and the force of the projectile. In the sixteenth century, arquebuses could sustain a rate of fire of only one shot every 2 min.[39] By the early nineteenth century, with flintlock muskets, as many as three shots per minute were possible.[40] But none of this would be reflected in conventional productivity measures. There is no allowance in these measures for the delivery of more effective violence by English armies and navies over the years.

There is no allowance also in the national productivity measure for improvements in the quality of literature, music, painting, and newspapers. These sources also do not reflect medical advances such as the one-third reduction in maternal childbirth mortality between 1600 and 1750.[41]

This makes it possible that the rate of technological advance in the economy measured just as a count of innovations and new ideas was actually increasing long before the breakthrough of the Industrial Revolution. But accidents of where these technological advances came in relation to mass consumer demand in the pre-industrial economy create the appearance of a technological discontinuity circa 1780. Suppose that prior to the Industrial Revolution, innovations were occurring randomly across various sectors of the economy – innovations in areas such as guns, gunpowder, spectacles, window glass, books, clocks, painting, new building techniques, improvements in shipping and navigation – but that just by chance, all these innovations occurred in areas of small expenditure. Then, the technological dynamism of the economy would not show up in terms of output per capita or in measured productivity in the years leading up to the Industrial Revolution.

To illustrate this, suppose we consider a consumer whose tastes were close to those of a modern university professor. Their consumption is much more heavily geared toward printed material, paper, spices, wine, sugar, manufactured goods, light, soap, and clothing than the average consumer in the preindustrial English economy. Based on their consumption, how would the efficiency growth rate of the economy 1250–1769 look compared to 1760–1869 and 1860–2009? Figure 10 shows the results, where efficiency is measured as an index on a log scale on the vertical axis so that the slope of the line measures the rate of efficiency growth. Thus,

[39]Shineberg (1971, p. 65).

[40]Townsend (1983, p. 6).

[41]Wrigley et al. (1997, p. 313).

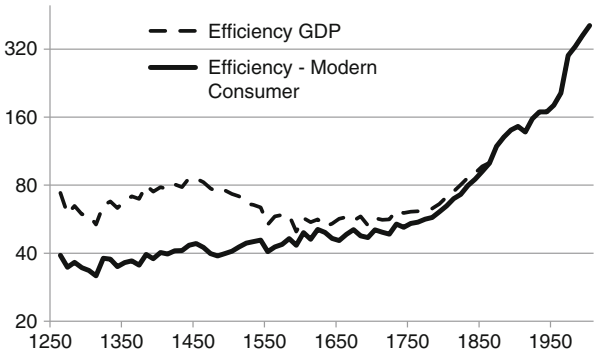


Fig. 11 Economic efficiency from the perspective of a modern consumer, England, 1250–2009. *Notes*: The weights in consumption for the modern consumer are assumed to be half from the consumption basket of the pre-industrial worker. But the other half is composed of books (.1), manufactured goods (.1), clothing (.1), sugar (.03), spices (.03), drink (.05), light (.05), soap (.02), and paper (.02) (Source: Clark 2010)

the upward slope of the line indicates efficiency growth rates. Now, in the years 1300–1770, there is an estimated efficiency growth rate of 0.09% per year for the goods consumed by a university professor. This is followed by efficiency growth rates of 0.6% per year 1760–1870 and 0.9% a year for 1860–2010. Estimated efficiency advance is still very slow for the preindustrial period, but there is a more than 50% increase in efficiency between 1300 and 1770. And this still excludes many of the gains that were discussed above. Thus, we can think of the economy in this period as going through a more protracted transition between preindustrial growth rates and modern growth rates (Fig. 11).

Framed in this way, the possibility opens of some more gradual transition to higher rates of technological advance starting in the medieval period or earlier. We can conceive of the Industrial Revolution as a more evolutionary affair with roots earlier than 1780. We can also think of the Dutch of the seventeenth century as having achieved significant technical progress though in areas such as painting, which leave little trace in aggregate measures of the efficiency of the economy.

Changes in People

Two things suggest that we should perhaps look at changes in people as the wellspring of the Industrial Revolution. The first is the lack of institutional or social barriers to innovation even in medieval England. The medieval economy was largely already a fairly laissez-faire system with modest taxation and few effective religious or social impediments to technological change. The second is the modest signs of any increase in returns to innovation at the time of the Industrial Revolution. If the barriers to innovation were unchanged and the financial rewards in England still modest and no greater than in seventeenth-century Holland or eighteenth-century

France, perhaps the transition was instead driven by changes in the aspirations and capabilities of economic agents.

We certainly see evidence within England that the behavior of economic agents had changed in significant ways between 1200 and 1800. Four changes stand out: a decline in impatience as revealed by a significant decline in the underlying interest rate, an increase in literacy and numeracy, a decline in interpersonal violence, and an increase in work hours. The levels of literacy and numeracy were high by the standards of the preindustrial world. Even the great civilizations of the past such as the Roman Empire or the city-states of the Italian Renaissance had general levels of literacy and numeracy that were surprisingly low by the standards of northwest Europe on the eve of the Industrial Revolution. Though, as noted in Table 6 above, while we can distinguish Industrial Revolution England in terms of literacy and numeracy from earlier societies and even from Japan and China in 1800, we cannot use such measures to explain why England, among many western European societies, was the creator of the Industrial Revolution.

Another caveat about the role of numeracy and literacy in the Industrial Revolution is that given the observed rates of return to schooling, the increased investment in countries like England in the Industrial Revolution period can account little for faster productivity growth rates. Thus, we can modify Eq. 1 to allow for investment in human capital to

$$g_y = a_k g_k + a_h g_h + g_A \tag{7}$$

where a_h is the share of income attributable to human capital investments and g_h is the growth rate of the stock of human capital. But the growth rate of the human capital stock in England from 1760 to 1860 implied by Fig. 9 is very modest: less than 0.4% per year. And even if we allowed one-third of all the 60% share of wage payments in income in Industrial Revolution England to be attributed to human capital, this would entail that human capital investments increased income growth rates by a mere 0.08% per year. If human capital lies at the heart of the Industrial Revolution, it must be because there are significant external benefits associated with human capital investments, as Lucas (1988) hypothesized.

We find interesting evidence that the average numeracy and literacy of even rich people in most earlier economies was surprisingly poor. A prosperous landowner in Roman Egypt, Isidorus Aurelius, for example, variously declared his age in legal documents in a less than 2 year span in 308–309 AD as 37, 40, 45, and 40. Clearly, Isidorus had no clear idea of his age. Other sources show he was illiterate (Duncan-Jones 1990, p. 80). A lack of knowledge of their true age was widespread among the Roman upper classes as evidenced by age declarations made by their survivors on tombstones. In populations where ages are recorded accurately, 20% of the recorded ages will end in 5 or 10. We can thus construct a score variable Z which measures the degree of "age heaping," where $Z = \frac{5}{4}(X - 20)$ and X is the percentage of age declarations ending in 5 or 10 to measure the percentage of the population whose real age is unknown. Z measures the percentage of people who did not know their true age, and this correlates moderately well in modern societies also with the degree of literacy.

Table 9 Age heaping

Place	Date	Type of community	Innumeracy rate
Ancient Rome	1–300	All	46
Medieval England	1270–1370	Landowners	61
Town of Florence	1427	Urban	32
Florentine Territory	1427	Rural	53
Corfe Castle, England	1790	Urban	8
Ardleigh, England	1796	Rural	30

Sources: Clark (2007a, Table 9.4, p. 178)

Among those wealthy enough to be commemorated by an inscribed tombstone in the Roman Empire, typically half had unknown ages. Age awareness did correlate with social class within the Roman Empire. More than 80% of officeholders' ages seem to have been known by their relatives. We can also look at the development of age awareness by looking at a census of the living, as in Table 9. Some of the earliest of these are for medieval Italy including the famous Florentine *Catasto* of 1427. Even though Florence was then one of the richest cities of the world and the center of the Renaissance, only 68% of the adult city population knew their age. Medieval England had even lower age awareness. The medieval Inquisitions post mortem, which were enquiries following the death of landholders holding property where the King had some feudal interest, show that the exact ages of the heirs to property was known in only 39% of cases. In comparison, a 1790 census of the small English borough of Corfe Castle in Dorset with a mere 1,239 inhabitants, most of them laborers, shows that all but 8% knew their age. In 1790, awareness correlates with measures of social class, universal knowledge among the higher-status families, and lower age awareness among the poor. But the poor of Corfe Castle or Ardleigh in Essex had as much age awareness as officeholders in the Roman Empire.

Another feature of the Roman tombstone age declarations is that ages seem to be greatly overstated for many adults. Thus, while we know that life expectancy in ancient Rome was probably in the order of 20–25 at birth, tombstones record people as dying at ages as high as 120. For North African tombstones, for example, 3% of the deceased are recorded as dying at age 100 or more.[42] Almost all of these 3% must have been 20–50 years younger than was recorded. Yet, their descendants did not detect any implausibility in recording these fabulous ages. In contrast, the Corfe Castle census records a highest age of 90, well within the range of possibilities given life expectancy in rural England in these years.

Why then did education levels rise in the centuries leading up to the Industrial Revolution? A theme of many of the previously mentioned economic models of the transition from Malthusian stagnation to modern growth is that there was a switch from quantity, or at least desired quantity, to quality in families as we moved to the modern world (see, example.g., Galor and Weil 2000; Galor 2011). This theme has been driven by the observation in modern cross sections, looking across countries,

[42]Hopkins (1966, p. 249).

that high-income, high-education societies are those with few children per woman. Also, within high-income societies, there was a period between 1890 and 1980 where lower-income families were those with more children.

Such theories face a number of challenges in modeling the actual world of Industrial Revolution England. The first challenge is that these theories are expressed always in terms of children surviving to adulthood. In the modern world, in most societies, child survival rates are high, and so, in practice, births and surviving children are closely equivalent. But in all known preindustrial societies including preindustrial England, large numbers of children did not survive even to their first year. In these cases, the distinction between births and surviving children becomes important. Measured in terms of births, Malthusian societies witnessed high fertility, with the average woman surviving to age 50 giving birth to 5 children. But in such societies, the average number of children surviving to adulthood could only be 2.

Further, since children who died in the preindustrial world tended to do so fairly early, the numbers of children in any household at any time in the preindustrial world would typically be 3 or less. For example, of 1,000 children born in England in 1700–1724, nearly 200 would be dead within 6 months (Wrigley et al. 1997). Preindustrial families would look similar to the families of the USA in the high-growth 1950s and 1960s. Preindustrial families thus faced remarkably similar trade-offs between the number and quality of children as do modern families. In some sense, there has been no change in fertility from the preindustrial to the modern world measured in net as opposed to gross terms.

The second challenge these theories face is that in England, the transition from high births per woman to lower levels of births per woman did not occur at the onset of the Industrial Revolution but only 100 years later – in the 1880s, after efficiency growth rates changed fundamentally.[43] Fertility in England did not show any decline at the aggregate level prior to 1880. Indeed, the opposite occurred, as Fig. 12 illustrates. Births per woman and also net fertility rose precisely in the period of the Industrial Revolution in England.

The third challenge is that in cross section in preindustrial England, there was a strong positive association between net fertility and the wealth or occupational status of families. Figure 13, for example, shows by 20 year periods the numbers of children alive at the time wills were made for married men in England between 1520 and 1879, where those leaving wills are divided into wealth terciles defined across the whole sample. The lowest tercile in wealth would still be men of above median wealth at death. Their implied net fertility is similar to that for men as a whole in England, as revealed by Fig. 13. But the men of the top wealth tercile marrying before 1780 were leaving on average 3.5–4 surviving children. The most educated and economically successful men in preindustrial England were those with the largest numbers of surviving offspring. Matching these men to parish records of births shows that this advantage in numbers of surviving children stems largely from

[43]France was the only country to experience a decline in fertility starting in the late eighteenth century, and France of course lagged Britain in terms of the onset of modern growth.

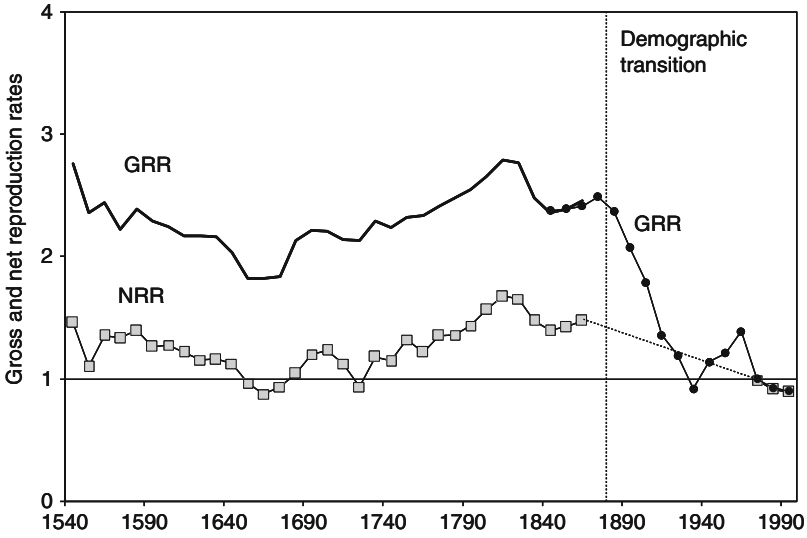


Fig. 12 The fertility history of England, 1540–2000 (Source: Clark 2007a, Fig. 14.6, p. 290)

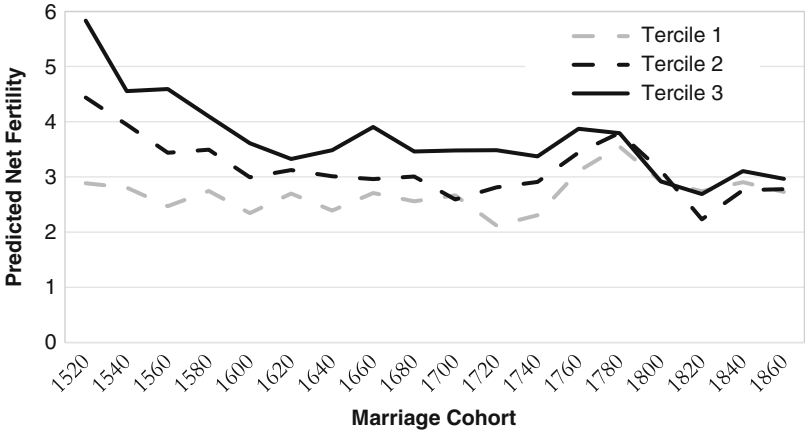


Fig. 13 Net fertility by wealth terciles, marriage cohorts, 1520–1879 (Source: Clark and Cummins 2015)

the greater fertility of the wives of richer men. Their gross fertility was equivalently higher. This positive association of economic status and fertility pre 1780 has been confirmed in an independent study of gross fertility in parish records in England from 1538 to 1837 by Boberg-Fazlic et al. (2011).

For marriages from 1780 to 1879, this pattern of high fertility by the rich and educated is more muted. Instead, we have an interval for most of the Industrial Revolution period where fertility is weakly positively linked to education, status, or wealth. Figure 14 shows the shift in pattern this represents, grouping married men by

Fig. 14 Net marital fertility by wealth decile, marriages 1500–1779 and 1780–1879. Note: The lines at the *top* of the *columns* indicate the 95% confidence interval for the net fertility of these groups relative to the decile of lowest asset income. All assets normalized by the average wage in the year of death from Clark 2010 (Source: Clark and Cummins 2015)

wealth deciles. However, the high overall fertility levels in England at the time of the Industrial Revolution meant that completed family size even for wealthier families continued to be large.

The delay in the decline in aggregate fertility levels in England until after the Industrial Revolution represents a formidable challenge for theories that seek to explain the Industrial Revolution through a quality–quantity trade-off and rising levels of human capital. For it implies that the agents making the Industrial Revolution were still typically from much larger families than in the modern world. Figure 15, for example, shows completed family size for marriages of richer families in England occurring from 1770 to 1879, where by completed size we mean children reaching at least age 21. While the average person in the modern middle class in Europe or the USA has only one sibling or less, this pattern of family sizes implies that the average middle-class person in England born before 1900 had four or more siblings. The engineers and the innovators who made the breakthrough of the Industrial Revolution were typically drawn from families that were very large by the standards of the modern era. Explaining the breakthrough of the Industrial Revolution by the operation of any quality–quantity trade-off in England thus seems a blind alley.

One reason the prosperous of Industrial Revolution England continued to have very large families may be that there is little sign that quantity had much effect on the quality of their children. For these richer families, one measure we have of quality in terms of human capital, though only for boys, is whether they enrolled at Oxford or Cambridge. We can thus estimate the coefficients of the regression

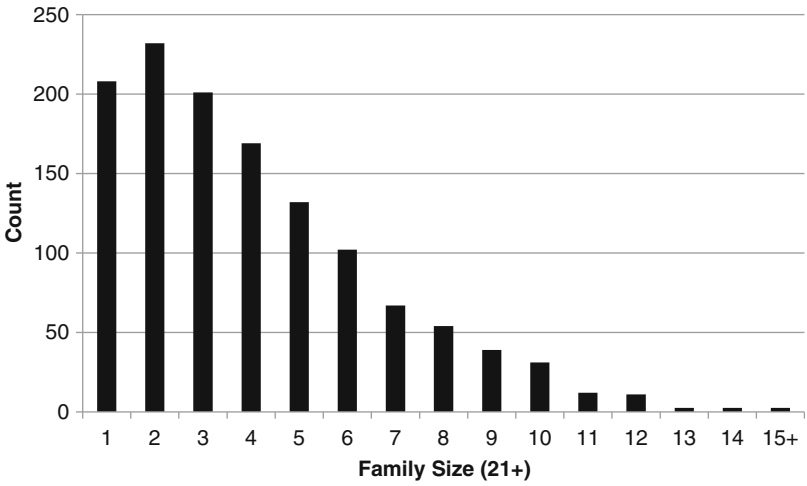


Fig. 15 The distribution of family size in english upper classes, marriages 1770–1879 (Source: Clark and Cummins 2014)

$$\mathrm{DOXB}_{t+1} = a + b_1\mathrm{DOXB}_t + b_2\ln(\mathrm{Wealth})_t + b_3N + e_t$$

where DOXB is an indicator variable for attending Oxford or Cambridge, t is the generation, Wealth is wealth at death, and N is the number of siblings. The estimated values of b_1 and b_2 are both positive and significant. Wealthier fathers who attended Oxford or Cambridge themselves are likelier to have sons who attend. But the coefficient on family size is insignificantly different from 0. There is no sign in education of any quantity–quality trade-off.

Longevity is another measure of child quality. In this sample, for example, sons who attend Oxbridge live on average nearly 4 years longer than those who do not attend. So we can also estimate the effect of family size on longevity, where in this case we can also include daughters. Thus, we estimate the coefficients of the regression

$$\mathrm{AGE}_{t+1} = a + b_0\mathrm{DFEM}_t + b_1\mathrm{PAGE}_t + b_2N + e_t$$

where AGE is child age at death (21+), DFEM is an indicator for a daughter, PAGE is the average of the parents' age at death, and N is again the number of siblings. There is a significant association between PAGE and AGE, but again, family size has no significant effect on age of death of children.

The final measure we have of child outcomes is wealth at death. Here, family size does play a role. We estimate the effect of family size on wealth through estimating the coefficients of the regression

$$\ln W_{t+1} = b_0 + b_1\ln W_t + b_2\ln N + b_3\mathrm{DFALIVE} + e_t$$

where N is family size, W indicates wealth, and DFALIVE is an indicator for when the father is still alive at the time of the son's death. DFALIVE is a control for the

effects of sons who die before fathers and thus likely receive smaller transfers of wealth from fathers. Such sons will also tend to be younger. And in this data, wealth rises monotonically with age until men are well past 60. With this formulation, b_2 is the elasticity of a son's asset income as a function of the number of surviving children the father left. N varies in the sample of fathers and children from 1 to 17. The coefficient b_1 shows the direct link between fathers' and sons' wealth independent of the size of the fathers' family.

For wealth, the coefficient on numbers of children is negative and strongly statistically significant. For sons, the estimated value is -0.43. However, it is much less in absolute value than -1. This implies that while additional children reduce the wealth at death of each sibling, inheritance cannot be the main force determining wealth at death for these children. Wealthy fathers tended to produce wealthy children independently of the actual expected value of the bequest to the child.

Hence, in the Industrial Revolution era, we find that family sizes for the educated classes remained large but seemingly had little effect on children's outcomes except for a partial dilution of their wealth at death. Thus, there is no sign that the Industrial Revolution was created by any move to reduce family size and correspondingly enhance child quality. The revolution was largely achieved in a world where average family sizes remained large for the educated share of the population.

These facts about the transition from preindustrial to modern fertility in England in the Industrial Revolution era represent a formidable challenge to those trying to model the Industrial Revolution in a child quality–quantity framework. Since some of these patterns such as the strong positive association of wealth and fertility in preindustrial England were discovered only in the last few years, many of these models fail to capture essential features of the fertility transitions (Clark and Cummins 2015; Boberg-Fazlic et al. 2011).

So, if changes in family size were not important, why were economic agents in England more effective after 1800? Clark (2007a) postulated that the excess fertility of the rich in the years 1250–1800 observed in England could itself be a factor changing the characteristics of the population across the preindustrial era. We know that all through English history, there is a strong correlation between children and their parents in economic success. So, as the population over many generations shifted toward the children of those achieving economic success, could this have raised the general economic abilities of the society? A full discussion of this issue would take us too far from the specific issue of the Industrial Revolution. But it remains an intriguing possibility that the societies we observe now all have a *deep history*, a set of social conditions over millennia that continue to exert influence on their current possibilities and abilities.

Conclusion

The Industrial Revolution remains one of history's great mysteries. At one level, the transformation is very clear and easily described. After millennia of extremely low rates of technological progress, England in the Industrial Revolution made the first

great step toward modern rates of technological advance. But why this monumental break with the patterns of millennia occurred on a small island of six million people on the periphery of Europe remains a mystery to this day. Attempts by economists to model this transition in terms of institutions and incentives have been so far largely unsuccessful. Changes in institutions and the incentives they generate seem to play little role in the transformation of England in these years. Such changes would also predict an Industrial Revolution much earlier in the seventeenth century in the Netherlands. There is also little sign of any major changes in the underlying parameters of the economy circa 1780 which would lead to changed behavior by individuals.

There is clear sign in England in the 500 years leading up to the Industrial Revolution that there were changes occurring in the basic behaviors of economic agents, changes that might explain an enhanced rate of technological advance. The underlying interest rate in England fell, for example, from 10% in 1300 to 4% by 1770. But what drove these changes remains mysterious. These changes occurred before any significant decline in realized family sizes for the English upper classes, so they do not represent the quantity–quality trade-off beloved by modern growth theorists. There is an intriguing possibility that they are the result of a logic inherent to the preindustrial Malthusian demographic regime, which predicts that the economically successful in any society will also be the demographically successful. But for this to explain an English Industrial Revolution, it would have to be the case that this process was more advanced in England than in other societies. This is an open-research question. We know, for example, that in Qing China, similar Malthusian demographic processes were at work, but we cannot yet quantify whether they had less force in China than in England.

So, 250 years after its first appearance, the Industrial Revolution remains one of the great puzzles of human history, a challenge for future generations of researchers in cliometric history.

References

Acemoglu D, Robinson JA (2012) Why nations fail: the origins of power, prosperity, and poverty. Crown Publishers, New York

Acemoglu D, Robinson JA, Johnson S (2001) The colonial origins of comparative economic development: an empirical investigation. Am Econ Rev 91:1369–1401

Acemoglu D, Robinson JA, Johnson S (2002) Reversal of fortune: geography and institutions in the making of the modern world. Q J Econ 117:1231–1294

Acemoglu D, Johnson S, Robinson JA (2005) The rise of Europe: Atlantic trade, institutional change and economic growth. Am Econ Rev 95:546–579

Aglionby W (1669) The present state of the United Provinces of the Low-Countries as the government, laws, forces, riches, manners, customes, revenue, and territory of the Dutch in three books: collected by W.A. Fellow of the Royal Society, London 1669. http://gateway.proquest.com/openurl?ctx_ver=Z39.88-2003&res_id=xri:eebo&rft_id=xri:eebo:image:64416

Allen RC (2009) The British industrial revolution in global perspective. Oxford University Press, Oxford

Boberg-Fazlic N, Sharp P, Weisdorf J (2011) Survival of the richest? Testing the Clark hypothesis using English pre-industrial data from family reconstitution records. Eur Rev Econ Hist 15(3):365–392
Bresnahan TF, Trajtenberg M (1996) General purpose technologies: engines of growth? J Econ Ann Econ 65:83–108
Broadberry S, Campbell B, Klein A, Overton M, van Leeuwen B (2014) British economic growth, 1270–1870. Cambridge University Press, Cambridge
Clark G (1996) The political foundations of modern economic growth: England, 1540–1800. J Interdiscip Hist 26:563–588
Clark G (1998) Commons sense: common property rights, efficiency, and institutional change. J Econ Hist 58(1):73–102
Clark G (2005) The condition of the working-class in England, 1209–2004. J Polit Econ 8 113(6):1307–1340
Clark G (2007a) A farewell to alms: a brief economic history of the world. Princeton University Press, Princeton
Clark G (2007b) A review of Avner Greif's, institutions, and the path to the modern economy. J Econ Lit 45:727–743
Clark G (2010) The macroeconomic aggregates for England, 1209–2008. Res Econ Hist 27:51–140
Clark G, Cummins N. (2015) Malthus to modernity: wealth, status, and fertility in England, 1500–1879. J Popul Econ: 3–29.
Clark G, Cummins N (2014) The child quality-quantity tradeoff and the industrial revolution. Working paper, University of California, Davis
Clark G, Jacks D (2007) Coal and the industrial revolution, 1700–1869. Eur Rev Econ Hist 11 (1):39–72
Clark G, Jamelske E (2005) The efficiency gains from site value taxes: the Tithe Commutation Act of 1836. Explor Econ Hist 42(2):282–309
Clark G, Levin P (2001) How different was the industrial revolution? The revolution in printing, 1350–1869. Working paper, University of California, Davis
Crafts NFR (1985) British economic growth during the industrial revolution. Oxford University Press, New York
Crafts NFR, Harley CK (1992) Output growth and the industrial revolution: a restatement of the Crafts-Harley view. Econ Hist Rev 45:703–730
De Vries J (1978) Barges and capitalism: passenger transportation in the Dutch Economy, 1632–1839. A.A.G. Bijdragen no. 21. Wageningen
De Vries J (2000) Dutch economic growth in comparative historical perspective, 1500–2000. De Economist 148:443–467
De Vries J, van der Woude AM (1997) The first modern economy. Success, failure, and perseverance of the Dutch economy from 1500 to 1815. Cambridge University Press, Cambridge
Deane P, Cole WA (1962) British economic growth 1688–1959. Cambridge University Press, Cambridge
DeLong BJ, Shleifer A (1993) Princes and merchants: European city growth before the industrial revolution. J Law Econ 36:671–702
Duncan-Jones R (1990) Structure and scale in the roman economy. Cambridge University Press, Cambridge
Freist D (2012) The "Dutch Century." In: European History Online (EGO). Leibniz Institute of European History (IEG), Mainz
Galor O (2011) Unified growth theory. Princeton University Press, Princeton
Galor O, Weil DN (2000) Population, technology and growth: from malthusian stagnation to the demographic transition and beyond. Am Econ Rev 90:806–828
Ganea P, Pattloch T (2005) Intellectual property law in China, vol 11, Max Planck series on Asian intellectual property law. Kluwer, New York
Greif A (2006) Institutions and the path to the modern economy: lessons from medieval trade. Cambridge University Press, Cambridge

Harley CK (1998) Cotton textile prices and the industrial revolution. Econ Hist Rev 51(1):49–83
Harley CK (2010) Prices and profits in cotton textiles during the industrial revolution. University of Oxford discussion papers in economic history, #81
Hopkins K (1966) On the probable age structure of the Roman population. Popul Stud 20(2):245–264
Jacob M (1997) Scientific culture and the making of the industrial West. Oxford University Press, Oxford
Jacob M (2013) How to think about culture in relation to economic development. Working paper, LSE
Jacob M (2014) The first knowledge economy: human capital and the European economy, 1750–1850. Cambridge University Press, Cambridge
Khan Z (2008) An economic history of patent institutions. In: Whaples R (ed) EH.Net encyclopedia. http://eh.net/encyclopedia/an-economic-history-of-patent-institutions/
Lindberg E (2009) Club goods and inefficient institutions: why Danzig and Lübeck failed in the early modern period. Econ Hist Rev N Ser 62(3):604–628
Long P (1991) Invention, authorship, 'intellectual property', and the origin of patents: notes towards a conceptual history. Technol Cult 32:846–884
Lucas R (1988) On the mechanics of economic development. J Monet Econ 22:3–42
Malthus TR (1798) An essay on the principle of population. J. Johnson, London
McCloskey DN (1981) The industrial revolution: 1780–1860, a survey. In: Floud R, McCloskey D (eds) The economic history of Britain since 1700. Cambridge University Press, Cambridge, pp 103–128
Mill J (1821) Elements of political economy. Baldwin, Cradock and Joy, London
Mitchell BR (1988) British historical statistics. Cambridge University Press, Cambridge
Mitchell BR, Deane P (1971) Abstract of British historical statistics. Cambridge University Press, Cambridge
Mokyr J (1999) The industrial revolution and the Netherlands: why did it not happen? Prepared for the 150th Anniversary conference organized by the Royal Dutch Economic Association, Amsterdam, 1999
Mokyr J (2003) Long-term economic growth and the history of technology. In: Aghion P, Durlauf S (eds) Handbook of economic growth. Elsevier, Amsterdam
Mokyr J (2012) The enlightened economy. An economic history of Britain 1700–1850. Yale University Press, New Haven
North DC (1981) Structure and change in economic history. Norton, New York
North DC (1994) Economic performance through time. Am Econ Rev 84(3):359–368
North DC, Thomas RP (1973) The rise of the western world. Cambridge University Press, Cambridge
North DC, Weingast BR (1989) Constitutions and commitment: evolution of institutions governing public choice in seventeenth century England. J Econ Hist 49:803–832
North DC, Wallis JJ, Weingast BR (2012) Violence and social orders: a conceptual framework for interpreting recorded human history. Cambridge University Press, Cambridge
Overton M (1985) The diffusion of agricultural innovations in early modern England: turnips and clover in Norfolk and Suffolk, 1580–1740. Trans Inst Br Geogr N Ser 10(2):205–221
Overton M (1991) The determinants of crop yields in early modem England. In: Campbell BMS, Overton M (eds) Land, labour and livestock. Manchester University Press, Manchester, pp 284–322
Pomeranz K (2000) The great divergence: China, Europe and the making of the modern world economy. Princeton University Press, Princeton
Prak M (1997) Burghers, citizens and popular politics in the Dutch Republic. Eighteenth Century Stud 30(4):443–448
Reis J (2005) Economic growth, human capital formation and consumption in Western Europe before 1800. In: Allen RC, Tommy B, Martin D (eds) Living standards in the past: new perspectives on well-being in Asia and Europe. Oxford University Press, Oxford, pp 195–226

Ricardo D (1821) On the principals of political economy and taxation, 3rd edn. John Murray, London
Romer PM (1986) Increasing returns and long-run growth. J Polit Econ 94:1002–1037
Rosenthal J-L (1992) The fruits of revolution, property rights, litigation and French agriculture (1700–1860). Cambridge University Press, Cambridge
Rubinstein WD (1981) Men of property: the very wealthy in Britain since the industrial revolution. Croom Helm, London
Shineberg D (1971) Guns and men in Melanesia. J Pac Hist 6:61–82
Smith A (1776) An inquiry into the nature and causes of the wealth of nations. W. Strahan and T. Cadell, London
Townsend JB (1983) Firearms against native arms: a study in comparative efficiencies with an Alaskan example. Arct Anthropol 20(2):1–33
van Zanden JL (2008) Prices and wages and the cost of living in the western part of the Netherlands, 1450–1800. Working paper, International Institute of Social History, Amsterdam. http://www.iisg.nl/hpw/brenv.php
van Zanden JL, Prak M (2006) Towards an economic interpretation of citizenship: the Dutch Republic between medievalcommunes and modern nation-states. Eur Rev Econ Hist 10(2):111–145
Vries PHH (2001) Are coal and colonies really crucial? Kenneth Pomeranz and the Great Divergence. J World Hist 12(2):407–446
Wicker ER (1957) A note on Jethro Tull: innovator or crank? Agric Hist 31(1):46–48
Wrigley EA (1988) Continuity, chance and change. Cambridge University Press, Cambridge
Wrigley EA, Davies RS, Oeppen JE, Schofield RS (1997) English population history from family reconstruction: 1580–1837. Cambridge University Press, Cambridge/New York

The Antebellum US Economy

Gavin Wright

Contents

Abstract

In the antebellum era, the economy of the United States underwent an acceleration of economic growth. Cliometric studies have established not only that growth predated the Civil War but that many features prominent in later years – commercialization of agriculture, urbanization, the rise of manufacturing, and mass European immigration – were clearly visible in the antebellum period. This chapter reviews and summarizes this body of research. A notable dimension of this history is that growth and development transpired under two distinct regimes: the slave economy of the southern states and the family-farm, wage

G. Wright (✉)
Stanford University, Stanford, CA, USA
e-mail: write@stanford.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_65

labor economy of the free states. This experiment in comparative development has also been a focus of cliometric attention and thus provides a second theme for the essay.

Keywords
Economic history · Economic growth · Antebellum · Transportation · Technology · Slavery

Introduction

From the Constitution of 1789 to the secession crisis of 1861, the economy of the United States underwent an acceleration of growth, reaching annual rates of 1.7% per capita by the late antebellum decades. Cliometric studies have been instrumental in establishing not only that growth predated the Civil War but that many features of the growth process prominent in later years – commercialization of agriculture, urbanization, the rise of manufacturing, mass European immigration, and rapid adoption of new technologies – were clearly visible in the antebellum period. This chapter reviews and summarizes this body of research. Because Douglass North's *The Economic Growth of the United States, 1790–1860*, first published in 1961, was something a landmark in the rise of cliometrics; that book proves helpful in defining the issues and consolidating the results of more than a half-century of scholarship since then.

A notable dimension of the antebellum US economy is that growth and development transpired under two distinct regimes: the slave economy of the southern states and the family-farm, wage labor economy of the free states to the north. Although both regions shared in the antebellum growth spurt, patterns of geographic settlement and investment were strikingly different between them, with important implications for subsequent eras. This historical experiment in comparative economic development has also been an important focus of cliometric attention and thus provides a second theme for this essay.

Estimates of Gross Domestic Product, 1790–1860

Compiling estimates of national income and product over extended historical periods was one of the earliest and most important contributions of cliometrics. Simon Kuznets received a Nobel Prize for developing the modern framework for national income accounts in the 1930s, extending his estimates back to 1869 in subsequent work. Robert E. Gallman, a former student of Kuznets, constructed new estimates for 1834–1909, using figures for commodity production (agriculture, mining, and manufacturing) from the federal censuses of 1839, 1849, and 1859 as "major" benchmarks and extending the series to other years on the basis of state censuses (Gallman 1960, 1966). The results clearly showed that high rates of per capita growth predated the Civil War.

Gallman's figures had a solid basis in the national surveys of agriculture and manufacturing that began with the US Census of 1840 (reporting economic data for 1839). Estimates for early years have to be derived from much more fragmentary sources, prompting Paul David to label this era a "statistical dark age" (David 1967). David constructed new "conjectural" estimates by combining evidence on average agricultural productivity from Towne and Rasmussen (1960) with estimates of the agricultural labor force from Stanley Lebergott (1964) and assuming that the productivity ratio between the two sectors was constant across the period. The results showed little or no growth between 1800 and 1820, followed by a sharp spurt between 1820 and 1840. Alternative labor force figures developed by Thomas Weiss (1992, 1993) and supplemented by evidence from Kenneth Sokoloff (1986) on productivity growth in manufacturing pointed to a more gradual acceleration of growth between 1800 and 1840.

The subsequent literature raised a number of questions about the underlying evidence and interpretation of these estimates. Nancy Folbre and Barnet Wegman (1993) noted that the conventional definition of GDP descended from Kuznets excludes housework, childcare, and other nonmarket outputs chiefly (though not exclusively) produced by women and that women also contributed to marketed production on farms, though they are excluded from the labor force by convention. These concerns are particularly pressing for this historical era, when farms were organized as family enterprises and a large fraction of their produce did not pass through markets. Thus Weiss presents two sets of estimates of farm productivity, a "broad" output measure that includes home manufactures and improvements and a "narrow" version that counts only the market value of farm products. In 1800 the gap between the two was 17.5%, and by 1900 it had virtually disappeared. Using market values may be said to overstate growth, because some of the expansion represents a resource transfer from nonmarket to market activity. But because commercialization was itself a contributor to growth, neither alternative is unambiguously "correct." Both series, it should be noted, show an acceleration of productivity growth after 1850.

Another interpretation is suggested by a later set of estimates by David (1996). Abandoning the notion of a productivity "gap" between agriculture and manufacturing as inappropriate for the early American land-abundant setting, David constructed detailed estimates of hours of labor by sector. His conclusion is that more than half of the average annual growth rate of 0.9% between 1800 and 1840 was attributable to an increase in hours worked per year: a labor force "working harder" rather than "working smarter."

A genuine infusion of new evidence into this context was provided by Joseph H. Davis (2004), who compiled an annual index of US Industrial Production beginning in 1790. Drawing upon previously untapped sources such as state inspection records and local trade organization reports, Davis collected a total of 43 quantity-based series for mining and manufacturing, most entirely new. Many of these date from 1790, such as copper smelting, firearms, fire engines, shipbuilding, newspaper publishing, and pipe organs, while others enter the index at later dates. The Davis index shows almost continuous growth from 1790 onward. It is of

relatively limited direct value in estimating GDP per capita, because the index measures total production rather than productivity and because nonagricultural occupations (only partially covered by "industrial production") accounted for less than 25% of total employment in 1800. Nonetheless, the series has been used to interpolate GDP estimates based on benchmark dates and to improve knowledge of the timing and magnitude of fluctuations in economic activity (Fig. 1). Even the list of new industries entering the index (presumably because production first began or first became notable enough to be counted) such as whale oil refining (1793), salt production (1797), milled wheat flour (1798), fish curing (1804), gunpowder (1804), wool stockings (1808), hog packing (1809) make it evident that an economic development process was underway.

The most recent attempt to re-estimate antebellum growth is by Peter H. Lindert and Jeffrey G. Williamson (2016). Whereas virtually all other estimates are constructed directly or indirectly from the production side, Lindert and Williamson instead build their new figures from the income side, using the venerable methodology of "social tables": average incomes for various classes of society, such as officials, merchants, artisans, farm operators, and laborers. The authors draw upon the federal tax assessment of 1798 to estimate incomes for 1800 and then use the federal censuses to generate new figures for 1850 and 1860. Thus, they cannot address the debate about the timing of growth acceleration. But the new estimates confirm the finding of strong per capita growth across the entire period. Indeed, their point estimate of 1.42% average annual growth between 1800 and 1860 is higher than any previous estimate and higher than the growth rate of any other country during this period. The reasons for these differences are not clear, but caution is advisable, because the estimates by other scholars and for other countries were not generated by similar methods and may not be conceptually comparable. A prudent assessment would be that Lindert and Williamson confirm the broad consensus

Fig. 1 The left-hand axis is in 1996 dollars. The right-hand axis is an index set at 1849–1850 = 100, divided by the resident population. (Source: Carter et al. (2006), Series Ca11, Ca14, Ca19. The original source is Davis (2004))

among cliometricians that aggregate US production and incomes per capita grew at accelerating rates prior to the Civil War, putting the country on a path toward world leadership by the end of the century.

Napoleonic Wars, Embargo, and the War of 1812

Within 3 years of the new constitution, economic activity along the Atlantic seaboard began to accelerate. The problem with relating institutional to economic change is that the primary stimulus was external: the advent of the long period of European conflict known as the Napoleonic Wars, during which American shippers and shipbuilders took advantage of the nation's neutral status. As North wrote: "One need look no further than to events in Europe to account for nearly every twist and turn in the fortunes of the American economy during these years" (1961, p. 36). The growth in US foreign earnings came primarily from the carrying trade (shipping services), shipbuilding, and reexports of goods from the West Indies and other tropical sources (Fig. 2).

Cliometricians have debated North's assertion that "the years 1793–1808 were years of unparalleled prosperity" (p. 53). Claudia Goldin and Frank Lewis (1980) estimated that the surge of foreign demand across these years augmented US growth by an average of 0.22 percentage points. In contrast, Donald Adams (1980) noted that this period saw no sustained increase in real domestic imports per capita, perhaps because higher freight rates caused by pressures on capacity acted as a tax on American consumers. In the absence of reliable annual income figures, this debate may be impossible to resolve. But, North also argued that the period of seaboard prosperity fostered long-term developments in financial and urban infrastructure, laying "important foundations for the growth of the economy after 1815"

Fig. 2 The vertical axis is in millions of dollars. (Source: Historical Statistics of the United States, Colonial Times to 1970 (1975): Series U190, 192, 193)

(p. 53). Examples include the rise of banking and insurance companies, much of which serviced the shipping industry, and construction of warehouses and turnpikes, driven by growth of the urban population. Richard Sylla (1998) has shown that capital markets were active in Boston, New York, Philadelphia, and Baltimore far earlier than previously believed, with a strong surge between 1797 and 1811. It seems evident that the capital demands of shipping and urban growth were important forces behind this development.

What we do know is that the era of open trade with the warring nations came to a crashing close with Jefferson's Embargo of December 1807, which closed down all foreign trade. This effort to assert American rights on the high seas gave way to agreements that partially restored trade over the next 4 years, but the seaboard economy never returned to its former glories, and ultimately the US itself entered the conflict by declaring war on Britain in June 1812. From an economic standpoint, 15 years of open economy were followed by 8 years of closed economy, the most severe isolation coming during the later war years, when Britain blockaded the entire US coast. Davis and Irwin (2003) show that the composition of US industrial production radically shifted, from trade-dependent industries such as shipbuilding (which contracted) to domestic infant industries such as cotton textiles (which boomed).

Joshua Rosenbloom (2004) makes the case that the rise of the antebellum US cotton textile industry exemplifies a "path-dependent" historical process, whereby the 8-year break from foreign competition triggered developments with enduring consequences long afterward. The discontinuity in industry growth supports this interpretation: Even after the transfer of Arkwright's spinning technology to America by Samuel Slater in 1790, by 1807 the United States still had only 8,000 spindles; that number jumped to 31,000 in 1809 and to 300,000 by the war's end in 1815. Despite the devastating effects of renewed British imports in 1816, the industry's capacity never declined thereafter; within a few years, growth had resumed.

An important question for this thesis is: What was the basis for the industry's survival and growth in the postwar era? One set of factors may be classified as "learning": gains in skills and experience on the part of workers, managers, and machine-makers. We have little detailed knowledge of these matters for this early period, but we can point to one technological innovation that was critical to the industry's success in later years: adoption and improvement of the power loom by mechanic Paul Moody, who was employed by Boston merchant Frances Lowell at the Boston Manufacturing Company. The BMC was established in 1813 during the wartime boom, and its business strategy embodied features that became hallmarks of a distinctive American approach: large capitalization, integrated spinning and weaving under one roof, and commitment to long production runs of uniform products, coarsely woven but durable. The firm also innovated in its labor policies, recruiting young, unmarried women from the New England countryside, to be well-housed and supervised in dormitories under what was known as the "Massachusetts system." All of these innovations may be understood as adaptations of British technologies to the labor-scarce, resource-rich American environment.

A difficulty in presenting this case as a clear illustration of the "infant industry effect" is the postwar competitive setting, which was also reshaped by new protective tariffs, beginning with the Tariff of 1816 – another aspect of Rosenbloom's case for path dependence. Although New England's representatives favored free trade before the war, protests from failing textile owners pushed many of them toward protection in that year. One of the leading voices was Frances Lowell himself, who was instrumental in enacting a "minimum valuation" on imports of 25 cents per square yard. Because this principle effectively constituted a prohibitive tax on Asian cloth made of Asian cotton, the measure won a number of Southern votes. The industry thus received several more years of breathing space in their favored product lines (Temin 1988). Tariffs were increased still further during the 1820s, as emerging patterns of regional specialization became clearer, pushing erstwhile supporters of mercantile interests (such as Daniel Webster) into the protectionist camp. The "Tariff of Abominations" in 1828 and the Nullification Crisis of 1832 made clear that the tariff had become a major issue of regional conflict.

Transportation Revolution

The dramatic fall in the costs of transportation, associated with rapid extension of settlement and markets, was a defining feature of the antebellum economy. This phenomenon was central to the interpretation of Douglass North, who called it "the most pervasive influence affecting the. . .shift from pioneer self-sufficiency to a market oriented agriculture" (1961, p. 143). An interest in the role of government in promoting such "internal improvements" was in the air during the 1950s, as economic historians sought to explore the lessons of the US experience for countries pursuing development programs in our own times. Carter Goodrich of Columbia University should be considered one of the founders of cliometrics, in that his students H. Jerome Cranmer, 1960, and Harvey H. Segal, 1961, assembled statistical evidence to track the flow of investments in canals over time and to estimate the benefits of those investments for the economy (Cranmer 1960; Goodrich 1961). This was the intellectual background for the two landmark railroad books by Albert Fishlow (1965) and Robert Fogel (1964). (Fogel's book dealt primarily with the postbellum economy and hence will not be reviewed here.)

Turnpikes

The drive to improve transportation long predated the large-scale canal projects of the 1820s and the 1830s. Beginning in the 1790s, a network of toll roads known as "turnpikes" crisscrossed the northern and middle states, constructed in most cases by corporations chartered by the states for specific ventures. More than 70 such companies were created by 1800, rising to as many as 800 by 1830, comprising more than 11,000 miles of roads and accounting for between one-fourth and one-half of all business incorporations in these states. From an economic standpoint, a remarkable

feature of these corporations is that although they were chartered as profit-seeking enterprises, nearly all of them lost money. It is not unusual for entrants into a competitive industry to experience losses, perhaps because of overoptimistic expectations or severe competition. But when hundreds of firms enter an activity with a proven track record of losses, one is driven to the conclusion that they must have known what they were getting into. Despite the adverse prospects, stock subscriptions were typically broad-based in the affected areas. Daniel Klein (1990) describes the ways in which turnpike promoters persuaded community members to invest, invoking the shared interests of farmers, merchants, and land speculators to overcome free rider problems. The patterns illustrate the American "cooperative spirit" that so impressed Tocqueville but also Veblen's characterization of the American town as a form of collusion among "interested parties" whose social cohesion was enhanced by a common interest in its real estate values. Compared to the more densely populated United Kingdom, these "indirect benefits" of transportation investments were far more important in the United States (Bogart and Majewski 2008).

Canals

Turnpikes extended the range of markets, but they could not overcome a basic fact of economic geography for this period: the dramatic 10- to 30-fold differential between ton-mile rates on overland relative to water modes of transportation (Carter et al., vol. 4, p. 781). This gap gave rise to a demand for canals, "artificial rivers" to augment the nation's natural waterways. These larger-scale projects were typically beyond the reach of local governments, and so the states took the lead in the post-1815 era, beginning with New York's 363-mile Erie Canal, linking the Great Lakes to the Atlantic Ocean via the Hudson River. Begun in 1817 and opened on November 4, 1825, the Erie was a decisive turning point in transportation history, spawning as many as a hundred imitative projects and 3,000 canal miles by 1840 (Goodrich 1961).

John Wallis and Barry Weingast (2018) theorize that the states played the leading role in the antebellum era because the federal government was locked in a position of "equilibrium impotence": the US Constitution required direct taxes to be allocated on the basis of population, so the federal government was unable to develop a system of "benefit taxation" in which costs were borne by the areas that gained the most from transportation project. As a result, federal infrastructure expenditures tended to be collections of small projects broadly distributed around the country, such as lighthouses and rivers and harbors. In contrast, state governments (who faced similar problems of sectional rivalries) did use schemes that coordinated taxes and benefits. For example, in 1817 New York created a special "canal tax" to be levied on counties along the canal, if additional funds were required to service state bonds. Ohio in 1826 and Indiana in 1836 created new ad valorem property taxes to allocate the burdens more equitably and to align these burdens more closely with increases in land values that were the consequence of state projects. Wallis (2003) shows that Indiana's Mammoth system of canals, railroads, and turnpikes followed directly on this change in public finance.

This political economy analysis should not be understood as an argument that construction of the American canal network was efficient. No state replicated New York's Erie success in generating revenues sufficient to pay off the bonds early – a performance the state subsequently squandered by extending canals through many interior counties. One-fifth of all canal ventures were costly failures, contributing to the wave of defaults on state bonds during the depression in 1839–1843. Even for the survivors, canal abandonments continued during the 1840s, as many were soon rendered obsolete by the railroads. With all of these defects, however, the canal era in many respects launched the nation on the path to economic development. Lee Craig et al. (1998) show that location on a canal or river in 1850 raised land values by $2.68 per acre, a result confirmed by Wallis (2003) for Indiana a decade earlier. Treatment-control and difference-in-difference methods were anticipated by Segal (1961), who showed that canal counties experienced more rapid settlement and occupational shifts into commerce and manufacturing compared to non-canal counties. Andrew Coleman (2012) uses New York and federal censuses to develop a panel of households for 1825 through 1845, confirming that farm families with access to the canal reduced home production of cloth, in favor of commercial activity and factory-made goods. Coleman adds the interesting detail that new arrivals to the canal zone had lower home production activity than long-established residents. Eventually, however, the old-timers copied their neighbors and reduced home production to similar levels.

Steamboats on Western Rivers

An innovation that augmented the shift from land to water transportation was the steamboat, which radically reduced the time and cost of upstream river routes (Carter et al. 2006, vol. 4, pp. 878–880). Neither the steam engine nor the steamboat were US inventions, but from the time of Robert Fulton's commercially successful passenger service on the Hudson River in 1807, steamboating on internal waterways became an American specialty. Fulton and his partner Robert Livingston were granted a monopoly by Louisiana in 1812, but this proved unenforceable even before the Supreme Court confirmed exclusive federal authority over interstate commerce in *Gibbons vs. Ogden* (1824). Thus driven by competitive forces, round-trip travel times between New Orleans and Louisville fell from 30 days in 1815–1819 to 11.7 days in 1860. Haites et al. (1975) estimate that total factor productivity in western river steamboating grew at an average annual rate of 4.6–5.5% over the entire period 1815–1860.

Steamboats became faster and cheaper, but they were also hazardous. Paul Paskoff (2007) tallies 1,166 wrecks on US waters between 1821 and 1860. Highly publicized fatal explosions led to federal regulatory acts in 1838 and 1852, of uncertain effect. Although the absolute number and tonnage of losses increased over time with growing commerce, rates of loss clearly declined. Whereas Haites et al. attributed efficiency gains on the rivers almost entirely to private enterprise, Paskoff builds a persuasive case for the efficacy of the federal government's support

for removal of snags and other natural hazards, under the Rivers and Harbors legislation. This relative success contributed not just to safety but to the growth of total factor productivity by reducing losses, encouraging larger vessel size, and increasing steamboat longevity. As with virtually all phases of the antebellum transportation revolution, a combination of private initiative and public support was at work.

Railroads

US railroad mileage in the antebellum era was small relative to later decades, barely one-sixth of the total for the nineteenth century as a whole. Yet this amount was greater than that of any other country at that time, greater indeed than the rest of the world combined, and its effects were transformative, as recounted in Albert Fishlow's now-classic 1965 book. Digging deeply into industry records as well as government sources, Fishlow assembled detailed new estimates of railroad output, receipts, capital expenditures, employment, and profits and deployed these to address basic historical issues. Among these were the direct benefits of railroad transportation services, forward and backward linkages, and the impact of the railroad on regional patterns of domestic commerce.

Probably the most influential chapter was Fishlow's critique of Joseph Schumpeter's dictum that antebellum railroad building "meant building ahead of demand in the boldest acceptance of that phrase" (quoted in Fishlow 1965, p. 165). Restating this assertion as "an initial disequilibrium that is self-correcting over time. . .by induced shifts in the demand schedule" (1965, p. 166), Fishlow developed a series of tests: ex ante risk premiums for investment, ex post profits, and patterns of sequential construction and governmental assistance. On each count, he found that the hypothesis failed. During the 1850s, railroads first entered counties that already had higher population densities and wheat production than their neighbors. Most were highly profitable from the start; small sections of line typically opened as soon as they were completed, to the cheers of expectant farmers. Most tellingly, governmental aid during this period came mainly from local rather than state or federal sources, from towns and counties eager for a new rail line and fearful that they might be bypassed. An active recruiting and promotional role for such communities is the opposite of what one would expect to see if railroads were built ahead of demand.

There was one complicating factor, however, the phenomenon Fishlow called "anticipatory settlement." Settlers may have been on the ground waiting when the railroads arrived, but their choice of location was heavily influenced by expectations regarding the railroad's path. Anticipating the impact on land values, farmer-speculators wanted to "get in on the ground floor" ahead of their rivals. This consideration undermines the simple conceptualization behind the empirical tests, the notion that the issue could be answered on the basis of timing. As Fishlow wrote: "In a larger sense, therefore, railroads *were* a leading sector. . .[but] this interpretation takes us far from a simple Schumpeterian world. Indeed, it almost turns that

model on its head: instead of a heroic role for the railroad investor or even the state, the beneficiary of railroad construction displays the crucial attributes of foresight" (1965, pp. 165, 197).

Many subsequent cliometric studies have pursued issues related to antebellum railroads. The most recent use geographic information system (GIS) databases drawn from digitized historical travel guides, allowing authors to pinpoint railroad locations and timing far more precisely. Atack et al. (2010) deploy difference-in-difference methods to compare railroad with non-railroad counties, showing that the entry of railroads had a small effect on population density but a large impact on urbanization. As a robustness check, the authors use instrumental variables based on early (pre-railroad) government surveys of potential transportation routes. Atack and Margo (2011) extend this methodology to document an effect on agricultural improvement and farmland values.

Note that these questions differ from Fishlow's. Fishlow asked whether railroads were built before they were commercially viable, i.e., whether they induced settlement in a dynamic sense, while the later scholars set out to identify the "treatment" effect of railroads on economic outcomes. Their approach is to look for "credible exogenous variation" in rail access, and for this purpose, endogeneity in the geography of rail location is a source of bias to be corrected through instrumentation. Indeed, the first stage IV regressions confirm the very endogeneity that Fishlow described – biasing the DID coefficients downward – but the objective of the articles is to purge this effect rather than to analyze it. The new results would not surprise Fishlow, who argued throughout that antebellum railroads were transformative.

Atack et al. (2014) come closer to capturing Fishlovian dynamics in their econometric procedure. These authors link the GIS database to a census of banking, exploring the detailed timing of railroad construction and bank entry. They show first that railroads tended to link existing financial-commercial centers but next that new bank entry quickly followed (within 1–3 years) in the wake of rail construction. Their conclusion is that "financial and real factors interact[ed] to put a virtuous cycle of economic development in motion" (p. 943). The deeper implication is that the grass-roots foundations for transportation improvements were broadly similar from the turnpikes to the railroad era.

Productivity Growth in Agriculture

Because the great majority of the antebellum labor force was in farming, acceleration in growth for the economy as a whole would have to entail productivity growth in agriculture. Extending a methodology utilized by the US Department of Agriculture, William N. Parker developed productivity estimates for individual crops, by decomposing output per laborer into labor requirements per acre (distinguishing preharvest, harvest, and postharvest labor) and yields per acre. The results clearly pointed toward mechanization as the primary source, as proxied by the increase in acres per laborer (Parker and Klein 1966). Although the research sought to track

productivity growth over the long period from 1840 to 1910, it is evident that the trends toward mechanization and larger acreages were underway well before the Civil War.

One of the most widely discussed studies in cliometrics is Paul David's analysis of the diffusion of the mechanical reaper in antebellum Illinois (David 1966). Although the machines invented by Obed Hussey and Cyrus McCormick were available as early as 1833 or 1834, they were not widely adopted until the mid-1850s. David's explanation for this 20-year lag is that most Midwestern grain farms were too small in harvestable acreage to justify the fixed cost of purchasing a reaper. David estimated the threshold scale for reaper profitability as 46.5 acres using the factor prices of 1849–1853, whereas Illinois farms in 1850 averaged only 15 or 16 acres of wheat, oats, and rye. Between 1850 and 1860, however, harvest acreage grew as farm families continued their land-clearing operations, while the threshold size fell as reaper prices declined and the wages of harvest labor rose during the boom decade of the 1850s. The transversal of a falling threshold and a rising farm size distribution generated the sigmoidal adoption pattern commonly seen in diffusion studies.

Critiques across several decades have revised nearly all parts of this narrative. Alan Olmstead (1975) showed that the threshold calculation is highly sensitive to parameter assumptions, so that modest adjustments nearly doubled the estimated crossover point. Drawing on a sample of nearly 12,000 farms from the 1860 census, Jeremy Atack and Fred Bateman calculated a range of thresholds and found that only 41,000 reapers would have been profitable in that year, or 150,000 reaper-mowing machines (1987). Yet, Olmstead showed that as many as one-fourth of the reapers sold by McCormick were purchased jointly by two or more individuals, indicating that sharing arrangements were not infeasible. Olmstead's explanation for the lag in adoption was that the performance and reliability of the reapers improved considerably across the 20-year period. Olmstead and Rhode (1995) presented new evidence based on more than 1,000 reaper purchasers matched to the manuscript censuses. Their data showed that although the relative propensity to buy a reaper increased with small-grain acreage, the pattern was by no means discontinuous at a specific threshold. The authors went on to question the threshold concept more thoroughly, drawing on farm diaries to suggest that individual operators typically drew assistance from neighbors at peak times, offering their own labor at other times in exchange. In this view, effective use of mechanized equipment was less a micro-level choice by self-contained farm households than a community undertaking, reflecting norms of cooperation not unlike those of the turnpike movement.

The significance of the mechanization debate goes well beyond the detailed timing of reaper adoption. At issue is the characterization of the contrast between agriculture in the free and slave states. As proposed by Heywood Fleisig (1976), an essential feature of slavery was to provide an elastic supply of labor to individual farms, allowing them to expand acreage and production without mechanization. In contrast, free labor was scarce and unreliable, long-term contracts legally unenforceable. Olmstead and Rhode (1995) show that farmers were not limited to members of their own family, but this evidence does not gainsay the broader observation that in

the absence of slavery, acreage could not expanded indefinitely on the basis of hired labor. As Fleisig shows, the size distribution of farms under slavery extended well beyond that of the free states, and larger free-state farms had greater investments in implements and machinery.

To be sure, one cannot necessarily compare cotton and wheat farms directly, because these crops were quite different both in seasonality and technology. It is possible, however, to compare northern wheat farms with the slave-using wheat farms of Virginia. Despite the common belief that slave labor was not well-suited for wheat, the two were in fact closely connected in the fertile Valley of Virginia and the Piedmont, so much so that James Irwin (1988) published an article "Exploring the Affinity of Wheat and Slavery in the Virginia Piedmont." Irwin's data showed that slavery was more closely tied to wheat than to tobacco in that area and that slave-using wheat farms were larger and more productive than their smaller counterparts. A plausible explanation is that slave owners could mobilize virtually the entire captive labor force at the time of the harvest peak. As one ex-slave recalled: "John Fallons had 'bout 150 servants an' he wasn't much on no special house servants. Put everybody in de field, he did, even de women. Growed mostly wheat on de plantation, an' de men would scythe and cradle while de women come along an' stack" (quoted in Wright 2006, p. 119). On this reading, slavery and mechanization were alternative methods of expanding cash crop acreage.

Biological Sources of Productivity Growth

A more fundamental challenge to mechanization as the exclusive driver of productivity was provided by Olmstead and Rhode, in their 2008a book *Creating Abundance: Biological Innovation and American Agricultural Development*. The relative constancy of land yields over time led many economists to infer that farmers devoted little attention to improving crops, but Olmstead and Rhode show that this was far from the case. An 1840 survey found 41 wheat varieties under cultivation in New York State, and in 1857 the Ohio State Board of Agriculture identified 111 varieties that had been grown in recent years. One reason for diversity was that the geographic centers of cultivation were constantly changing, generating demands for new varieties adapted to new geo-climatic areas. Even within localities, static plants came under unrelenting attack from pathogens (such as rusts and other fungi) and pests, such as the Hessian fly, the grain midge, and the chinch bug. Maintaining constancy in land yields was itself a productive achievement requiring continuing effort – a phenomenon the authors call the curse of the Red Queen, after the Lewis Carroll character who declared that "HERE, you see, it takes all the running YOU can do, to keep in the same place." Olmstead and Rhode deploy their estimates of biological innovation to revise Parker and Klein on the sources of productivity growth, but the broader message is that such accounting exercises are misleading when inputs interact with each other in complex ways.

Olmstead and Rhode (2008b) extend their biological theme to the antebellum cotton economy. Economic historians have long noted the growth of cotton

production relative to the slave labor force, but this simple ratio reflects changes in the mix of crops and in the geography of cotton, as well as true productivity growth. Drawing from the records of 142 plantations, Olmstead and Rhode develop a panel dataset of picking rates between 1801 and 1862, reporting a fourfold increase, or a growth rate of 2.3% per year. To be sure, picking was not the only important labor activity in cotton farming, but it seems to have been the constraining task in most times and places. The authors argue that the principal force behind this productivity growth was the advent and diffusion of new cotton seeds, primarily those adapted from Mexican imports during the 1820s, and later hybrids. The new varieties sharply increased lint-to-seed ratios, while counteracting the adverse effects of plant disease (such as the "rot") and destructive insects (such as the boll worm and the cotton worm). This evidence confirms that cotton farmers were entrepreneurial and adaptive like their northern counterparts. It does not undermine the assertion that a substantial portion of southern growth was attributable to a shift onto superior cotton lands, but the authors argue that these natural advantages were augmented by technological changes that were both picker-friendly and regionally biased. This new quantitative evidence has played an important role in ongoing debates between cliometricians and the "new historians of capitalism" (Olmstead and Rhode 2018).

Manufacturing and American Technology

The biggest splash in the antebellum economy was the rise of manufacturing. From the barest of beginnings before 1815, manufacturing output grew to more than 30% of national commodity output by 1859 according to Gallman (1966), roughly 20% of GDP. Still more impressive, many of these industries displayed novel technologies by mid-century, drawing international attention at the Great Crystal Palace Exhibition of 1851. In at least one case (small arms), technology flowed eastward across the Atlantic in the 1850s, sparking discussions of a new American System of Manufactures.

Economic historians have analyzed this American Industrial Revolution from the beginning, but quantitative evidence has been unsystematic and scarce. A breakthrough came from Kenneth Sokoloff (1986), who used data from the 1820 Census of Manufactures, the 1832 McLane Report, and samples from the 1850 and 1860 Censuses of Manufactures (see Atack et al. 2006) to construct estimates of labor productivity and total-factor productivity for northeastern manufacturing firms, 1820–1860 (Table 1). The results showed positive productivity growth for all 13 industries studied, averaging 2.2–2.5% per year across the 40-year period. The figures suggested an acceleration of productivity growth during the 1850s, particularly in boots and shoes, woodworking, iron, flour milling, tobacco, and wool. Sokoloff suggested an evolution of productivity over time, from an early phase based on organizational change (such as the advent of the factory and the employment of women) to an acceleration based on sophisticated capital equipment embodying new technologies.

Table 1 Growth rates of labor productivity in selected manufacturing industries, 1820–1860 (percent per year). The figures are average annual percentage change in an index of real value added per equivalent worker. The ranges of estimates reflect the difference between figures derived from firm data and those based on aggregate data. Averages based on fewer than 13 industries (affected by missing values) are reported in brackets. (Source: Sokoloff 1986, p. 698)

	1820–1850	1850–1860	1820–1860
Boots/shoes	1.0–1.2	4.4–5.2	2.0–2.1
Coaches/harnesses	1.8–2.0	2.7	3.7
Cotton textiles	2.3–3.5	1.9–2.9	2.2–3.3
Furniture/woodwork	1.9–2.7	3.8–6.2	2.9–3.0
Glass	3.7	−0.7	2.5
Hats	2.1–2.3	3.0–3.4	2.4–2.5
Iron	0.5–0.7	4.6–4.6	1.5–1.7
Liquors	0.5–1.9	2.0–5.3	1.7–1.9
Flour/grist mills	−0.4–0.4	1.6–3.8	0.6–0.7
Paper	6.0–6.2	−1.2–4.0	4.3–5.5
Tanning	1.7–3.0	−1.7–0.4	1.2–1.7
Tobacco	0.1–1.0	6.3–8.1	2.1–2.4
Wool textiles	1.3–2.3	4.4–7.0	2.7–2.8
Weighted average	[1.8]–2.3	3.2–[3.2]	[2.2]–2.5

What were the characteristics of these new technologies? This question has been the object of a long-running debate, following the thesis advanced by Habakkuk (1962) that American producers developed labor-saving, capital-intensive techniques in response to pervasive labor scarcity. The discussion is too lengthy for full review, but a broad consensus has emerged that Habakkuk's simple two-factor framework is not adequate to capture the distinguishing features of American technology. What has emerged instead is a set of identifiable traits that differentiated American from British manufacturing practice. Some examples include:

1. Greater use of natural resources relative to both capital and labor. In the antebellum era, American machines and products tended to be made of wood, and the country became an early leader in woodworking technology (Rosenberg 1976). A case in point was the Blanchard lathe, invented in 1818 for shaping gunstocks but adapted over time for reproducing other irregular shapes such as shoe lasts, hat blocks, spokes, and oars.
2. Use of special-purpose machinery for long production runs of standardized commodities (Ames and Rosenberg 1968). Such products were well-suited to the utilitarian tastes of American consumers and the relatively equal distribution of income.
3. American manufacturing processes operated *faster* than their European counterparts, from machine speeds to the intensity of the work pace (Field 1983). Intense use of the capital stock foreshadowed the "high throughput" production systems of later eras.

4. American manufacturers did not necessarily substitute capital for "labor" generally, but deployed machinery to substitute unskilled for skilled craft labor. Early American factories relied heavily on women and children, subsequently on immigrants (Goldin and Sokoloff 1982). High mobility and job turnover were the norms, to which mechanization was a response.

These attributes and others may be considered adaptations of European technologies to the American setting. The crucial development was an indigenous technological community, its members pursuing individual goals but collectively shaping a new body of techniques and practices. American technological progress was a network phenomenon, unfolding over a longer time period but taking its shape and direction during the antebellum era.

Network relationships do not readily lend themselves to cliometric analysis. A classic work in this field is Nathan Rosenberg (1963), which described how specialized machine tools firms spun off from the machine shops associated with early textile mills, applying their skills to diverse new industries – sewing machines, locomotives, steam engines, turbine waterwheels, bicycles, and eventually the automobile. Although Rosenberg's paper had been invited by and presented to the Conference on Income and Wealth, it was dropped from the conference volume because it had no numbers! Yet the article, subsequently published in the *Journal of Economic History*, has been described as "the single most influential essay ever written" in the history of technology. What Rosenberg called "technological convergence" – the application of common technical principles to a wide range of industries – might today be reframed by identifying metal working and machine building as general purpose technologies. Equally important, however, was Rosenberg's depiction of machine tools firms as aggressively promoting and seeking out new markets for their accumulated knowledge and expertise.

Cliometric students of technology frequently work with data on patents. Kenneth Sokoloff and Zorina Khan (1990) credit the patent system for the "democratization of invention" in the United States compared to the United Kingdom and for the acceleration of productivity change in the late antebellum period. Whereas the British system was expensive, administratively cumbersome, and elitist, US patents were inexpensive and accessible to a much broader segment of the population. Particularly after the patent reform of 1836, which restored the examination system and established "the world's first modern patent institution" (Khan 2005, p. 53), US patenting rates per capita surpassed those of the United Kingdom. According to Ross Thomson (2009), as many as 36,000 individuals obtained patents in the United States between 1790 and 1865.

Patenting rates are a very imperfect measure of technological progress. Not only are many inventions patented but never used, but many important innovations were not patented. In a study of the 1851 London Crystal Palace Exhibition, Petra Moser (2005) found that only 11% of British exhibits and 16% of American exhibits were patented. Patenting rates varied significantly by industry, the highest being machinery at 36.4%. Thomson found somewhat higher patenting rates at the 1853 New York Crystal Palace Exhibition: three-fifths of Americans displayed patented

products, the highest shares in specialized machines for harvesting, woodworking, transportation, and firearms. Surveying international patterns, Moser suggests that patenting laws may have influenced the *direction* as opposed to the pace of technological change. But many types of innovation were not patentable during this era (such as seed varieties), and secrecy was a viable alternative to patenting in many others (such as chemicals and dyestuffs).

Despite their limitations, patenting data have proven useful for tracing network patterns among inventors and technologies. Thomson (2009) follows intricate lines of influence among textile machinists and steam engine innovators, while Meyer (2005) presents maps showing the clustering of innovation centers and paths of mobility for machinists and engineers. Thomson uses patent evidence to document a shift in the 1830s from own-industry inventions to crossover linkages passing through machine tools, largely confirming Rosenberg's account. Evidence of complex supply chains of specialized machinery, and of high mobility and job-hopping for skilled machinists, convey a powerful resemblance to modern "knowledge economy" clusters such as Silicon Valley. But, whereas Silicon Valley has remained highly concentrated geographically, the antebellum knowledge network was also a vehicle for diffusing technology and industry from the Northeast to the Midwest, which was developing an advanced manufacturing sector even in the antebellum era.

Economic Growth in Slave South and Free North

Accelerated growth in the antebellum era was shared by southern as well as northern regions. As North saw it, the cotton trade was the prime mover for the economy as a whole. He wrote:

> Cotton was strategic because it was the major independent variable in the interdependent structure of internal and international trade. The demands for western foodstuffs and northeastern services and manufactures were basically dependent upon the income received from the cotton trade. . .it was cotton which was the most important influence in the growth in the market size and consequent expansion of the economy. . .Cotton played the leading role (1961, pp. 67–68, 194).

This analysis has been revived by practitioners of the New History of Capitalism, in their efforts to show the centrality of slavery for American development. But it has largely been rejected by cliometric research.

Drawing on contemporary southern newspapers, railroad reports, and periodicals, Diane Lindstrom (1970) confirmed Fishlow's finding that the South provided only a limited market for imported foodstuffs: "the needs of the lower South for flour and corn were insufficient to absorb the output of these products from the upper South, to say nothing of their serving as a major outlet for western produce" (p. 113). The reason for this pattern is that most cotton plantations were themselves self-sufficient in food, planting ample corn crops to spread the fixed costs of slave labor across the

year and maintaining swine to feed the residents. The landmark Parker-Gallman sample of 5,229 farms from the 1860 manuscript census was developed specifically to test North's interregional trade hypothesis (Parker and Gallman 1992). (For a description of the sample, see Parker 1970.) Drawing upon the sample, Gallman (1970) concluded that "large plantations as a group produced more than enough grain to meet their requirements" (p. 20) and that "large planters in the cotton South were attempting to be self-sufficient in basic foodstuffs" (p. 22), typically satisfying this objective. Taken together, the evidence rejects the claim that "the growth of the market for western foodstuffs was geared to the expansion of the southern cotton economy" (North 1961, p. 68).

As a market for northeastern manufactured goods, the South was sizeable in the immediate aftermath of the War of 1812, but its role was never dominant, and it diminished over time. Using capture-recapture methods to analyze the coastal trade from New York City, Lawrence Herbst (1978) estimated that no more than 16.4% of northern manufacturing output went South in 1839, of which only a subset was attributable to surging exports of cotton. In her study of economic development in the Philadelphia region, Lindstrom (1978) found that manufacturers rarely sold goods in distant markets before 1840, and when they did, these markets were normally in the East. Longer-distance trade grew over time, but primarily along east-west lines. The transportation revolution hastened both western settlement and commercialization, together comprising the majority of demand growth for US manufactures. Financial connections between the slave South and the money markets of New York City were undoubtedly important. But northern financial centers were active and growing long before cotton became significant (Sylla 1998). Investment opportunities in western railroads were at least as attractive as southern slave plantations to northern capitalists. In short, the free states had their own growth dynamic, in which the South was largely peripheral.

If the two systems were increasingly divergent in their features, attention shifts to their comparative economic performance. Surprising as it may seem, economic historians have debated this issue for years, in a sense continuing an exchange that goes back to antebellum times. Northern critics observed that the South was far behind in cities and towns, transportation, manufacturing, population growth, and education. Southern apologists pointed out that the slave system avoided many unappealing features of capitalism (such as large cities, factories, and mass immigration), while generating large fortunes that free-state residents could only dream of. The cliometric debate has largely revolved around a set of regional income estimates developed by Richard Easterlin (1961). The data showed that in 1840, southern per capita income was only three-fourths the national average, while average incomes in the South Atlantic and East South Central regions were barely half those in the northeast. The South as a whole grew rapidly between 1840 and 1860, but much of this spurt was attributable to the shift of population from the low-income southeast to the high-income West South Central area, propelled in turn by surging world demand for cotton. True, southern per capita income levels would look more favorable if only incomes of the free population were considered, but dividing by total population is more appropriate for assessing the legacy of slavery for long-term regional development.

Lindert and Williamson (2016) have recently used their "social tables" approach to develop new regional income estimates. Although the resulting figures are not clearly comparable to those generated from the output side, patterns of relative regional income may still be informative. An interesting feature of Lindert and Williamson's account is the long-term relative decline of the South. After the region bore the brunt of the post-Revolutionary War economic collapse, southern per capita income grew at only 0.9% per year between 1800 and 1860, compared to 1.94% in New England and 1.66% in the Middle Atlantic. The authors attribute this relative slippage not directly to slavery, but to the emergence of a poor white class within the region, aggravating inequality within the free population. Because the components of regional performance in this new formulation are not yet clear, their significance for prevailing interpretations remains to be determined.

An alternative approach is to move away from volatile current-income estimates in favor of tracking the accumulation of tangible wealth, which may be viewed as a market valuation of future income streams from assets. Much of the dynamic of antebellum growth was driven by efforts to capture capital gains on land, as settlement and commercial development proceeded. Although this underlying motivation was presumably similar throughout the country, there was one major difference between the regions: southerners could accumulate wealth in the form of slave property while northerners could not. In 1860, the value of slaves accounted for nearly half of all tangible wealth held in the South.

According to the research of Alice Hanson Jones (1980) based on probate records, northern and southern colonies were virtually equal in nonhuman wealth per capita (real estate, livestock, equipment and inventories of producers' and consumers' goods) as of 1774 (Wright 2006). The South, however, had the advantage in total wealth, 20% greater in aggregate, nearly 40% on a per capita basis. At that point in history, one could not say that the use of slave labor had retarded southern economic progress in the colonial era.

Seventy-five years later, North and South were virtually equal in total wealth per capita, consistent with Lindert and Williamson's finding that northern growth was faster across this epoch. During the cotton-boom decade of the 1850s, the South surged to an 18% lead by this metric (Table 2). If slave values are included, the South

Table 2 Regional wealth 1850 and 1860 (in dollars). (Source: Wright 2006, p. 60)

	1850		1860	
	North	South	North	South
Physical wealth (billions of dollars)	4,474	2,844	9,786	6,332
Value of slaves (billions of dollars)		1,286		3,059
Nonslave wealth (billions of dollars)	4,474	1,559	9,786	3,273
Wealth per capita	$315	$316	$482	$569
Nonslave wealth per capita	$315	$174	$482	$294
Nonslave wealth per free capita	$315	$266	$482	$449
Wealth per free capita	$315	$483	$482	$868

was the wealthiest region. Note, however, that in both years the South was well behind in nonslave wealth per capita, a metric that better indicates how the regions would compare economically after slavery was abolished. The south-to-north ratios are 55% in 1850 and 61% in the peak census year 1860. These are not very different from the ratios of regional per capita income that prevailed in the postbellum era.

Perhaps more telling than the aggregate ratios is a calculation showing that wealth per slave owner was more than $2,000 in 1860, or more than four times the average for northerners. The average wealth of nonslaveholding white southerners, by contrast, was barely half that of their northern counterparts. This discussion has omitted intangible forms of wealth such as education, an indicator that would only accentuate the contrast between the regions – most obviously for the slaves but also for the free populations. On this reading, the origins of southern backwardness were firmly located in the antebellum era.

Conclusion

Cliometric studies have clearly shown the acceleration of economic growth during the antebellum era, for aggregated national measures of production and income per capita. Both slave and free regions shared in this growth, but the components and direction of economic change increasingly diverged. In the free states, growth was characterized by extensive investments in transportation and population recruitment, commercialization and mechanization of agriculture, and the emergence of a manufacturing sector with a distinctive technology adapted to American conditions, stretching over time from Northeast to Midwest. The slave South was highly successful in expanding cotton production and in generating wealth for the owners of slaves. But the region lagged in other dimensions of economic development, leaving it poorly positioned for ongoing progress after Civil War and emancipation.

Cross-References

- ▶ Cliometrics and Antebellum Banking
- ▶ Douglass North and Cliometrics
- ▶ Historical Measures of Economic Output
- ▶ Origins of the U.S. Financial System

References

Adams DR (1980) American neutrality and prosperity, 1793–1808: a reconsideration. J Econ Hist 40:713–737

Ames E, Rosenberg N (1968) The Enfield Arsenal in theory and history. Econ J 78:827–842

Atack J, Bateman F (1987) To their own soil: agriculture in the antebellum North. Iowa State University Press, Ames

Atack J, Margo RA (2011) The impact of access to rail transportation on agricultural improvement: the American Midwest as a test case. J Transp Land Use 4:5–18

Atack J, Bateman F, Weiss T (2006) National samples from the census of manufacturing: 1850, 1860, and 1870. Inter-University Consortium for Political and Social Research, Ann Arbor
Atack J, Bateman F, Haines M, Margo RA (2010) Did railroads induce or follow economic growth? Urbanization and population growth in the American Midwest, 1850–1860. Soc Sci Hist 34:171–197
Atack J, Jaremski M, Rousseau PL (2014) American banking and the transportation revolution before the Civil War. J Econ Hist 74:943–986
Bogart D, Majewski J (2008) Two roads to the transportation revolution: early corporations in the United Kingdom and the United States. In: Costa D, Lamoreaux NR (eds) Understanding long-run economic growth: geography, institutions and the knowledge economy. National Bureau of Economic Research. University of Chicago Press, Chicago
Carter SB, Gartner SS, Haines MR, Olmstead AL, Sutch R, Wright G (eds) (2006) Historical statistics of the United States, earliest times to the present: millennial edition. Cambridge University Press, New York
Coleman A (2012) The effect of transport infrastructure on home production activity: evidence from rural New York, 1825–1845. Motu working paper 12-01. Motu Economic and Policy Research, New Zealand
Craig LA, Palmquist RB, Weiss T (1998) Transportation improvements and land values in the antebellum United States: a hedonic approach. J Real Estate Financ Econ 16:173–189
Cranmer HJ (1960) Canal investment, 1815–1860. In: Parker WN (ed) Trends in the American economy in the nineteenth century. Princeton University Press, Princeton
David PA (1966) The mechanization of reaping in the ante-bellum Midwest. In: Rosovsky H (ed) Industrialization in two systems: essays in honor of Alexander Gerschenkron. John Wiley and Dons, New York
David PA (1967) The growth of real product in the United States before 1840: new evidence, controlled conjectures. J Econ Hist 27:151–195
David PA (1996) Real income and economic welfare growth in the early republic or, another try at getting the American story straight. Discussion papers in economic and social history, vol 5. University of Oxford, Oxford
Davis JH (2004) An annual index of U.S. industrial production, 1790–1915. Q J Econ 119:1177–1215
Davis JH, Irwin DA (2003) Trade disruptions and America's early industrialization. NBER working paper no. 9944. National Bureau of Economic Research, Cambridge
Easterlin RA (1961) Regional income trends, 1840–1950. In: Harris SE (ed) American economic history. McGraw-Hill, New York
Field AJ (1983) Land abundance, interest/profit rates, and nineteenth-century American and British technology. J Econ Hist 43:405–431
Fishlow A (1965) American railroads and the transformation of the antebellum economy. Harvard University Press, Cambridge, MA
Fleisig H (1976) Slavery, the supply of agricultural labor, and the industrialization of the South. J Econ Hist 36:572–597
Fogel RW (1964) Railroads and American economic growth: essays in econometric history. Johns Hopkins University Press, Baltimore
Folbre N, Wagman B (1993) Counting housework: new estimates of real product in the United States, 1800–1860. J Econ Hist 53:275–288
Gallman RE (1960) Commodity output, 1839–1899. In: Parker WN (ed) Trends in the American economy in the nineteenth century. Princeton University Press, Princeton
Gallman RE (1966) Gross national product in the United States, 1834–1909. In: Conference on Research in Income and Wealth (ed) Output, employment, and productivity in the United States after 1800. Columbia University Press, New York
Gallman RE (1970) Self-sufficiency of the cotton economy of the antebellum South. In: Parker WN (ed) The structure of the cotton economy of the antebellum South. The Agricultural History Society, Washington, DC
Goldin CD, Lewis FD (1980) The role of exports in American economic growth during the Napoleonic Wars, 1793–1807. Explor Econ Hist 17:6–25

Goldin CD, Sokoloff K (1982) Women, children, and industrialization in the early republic. J Econ Hist 42:741–774
Goodrich C (ed) (1961) Canals and American economic development. Columbia University Press, New York
Habakkuk HJ (1962) American and British technology in the nineteenth century. Cambridge University Press, Cambridge
Haites AF, Mak J, Walton GM (1975) Western river transportation: the era of early internal development, 1810–1860. Johns Hopkins University Press, Baltimore
Herbst LA (1978) Interregional commodity trade from the North to the South and American economic development in the antebellum period. Arno Press, New York
Irwin J (1988) Exploring the affinity of wheat and slavery in the Virginia Piedmont. Explor Econ Hist 25:295–322
Jones AH (1980) Wealth of a nation to be: the American colonies on the eve of the revolution. Columbia University Press, New York
Khan BZ (2005) The democratization of invention: patents and copyrights in American economic development, 1790–1920. Cambridge University Press, New York
Klein DB (1990) The voluntary provision of public goods? The turnpike companies of early America. Econ Inq 28:788–812
Lebergott S (1964) Manpower in economic growth: the American record since 1800. McGraw-Hill, New York
Lindert PH, Williamson JG (2016) Unequal gains: American growth and inequality since 1700. Princeton University Press, Princeton
Lindstrom DL (1970) Southern dependence upon interregional grain supplies: a review of the trade flows, 1840–1860. In: Parker WN (ed) The structure of the cotton economy of the antebellum South. The Agricultural History Society, Washington, DC
Lindstrom DL (1978) Economic development in the Philadelphia region, 1810–1850. Columbia University Press, New York
Meyer D (2005) Networked machinists: high-technology industries in antebellum America. The Johns Hopkins University Press, Baltimore
Moser P (2005) How Do Patent Laws Influence Invention? Evidence from nineteenth-century world's fairs. Am Econ Rev 95:1214–1236
North DC (1966) The economic growth of the United States, 1790–1860. W.W. Norton, New York. First published 1961
Olmstead AL (1975) The mechanization of reaping and mowing in American agriculture, 1833–1870. J Econ Hist 35:327–352
Olmstead AL, Rhode PW (1995) Beyond the threshold: an analysis of the characteristics and behavior of early reaper adopters. J Econ Hist 55:27–57
Olmstead AL, Rhode PW (2008a) Creating abundance: biological innovation and American agricultural development. Cambridge University Press, New York
Olmstead AL, Rhode PW (2008b) Biological innovation and productivity growth in the antebellum cotton economy. J Econ Hist 68:1123–1171
Olmstead AL, Rhode PW (2018) Cotton, slavery and the new history of capitalism. Explor Econ Hist 67:1–17
Parker WN (ed) (1970) The structure of the cotton economy in the antebellum South. The Agricultural History Society, Washington, DC
Parker WN, Gallman RE (1992) Southern farms study, 1860. Inter-University Consortium for Social and Political Research, Ann Arbor
Parker WN, Klein JLV (1966) Productivity growth in grain production in the United States, 1840–60 and 1900–10. In: Output, employment and productivity in the United States after 1800. Columbia University Press, New York
Paskoff PF (2007) Troubled waters: steamboat disasters, river improvements, and American public policy, 1821–1860. Louisiana State University Press, Baton Rouge

Rosenberg N (1963) Technological change in the machine tool industry, 1840–1910. J Econ Hist 23:414–443
Rosenberg N (1976) Perspectives on technology. Cambridge University Press, Cambridge
Rosenbloom JL (2004) Path dependence and the origins of cotton textile manufacturing in New England. In: Farnie DA, Jeremy DJ (eds) The fibre that changed the world: the cotton industry in international perspective, 1600–1990s. Oxford University Press, Oxford
Segal H (1961) Cycles of canal construction. In: Goodrich C (ed) Canals and American economic development. Columbia University Press, New York
Sokoloff KL (1986) Productivity growth in manufacturing during early industrialization: evidence from the American Northeast, 1820–1860. In: Engerman SL, Gallman RE (eds) Long-term factors in American economic growth. University of Chicago Press, Chicago
Sokoloff K, Zorina Khan B (1990) The democratization of invention during early industrialization: evidence for the United States, 1790–1846. J Econ Hist 50:363–378
Sylla R (1998) U.S. securities markets and the banking system, 1790–1840. Fed Reserve Bank St. Louis Rev 80:83–98
Temin P (1988) Product quality and vertical integration in the early textile industry. J Econ Hist 48:891–907
Thomson R (2009) Structures of change in the mechanical age: technological innovation in the United States, 1790–1865. The Johns Hopkins University Press, Baltimore
Towne MW, Rasmussen WD (1960) Farm gross product and gross investment in the nineteenth century. In: Parker WN (ed) Trends in the American economy in the nineteenth century. Princeton University Press, Princeton
Wallis JJ (2003) The property tax as a coordinating device: financing Indiana's mammoth internal improvement system, 1835–1842. Explor Econ Hist 40:223–250
Wallis JJ, Weingast BR (2018) Equilibrium federal impotence: why the states and not the American national government financed economic development in the antebellum, era. J Public Financ Public Choice 33:19–44
Weiss T (1992) U.S. labor force estimates and economic growth, 1800–1860. In: Gallman RE, Wallis JJ (eds) American economic growth and standards of living before the Civil War. University of Chicago Press, Chicago
Weiss T (1993) Economic growth before 1860: revised conjectures. In: Weiss T, Schaefer D (eds) American economic development in historical perspective. Stanford University Press, Stanford
Wright G (2006) Slavery and American economic development. Louisiana State University Press, Baton Rouge

Economic-Demographic Interactions in the European Long Run Growth

James Foreman-Peck

Contents

Abstract

Cliometrics confirms that Malthus's model of the preindustrial economy is a good description for much of demographic-economic history; increases in productivity raise population, but higher population drives down wages. A contributor to the Malthusian equilibrium was the Western European marriage pattern, the late age of female first marriage, which promised to retard the fall of living standards by restricting fertility. The demographic transition and the transition from Malthusian economies to modern economic growth attracted many cliometric models surveyed here. A popular model component is that lower levels of mortality over many centuries increased the returns to, or preference for, human capital investment so that technical progress eventually accelerated. This initially boosted birth rates and population growth accelerated. Fertility decline was earliest and most striking in late-eighteenth-century France. By the 1830s, the fall in French marital fertility is consistent with a response to the rising opportunity cost of children. The rest of Europe did not begin to follow until near the end

J. Foreman-Peck (✉)
Cardiff University, Cardiff, UK
e-mail: foreman-peckj@cardiff.ac.uk

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_17

of the nineteenth century. Interactions between the economy and migration, mainly focused on the long nineteenth century, have been modeled with cliometric structures closely related to those of natural increase and the economy. Wages were driven up by emigration from Europe and reduced in the economies receiving immigrants.

Keywords
Demographic transition · Economic growth · Malthusian economy · Migration

Introduction

For most of history until the industrial revolution and the onset of modern economic growth, living standards have been stagnant or periodically falling and rising around a stationary level in response to wars, famines, plagues, and climate change. In the absence of technological or social change, population has also tended to a long-term balance. Demographic transition is a stylized description of a shift from one type of social order to another. In phase 1, mortality and fertility are high and population is broadly stable. Mortality falls in phase 2, but fertility does not; consequently, population explodes. In phase 3, fertility drops so that population begins to stabilize at a much higher level (Chesnais 1992). Cliometric analysis (not necessarily under this banner) attempts to link these demographic changes with economic development by explaining fertility and human capital accumulation as the outcome of household decisions that are rational in their environments but that also influence the way their environments evolve. In so doing, the aim is to model the transition from a low living standard "Malthusian" economy to one with rising output per capita.

A country's population is potentially influenced not merely by natural increase – the excess of fertility over mortality – but also by migration. Simply establishing the size of historical populations has been a major challenge for historians, with migration an almost unknown magnitude, primarily seen as a source of bias in estimates of birth rates, marriage rates, and death rates. But from the nineteenth century, with more effective state control and interest in statistics, usable migration data becomes available. In this field, cliometric analysis has focused primarily on the vast movements across the Atlantic before the First World War. Because a survey has been recently published (Hatton 2010), the central concern here is with the commonalities of modelling population movement and population change through natural increase.

These themes matter because the size and quality of an economy's population have been critical for its military and economic success and even its survival. Population is a tax base and source of military recruitment. Some economies at various times have considered themselves overpopulated and encouraged emigration. Others have believed they were underpopulated and promoted immigration or family-friendly policies. The natural increase of national or ethnic groups has been and can be a source of social tension, as has immigration; one response has been legal restrictions on immigration, pressure for which has not disappeared.

Unusual in the early acknowledgement of some state responsibility for sections of the population in difficulty was England's 1601 Poor Law. Elsewhere in Europe, relief of these people was primarily left to charity or religious organizations. A question that exercised English social thinkers in the late eighteenth century still concerns some today; if the state gives financial support for families with low earnings, will this encourage indigence and a larger population? A related policy concern has been "will encouraging or permitting immigration lower domestic wages and employment and will encouraging emigration raise domestic wages and job opportunities?"

In the following survey, section "Data" discusses the sources and data used by the cliometric literature; section "Population, Natural Increase, and the Economy" presents the literature on natural increase of population in a simple Malthusian model framework; section "Demographic Transition and Economic Growth" considers the possible ways that a shift could be made to modern economic growth; section "Migration and the Economy" extends the Malthusian framework to selected cliometric migration literature; section "Identification and Estimation" considers approaches adopted to inferring parameters of interest from the historical data; and section "Time Series Analyses" outlines the distinctive approaches of some recent studies that use methods for data indexed in time order.

Data

How far cliometrics can proceed substantially depends upon the nature of the available evidence. The more distant past is generally more problematic in this respect than the more recent. Much effort over many years has been devoted to establishing the historical course of demographic and economic variables, with some surprising results. The Emperor Diocletian's Price Edict of AD 301 has allowed a comparison of Roman living standards with those of the medieval and modern worlds (Allen 2009). The culmination of work by Thorold Rogers (1866), Beveridge (1939), and Phelps-Brown and Hopkins (1955, 1956, and 1981) in Clark's (2005) annual English real wage (for builders, coal miners, and agricultural workers) series beginning in 1209 is no less remarkable. English real wages were apparently higher in 1209 than in 1800, and in 1450, they were higher than at both these dates.

Swedish data on real wages show even more extreme contrasts, suggesting that unskilled laborers were better off in the Late Middle Ages than in the mid-nineteenth century. After about 1540, the trend in real wages in Stockholm is downward so that by 1600, real wages were some 40% lower than in 1540 (Söderberg 2010). This remarkably strong fall occurs elsewhere in Europe as well (Allen 2001). The difference between Britain and the Netherlands on the one hand and Southern Europe on the other is that the first two countries recovered the 1450 peak by the mid-nineteenth century, but the south did not. Like the Swedish series, the German real wage series beginning in 1500 are derived from the builders' wages (Pfister et al. 2012). But experience in Germany differed because of the labor scarcity created by the devastating demographic consequences of the Thirty Years War. Over the first

half of the seventeenth century, the war period, the German real wage rose by 40%. By the end of the third quarter century, wages had climbed to the level at the beginning of the sixteenth century.

The foregoing wage series all refer to men's earnings. Humphries and Weisdorf (2015) have constructed indices of women's casual and contractual wages in England between 1260 and 1850 which show a different pattern from men's wages. Women's annual wages were held down by regulation more than males in the labor scarcity period of the fourteenth and fifteenth centuries, reducing the incentive to postpone marriage and reduce fertility. Married women gained from more buoyant casual wages and by accessing the better paid male labor market through their husbands. When industrialization began to impact on the labor market strongly, the position of the two types of female employment was reversed. Women who could not commit to annual contracts fared less well, increasing the reliance of married women on the male earner.

Human capital indicator series, that might eventually reflect the level of skill, are much more fragmentary, but no less important. Literacy, book production, subscribers to the Encyclopédie, and age heaping have all been enlisted as indicators (A'Hearn et al. 2009; Baten and van Zanden 2008; Squicciarini and Voigtländer 2015). Most series are distinguished by their tendency to rise strongly in early modern Europe – before the onset of sustained increases in national income. This may be why Allen (2003) finds no effect of literacy on economic development measured by wages, though perhaps the impact is concealed by the close link with urbanization which he does find important. In France, illiteracy rates fall sharply from 65% to 5% between 1720 and 1880 (Diebolt and Perrin 2013a).

Tax surveys or returns, private or national censuses, and listings for military, civic, or religious purposes have all provided the raw material for constructing estimates of population, fertility, and mortality.[1]

Perhaps the best known surviving early European census of population and assets is that of 1086 in England, in the Domesday Book. Prussia conducted a national official population census in 1725 (Wilke 2004) and Sweden followed in 1749. By the beginning of the nineteenth century, governments were beginning to undertake censuses regularly; the first French and British official censuses took place in 1801. But for the sixteenth to the nineteenth centuries, parish registers of baptisms, marriages, and deaths have provided the principal sources of key demographic variables through aggregation and family reconstitutions in Europe.

Family reconstitution entails tracing individuals from birth, through marriage and births of children to death. Adding up these individual reconstitutions across a parish potentially provides measures of life expectation, age at marriage, fertility, and other indicators. Drawbacks include the difficulty of linking names in the different

[1]When interpreting these materials, it is important to appreciate that aggregated data can conceal relations that are apparent in more disaggregated sources of information (Brown and Guinnane 2007). The Princeton Fertility Project in particular (for instance, Coale and Watkins 1986) has been criticized for drawing incorrect inferences from excessively aggregated data.

registers, migration reducing the continuity of experience, and the possibility that substantial portions of the local population were not included in the registrations, either because of shortcomings of the recorder or because some people were able to avoid registration, perhaps for religious reasons. Migration from parishes of birth can bias estimates of mean marriage age and life expectancy even when demographic characteristics of migrants and nonmigrants are similar (Ruggles 1992). Only those married in their place of birth are included in the family reconstitution. Late marriage is more likely to take place after migration and so to be systematically excluded from the data. Early deaths are overrepresented in reconstitutions because those who live longer and have time to migrate will die elsewhere – and be omitted.

Good-quality parish registers have a high survival rate in France, Spain, and Italy. Sweden and Finland supplement registers with listings of inhabitants by house with notes on religion, reading ability, and migration in the seventeenth century (Flinn 1981). For England, Wrigley and Schofield (1981) established mortality and fertility rates by family reconstitution of 404 Anglican parish registers of baptism, marriage, and burial, for which the earliest date from the beginning of registration in 1541.[2] Wrigley and Schofield also used the data from parish registers to calculate population by "Back Projection."[3] This involved back dating and revising the age structure of the 1871 population census at 5-year intervals by taking into account flows of previously occurring "events." Age at death was not stated in the register before 1813, so deaths were allocated to age groups by a model mortality schedule.

The high famine and plague-induced mortality of the fourteenth and fifteenth centuries is a critical demographic event for Western Europe.[4] The Great European Famine of 1315–1317 and subsequently in England, diseases of cattle and sheep, and drought (Dodds 2004; Stone 2014), followed by the Black Death of 1348, drastically reduced populations. Moreover, the plague returned in 1361 and 1369, as well as later, with diminished force. In Sweden, population fell from 1.1 million to less than one-third between 1300 and 1413. Not until the mid-seventeenth century did the population recover, and sustained growth only resumed in the later part of the century (historical statistics.org). In England, clearly there was a massive decline in population, perhaps until the middle of the fifteenth century. In the Durham area, tenant numbers imply that population fell to 45% of pre-Black Death levels by the end of the fourteenth century, and tithe evidence indicates a similar collapse of output (Dodds 2004). Clark's (2007) calculations from wage data reach a broadly comparable conclusion for aggregate English population. Tuscan population decline was apparently even more severe; between 1244 and 1404, the population of the Pistoian countryside fell to less than one-third of its former level, and the city

[2]The Cambridge team also published a much more detailed analysis based upon 26 English parishes (Wrigley et al. 1997).

[3]Lee and Anderson (2002) contend that the resulting population estimates are inaccurate for taking into account international migration, but a fair representation of population excluding migration.

[4]For an interpretation on film of the impact, see Ingmar Bergman's *The Seventh Seal*.

population fell to one-half (Herlihy 1965). In Germany, the Thirty Years War of the early seventeenth century was a comparable mortality crisis, with population falling by more than one-half (Pfister and Fertig 2010).

Population, Natural Increase, and the Economy

The literature survey is initially structured around Malthus's (1970) fundamental treatment, presented as a two-equation model. This allows a wide range of cliometric and related literature to be interpreted. Such was the importance of Malthus's theorizing about the relation of the economy and population that he earned the rare accolade of posthumous transformation into an adjective. His work is therefore the natural beginning for a historical survey of the interaction of demography and economy.

Malthus behaved like a true social scientist, combining empirical evidence and theory. Among other data sources, he utilized Alexander von Humboldt's observations of Spanish American population behavior and, indirectly, the US census to contrast with European demographics. Malthus's focus on the geometrical progression of population increase compared with the arithmetic progression (at best) of food increase, when there was little extra land that could be brought into cultivation, proved compelling. Whereas in the Americas, population doubled every 25 years because land was abundant, in the hilly or mountainous parts of Europe, like Switzerland or Wales, there was virtually no increase, because of "positive" or "preventive" checks.

Positive checks shorten the natural duration of life; they include poverty, famine, pestilence, great cities (with their high mortality induced by work, living, and leisure styles), and war. Bubonic plague was the biggest killer in Western Europe from the fourteenth to the seventeenth centuries. Thereafter, quarantine regulations kept it at bay. Typhus and smallpox then assumed preeminence (Flinn 1981, Chap. 4). Urbanization must also have contributed to holding up mortality as the severity and frequency of bubonic plague declined. In 1500, London was estimated to contain 40,000 people and Paris 100,000, but the populations of both cities exceeded half a million by 1700 (de Vries 1984). Another type of mortality check originated from the failure of two or more harvests in a row, as in Finland by 1697 or Ireland by 1846. Poor transport infrastructure and the high cost of moving food meant that these crises were likely to be localized, although the Great European famine of 1315–1317 was an exception. Movement of armies could generate mortality crises even in the absence of fighting. The Thirty Years War in Germany was as lethal for the civilian population as the Black Death, because of disease spread, crop and livestock destruction, and confiscation, by marauding troops.

Preventive checks act on the effective birth rate; for Malthus, these included delayed age at marriage, "unnatural passions," and abortion. Some combination of these checks constrains population to the fixed land resources of long-settled regions. Subsequent research (Hajnal 1965) showed that substantially delayed age at first marriage of females until around an average age of 25 was indeed the norm in

Western Europe at the time Malthus was writing and for some centuries before (de Moor and van Zanden 2010). Moreover, this custom was unique to Western Europe. Everywhere else, the average marriage age was lower. The customary justification for "restraint" in Western Europe was the need to accumulate or acquire sufficient resources to create a separate household for a married couple. The other form of preventive check, or "moral restraint," in Europe that was unusual by world standards was that perhaps 10% or more of females never married at all. In conjunction with social sanction that held illegitimacy to low levels – perhaps 2–5% – these "moral restraints" limited population growth and the level of population below the rates that otherwise would have prevailed. Using English data, Crafts and Ireland (1976) suggest that in the late eighteenth century, a rise of 3 years in the age at marriage could have at least halved the population growth rate.

A policy implication, Malthus maintained, was that raising Poor Law payments with the number of children in the family would incentivize a larger population and undermine independence. The attractions of the ale house, and of large families, he contended, would be diminished if the laborer knew there were no state handouts to fall back on. Cross-section regression of English parishes by Boyer (1989) indeed indicated not only that higher wages were associated with more births (an elasticity of 0.4) but child allowances stimulated more births; parishes that paid allowances with the third child had 25% more births than those that paid no allowances. Boyer tests Huzel's (1980) suggestion that the allowance system was more likely a response to population increase and finds, on the contrary, that the allowance system was exogenous to births.

On the other hand, Kelly and O'Grada (2014) establish that the disappearance of the positive check coincided with the introduction of systematic poor relief. They cite Malthus himself as an authority for the likelihood that government action contributed to breaking the link between harvest failure and mass mortality. On the European continent, where there was no Poor Law, men and women could not count on relief from hardship, unlike in England, and this had profound consequences for economic development (Solar 1995). They were unwilling to break their ties with the land and become the labor force of an industrial revolution.

Not only did Malthus identify long-run equilibria, or steady states of population and wages, but he also noted the likelihood of a population and wage cycle. "Overpopulation" drives down money wages and pushes up food prices. This reduction of real wages discourages marriage, and so population stagnates or declines. But low wages encourage the extension of cultivation and improvement of land already farmed until real wages recover with the stronger demand for labor, and population expansion resumes.

"This sort of oscillation. . ..may be difficult even for the most penetrating minds to calculate its periods." (Malthus 1970, p. 77)

These oscillations can be simulated in response to a mortality shock, for instance, to show a key feature of the Malthusian model. As an exercise in "deterministic calibration" they illustrate an approach that has proved popular in demographic-economic interaction modeling in recent years (albeit with more complex models). In a stylized discrete time model of the Malthusian process, the relevant single period

may be at least 15 years but perhaps double that. The wage in the current period (w_t) falls with the population in the current period (P_t) (which brings the labor force on to the market) due to the diminishing marginal product of labor, and the fixed available land.[5] Where u_t is a random disturbance term with mean zero and a and b are parameters,

$$w_t = a - b.P_t + u_t \quad a,b > 0 \tag{1}$$

Perhaps the strongest evidence for this relationship is the terrible European population mortality from plague in the fourteenth century which boosted wages for instance in England to levels not seen again until the nineteenth century (Clark 2005).[6] Pfister et al. (2012) for Germany from 1500 find a strong negative relationship between population and the real wage until the middle of the seventeenth century that probably reflects this relationship. A plausible estimate of b is 0.5 according to Lee and Anderson (2002). (This assumes an elasticity of substitution of about 1 and labor's share in national income of 0.5.[7]) Allen (2003) measures the effects of population in this type of equation by the land-to-labor ratio in a cross-European country panel beginning in 1300 and ending in 1800. He estimates an elasticity of 0.4, when the a parameter in (Eq. 1) above is a function of urbanization and total factor productivity in agriculture. Crafts and Mills (2009), using time series English data from 1540, estimate a much higher elasticity of 0.95 for b (compared with Lee and Anderson's (2002) $b = 1$). One source of the difference from Allen may be that the time series approach captures short-term relationships which are less responsive than the long-term coefficients obtained from the panel data. Another possible reason is that Allen includes a wider range of explanatory variables in his model, leaving less wage variation to be explained by population. Crafts and Mills also find the shift in a up to 1800 is an average rate of technological progress of 0.75% per annum using Wrigley and Schofield's real wage series and 0.4% when Clark's more broadly based series is employed. Cervelatti and Sunde (2005) extend Eq. 1 by distinguishing two wages, arising from the demand for skilled labor and the demand for unskilled labor, an approach also followed by Diebolt and Perrin (2013a).

A second, quite different, relation connects population and wages. Population in the current period *Pt* increases with the previous period wage because higher wages

[5]In practice, cultivated land area expanded a little with population in Western Europe, as less productive soils were brought into use. Broadberry et al. (2015) Table 2.10 estimate that in England, the cultivated land area only exceeded the medieval peak of 1290 by 1836, when population was several times greater than at the earlier date.

[6]Spain appears to be an exception in Western Europe (Alvarez-Nogal and Prados de la Escosura 2013)

[7]Defining W as log w and p as log P (the labor force), the marginal productivity condition is $W = a - 0.5(p - q)$ where q is the log of output and the elasticity of substitution between factor inputs is unity. An additional assumption is that there should be close-to-perfect competition in labor markets.

encourage earlier marriage and because children are "normal goods" (the diminution of the preventive check); as household income rises, more children become desirable.[8] As well as the effect on births, at low living standards, higher wages reduce positive checks to premature death.

For the moment, no theoretically distinction will be made between positive and preventive checks; both, and their net effect, may depend upon the level of wages.[9] Where v_t is a random disturbance term with mean of zero and c and d are parameters,

$$P_t = c + d.w_{t-1} + v_t \quad c,d > 0 \tag{2}$$

Allen (2003) estimates a long-run equation of this form for early modern Europe and finds a positive coefficient d for the Netherlands and England and Wales but no long-run response for the other European countries in his sample. Evidence from a version of (Eq. 2), distinguishing the effect of wages on birth rate from that on death rate, includes a median of 14 European countries' fertility response to wages from 1540 to 1870, estimated at an elasticity of 0.14 and for England, of 0.12 (Lee and Anderson 2002; Table 2). Mortality elasticities for England go down to −0.076 with an indication of higher rates – perhaps −0.16 – elsewhere in Europe. These numbers imply a positive population response to wages in England. In the long run, the coefficient d may tend to infinity, so that wages eventually return to some customary subsistence level after an increase in productivity. This would be consistent with Ashraf and Galor's (2011) findings.[10]

An alternative measure of (the inverse of) wages, which has the merit of exogeneity to annual population and birth and death rates, is food prices. After 1740, there was no response of death rates to food prices in France, perhaps a century later than in England (Weir 1984). French marriages were more responsive to price shocks than the English, but in the nineteenth century, there was a weakening of this French preventive check. In eighteenth-century Sweden, a 15% rise in rye prices was associated with at least a 3% increase in mortality the following year (Bengtsson 1993). Sweden also showed evidence of the preventive check in both centuries, with higher rye prices reducing marriage rates and fertility.

Lagerlof (2003) postulates that v_t (in Eq. 2 above) is primarily due to mortality shocks, the values of which play a critical role in the breakout from the Malthusian equilibrium. He also derives an analogy to Eq. 2 from household optimization of a preference function for surviving children, human capital, and goods. This theme is developed by Foreman-Peck (2011) who shows that a fall in child mortality

[8] In reality, there may be longer lags in this relationship, which in turn lengthens the periodicity of the cycle discussed below. Autocorrelated shocks or disturbances have the same effect.

[9] When birth and/or death rates respond to wages, as, for example, in Lee (1973), then Eq. 2 explains the change in population and should be modified by the addition of $-P_{t-1}$ to the right-hand side. In the interests of simplicity, this modification is not implemented here.

[10] And with Arthur Lewis' (1954) model of economic development with unlimited supplies of labor, although here, the perfectly elastic supply of labor comes from migration, rather than natural increase.

theoretically reduces target births but increases desired family size and population. Across late-nineteenth-century Europe and across English counties, lower mortality rates are actually associated with lower birth rates.

From Eqs. 1 and 2 (substituting out wages) and assuming the disturbance terms take their mean values, we obtain a first-order difference equation for P:

$$P_t + bd.P_{t-1} = c + ad \tag{3}$$

Given the initial condition, the population in the base year P_0, we can solve the difference equation:

$$P_t = \frac{c + ad}{1 + bd} + \left(P_0 - \frac{c + ad}{1 + bd}\right)(-bd)^t$$

The first right-hand-side component is the particular solution. In the limit, this is the steady-state value of population. The second right-hand-side term is the complementary function with a characteristic root equal to $-bd$. Population will oscillate around the steady state (particular solution) every period until it converges to the equilibrium level (as long as $|bd| < 1$).

The wage equation corresponding to the population Eq. 3 is

$$w_t + bd \;\; w_{t-1} = a - bc. \tag{4}$$

The solution to this wage difference equation is

$$w_t = \frac{a - bc}{1 + bd} + \left(w_0 - \frac{a - bc}{1 + bd}\right)(-bd)^t.$$

Higher living standards are achieved by larger values of a and lower values of c, b, or d. A higher marriage age lowers the population by reducing c and d. Exogenous population growth, measured by the growth of c, drags down wages. This could be due to falling mortality, as Boucekkine et al. (2003) postulate when they calibrate their model with mortality schedules from Venice 1600–1700 and Geneva 1625–1825. Quarantine regulations in this period were supposedly increasingly successful in diminishing outbreaks of plague in Europe (Chesnais 1992, p. 141). Conversely, if urbanization was sufficiently important to raise national mortality rates (Voigtländer and Voth 2013b), c would fall and wages would rise. Wrigley and Schofield (1989, p. 475) maintained that in the half century after 1820, the rapid increase in the proportion of the population urbanized contributed substantially to the failure of English life expectations to rise significantly (though this is after the Voth period).

Deterministic calibration typically chooses values for parameters $a - d$ so that the model tracks the historical series of interest. More ambitiously, the researcher may adopt parameter values that have been estimated. Based on the calibration $a = 1$,

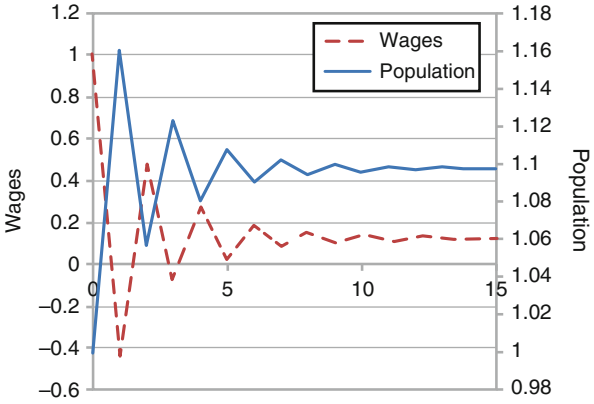


Fig. 1 Simulated Malthusian cycles

$b = 0.8$, $c = 1$, and $d = 0.8$, the steady state for wages to which the system converges is $w^* = 0.12$ and for population $P^* = 1.097$. The dynamics of the Malthusian process can be shown with an Excel spreadsheet[11] and by rearranging the system as follows:

$$P_t + 064P_{t-1} = 1.8$$

$$w_t + 0.64\ w_{t-1} = 0.2$$

Starting with values of 1, Fig. 1 shows the inverse fluctuations in population and wages in response to very large initial shocks, cutting population and boosting wages. The figure shows convergence on $w^* = 0.12$ and $P^* = 1.097$, getting quite close over 10 periods, perhaps 150 or 300 years (in the absence of other shocks, if each period is 15 or 30 years). The initial levels can be considered to represent a positive mortality shock, such as those of the fourteenth and fifteenth centuries in England or the Thirty Years War in Germany, cutting the population and boosting wages.

In summary, two features of this simple dynamic model are the very long-lasting oscillation and the inverse movements of wage and population. It is a homeostatic system returning to an equilibrium of population and wages. Are the theoretical cycles realistic? Lee (1993) maintains that at the macroeconomic level, homeostasis has only been a weak background force. The approximately 250-year European

[11]Enter the parameters of the population difference equation in say cells A1 and A2 (respectively, 0.64 and 1.8 in this case). Fill a column (say B) with a series starting at zero and increasing by one with each subsequent cell. Assign the column next to B for $P\,t$. The first value depends upon the shock to be considered. As a positive shock, use any number greater than 1.097 here. So entering 1 as the first cell in the C column will be a negative population shock. In cell C2, enter "= − A1*c1 + A2" and fill down column C. The series rises above the equilibrium level in period 1 and falls below it in period 2. The behavior of the equation can be studied by changing the parameters assigned to cells A1 and A2.

cycle was mainly driven by exogenous and probably autocorrelated shocks. On the other hand, there is much evidence that for most of history, there has been a stable Malthusian equilibrium of wages and population (Ashraf and Galor 2011). Across countries, land productivity and the technological level affected population density in the first to the sixteenth centuries, whereas the effects of land productivity and technology on income per capita in these years were not significantly different from zero.

Demographic Transition and Economic Growth

This Malthusian equilibrium might be defined as phase 1 of the interaction between demography and the economy. Phase 2 of the transition in Europe refers to the eighteenth and nineteenth centuries, when population generally increased strongly; real wages eventually no longer fell, and in due course began to rise. This last suggests an acceleration in the pace of technical change, represented in Eq. 1 by an increase in the a coefficient (and perhaps a fall in b). Malthus predicted that technical progress would be absorbed by greater populations, as birth rates rose. But real wages do not fall in this phase because advances in productivity offset the diminishing returns from population growth.

Lee and Anderson (2002) define the "population absorption rate" of an economy as the rate at which population can grow without a fall in real wages. This depends upon the growth in a relative to the growth in c (in Eq. 2). The balance was likely to change, and at diverse times in different countries. In eighteenth century England, strong growth in population in the eighteenth century no longer reduces wages. In Germany, the negative relation between wage and population size (Eq. 1) was weaker in the eighteenth than in the sixteenth century; the fall of the marginal product of labor was less pronounced, and the beginning of the eighteenth century saw a marked increase in labor demand (Pfister et al. 2012). German labor productivity experienced a strong positive shock during the late 1810s and early 1820s and continued to rise at a weaker pace during the following decades. Sustained economic growth began well before the beginnings of German industrialization, in the third quarter of the nineteenth century. French exceptionalism appears with the alternative to accelerating technical change; the widespread evidence of marital fertility control in France in the 1790s (reducing c and d), nearly 100 years before comparable evidence in England (Weir 1994).

In a version of unified growth theory (Galor and Weil 2000), rising population boosts technological advance, offsetting diminishing returns. The interpretation could be extended to include market widening, such as the European discovery of the Americas – Smithian growth (cf Acemoglu et al. 2005). Natural selection in favor of higher child quality (Galor and Moav 2002) cumulatively also has this consequence of raising technical progress, or increasing a.

By contrast, a permanent mortality shock reducing death rates could trigger offsetting behavior. Higher chances of child survival would require fewer births for desired final family numbers. And because fewer resources would be lost

investing in children dying before adulthood, larger family sizes can be afforded. The outcome would then be eventually lower wages as population increased, unless positive or preventive checks intervened (Doepke 2005).

However, longer life expectancy can raise the returns to investment in human capital because there is a greater period over which the benefits accrue. Eventually, accumulation can trigger an acceleration of technical progress (Boucekkine et al. 2003; Lagerlof 2003; Cervellati and Sunde 2005). Or higher child survival chances simply might increase parental investment in "child quality" (Foreman-Peck 2011). In this last case, higher female marriage age is hypothesized to increase the rate of human capital accumulation through greater investment in "child quality" as well. Assuming literacy is a partial measure of child quality and human capital, the hypothesis is supported by the finding that the proportion of single females aged between 25 and 29 negatively predicts illiteracy across Europe, controlling for schooling. Moreover, illiteracy in English counties in 1885 is negatively associated with age at marriage. Controlling for prior school attendance, literacy appears to boost output in a cross-Europe production function for the years 1870 to 1910.

The contribution of the later female first marriage preventive check to the breakout from the Malthusian equilibrium in Western Europe has been controversial (Dennison and Ogilvie 2014, Carmichael et al. 2016). Another hypothesized chain of events where marriage age matters is that mortality shocks of the fourteenth and fifteenth centuries induced higher male wages that switched demand towards meat, thereby encouraging the expansion of pasture at the expense of arable farming. Pasture was supposedly more conducive to higher women's wages which encouraged later age at marriage and lower fertility (Voigtländer and Voth 2013a). But women's wages in England do not seem to have conformed to this pattern (Humphries and Weisdorf 2015). A third suggested role for later age at marriage of females has been the general empowerment of females and this supposedly led to economic development (De Moor and Van Zanden 2010).

Inspection of Eq. 3 suggests Phase 2 of the transition occurs if a increases continuously; then population will eventually "take off." The same applies to the wage difference equation. Figure 2 shows the effect of continuous technical progress represented by a linear time trend superimposed upon the Malthusian population cycle. Initially, there is no noticeable effect; the cyclical response to the initial shock is similar. After period six, population is growing continuously, though obviously the strength of the time trend determines when the breakout occurs.

The fertility transition of the next phase, 3, has been interpreted either as an innovation – the spread of information about contraception – or as an adjustment, a response to changes in motivation and structural factors affecting preferences and choices (Carlsson 1966). These last include new economic conditions such as prohibitions of, or restrictions on, child labor, greater child costs, cultural change, higher incomes, and falls in mortality. Mortality reductions may *increase* target family size by lowering the number of births necessary to achieve a given target and thereby cutting the costs or price of a surviving child. But a fall in death rates may also contribute to fertility *decline*, simply because fewer births are necessary to achieve a given completed family size.

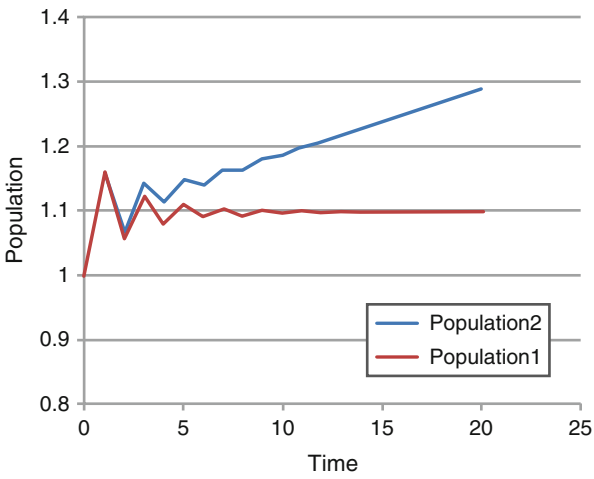


Fig. 2 Population in the break out from the Malthusian cycle

Galor (2012) maintains that a rise in the demand for human capital was the main trigger for the fertility transition. In terms of the simple model, a rise in demand for human capital stems from technological progress, reflected in continuous increases in a, which must raise real wages continuously. Since real wages are growing, the opportunity cost of children will be rising. This may reduce d, the response of population to wages. Other reasons for falling values of d include those mentioned above. When $d = 0$, population stops growing. However, wages continue to grow at the rate given by the growth of a. A "demographic transition" then is completed by the fall in d and perhaps c in Phase 3.

After 1830, French behavior is consistent with an adjustment; rising living standards reducing the demand for children for a slightly rising age at marriage was accompanied by falling marital fertility (Weir 1984). Not only living standards were driving the decline (Diebolt and Perrin 2013a, b). Empowerment increased the amount of time invested by women in their education, and fertility declined for this reason. Female literacy was associated with falling fertility in France in the period 1881–1911. In much of Bavaria, unlike France, only after 1900 did sustained fertility decline begin, when the rising opportunity costs of children became apparent (Brown and Guinnane 2002). Textile employment, a measure of nonagricultural opportunities for women, markedly cut fertility, as did higher women's wages. Conversely, women on small farms that relied primarily on family labor were more likely to have more children. For the British fertility decline Tzannatos and Symons (1989) find that fertility responded positively to income but was more than offset by the rise in cost of female time – that closely followed women's education (and implicitly, wages).

Belgian fertility decline was earlier among groups voting socialist, liberal, and communist in 1919 and later among those paying Easter dues to the Roman Catholic Church (Lesthaeghe 1977). These experiences suggest a role in fertility decline for ideology and ideological change creating a willingness to adopt more effective

contraception (Crafts 1984; Bhattacharya and Chakraborty 2017). Crafts (1984) maintained that the cheaper or more widely understood was contraception, the lower would be both illegitimacy and the legitimate fertility. Illegitimacy is then assumed to be an indirect measure of contraceptive costs. The approach shows that economic choice model of fertility can synthesize the "innovation" and "adjustment" processes, integrating contraceptive costs (broadly defined) into the decision to adopt birth control techniques.

Migration and the Economy

The Malthusian scheme also provides a conceptual framework to assess the impact of labor force/population growth induced by European migration. In the growing international economy of the nineteenth century, millions sailed from Europe to new lives across the Atlantic. These bursts of migration triggered lagged increases in building activity to absorb them and more rapid (extensive) economic growth than could be supported by natural increase alone. Much of the cliometric literature has been concerned with the forces of "push" from the country of origin, or "pull" by the destination, behind migration (Thomas 1973; Hatton and Williamson 1998, 2006; O'Rourke and Williamson 1999; Hatton 2010).

A technology shock (perhaps railway building) in the region of recent European settlement increases the demand for labor (a rises) and raises wages there:

$$w_t = a - b.P_t + u_t \tag{1}$$

Higher wages eventually mean (a "pull" for) higher immigration and therefore higher population in the recipient region:

$$P_t = c + d.w_{t-1} + v_t \tag{2}$$

This is the same model as that used to represent the Malthusian economy, but by the later nineteenth century, the Atlantic economy was expanding fast, with a continuously rising a parameter. There is a positive correlation of the real wage and the upswing. In the region without the positive shock, less labor and capital are supplied, because better returns are to be had elsewhere. In the booming region, the time necessary for building the infrastructure to take full advantage of the technology means that the flow of labor and capital continues for some years, until marginal returns are equalized again between regions (allowing for nonpecuniary differences and costs of migration), or other shocks occur.

An early cliometric calibrated cyclical model of this process for an export economy was presented by Parry Lewis (1960). The region of immigration exports "coal" (Parry Lewis had in mind later nineteenth-century Wales but probably today, the term would be replaced by "tradable goods"). Also, the economy has a "building" sector ("non-tradable (capital) goods"). When conditions abroad cause the demand for exports to oscillate or grow in a specific pattern, then the cycles and

growth would be reflected in building. If, in addition, exogenous population pressure abroad ("push") causes waves of emigration, then the building sector will fluctuate similarly. This endogenous cycle is heavily damped but endogenous immigration reduces the degree of damping.

Demographic impulses ("push" from the origin countries) as well as technology shocks promote the distinctive inverse cycles between the regions.[12] A case in point is the Napoleonic war "baby boomers" (a rise in c) that, in due course Thomas (1973) maintains, created the "hungry forties." Malthusian pressure in Europe pushed migrants to the USA; capital tended to follow them and the demand for housing in the USA rose (even though, in contrast to the positive technology shock, real wages in the receiving region fall, relative to what they would have been).

To the extent that the immigrants are complementary to the indigenous work force, wages will rise. More likely, as assumed above, is that some wages (say skilled) will rise and others fall – if, for instance, immigrants are unskilled. The more capital that flowed with the migrants, the stronger the economic growth they promoted and the less adverse the impact on wages. Taylor and Williamson (1997) calculated that in the absence of mass migration after 1870, real wages in 1910 would have been higher by 27% in Argentina, by 17% in Australia, and by 9% in the USA. The pervasive nineteenth-century innovation of railways was a major shift in technology for which immigration amplified the impact on output, while reducing output per head in Malthusian fashion (Foreman-Peck 1991, pp. 87–88). The more responsive immigration was to wages, the lower was steady-state output per head.

In the sending region, migration was a partial alternative to mortality as a positive check, for instance, in Ireland and Germany in the 1840s. In a Malthusian economy, emigration simply made space for a higher natural increase. In a neoclassical growth model economy on the other hand, with exogenous population growth and an unchanged savings ratio ($d = 0$ in the Malthusian scheme), output per head and wages would be raised by emigration. Emigration explained about half of the rise in wages across Swedish counties between 1870 and 1910 (Ljungberg 1997). Taylor and Williamson (1997) estimate that in the absence of emigration, real wages would have been lower by 24% in Ireland and by 22% in Italy but by only 5% in Great Britain and 2% in Germany.

Immigration restrictions in the two-equation model discussed have a similar effect to Malthus's prediction of a reduction in child benefit under the 1601 Poor Law. They reduce the population response to higher wages (reduce the value of d in Eq. 2). From the end of the US Civil War to the 1920s, European immigration provided strong competition to internal US migration from the southern states to the urban North and West. So, US immigration restrictions of the 1920s favored black

[12]Both types of shocks may be classified as originating on the supply side and as "real" rather than "monetary," consistent with real business cycle theory (Kydland and Prescott 1982).

migrants from the South who gained from the elimination of European competition (Thomas 1973; Williamson 2005).[13]

In Europe, the countries of emigration, the effects of the closure of the USA were less benign. Agricultural protectionism in Europe was encouraged by redundant work forces, unable to move to the USA, who were instead employed growing subsidized crops (Thomas 1973). Migrants were also diverted to Canada and South America, boosting output there.

Identification and Estimation

Much of the cliometric literature is inevitably concerned with how the values of the parameters of the favored model can be known or estimated. If an association in time series data between wages and population can be found, in principle, it could reflect relations generated by one or both of the equations in the model discussed above. The parameter values *a, b, c*, and *d* cannot be inferred from the estimated relation without further information; the original equations are not identified.

If shocks or variables affecting one equation, that do not affect the other, can be distinguished then there is a chance of identifying the parameters. A mortality shock because of plague might affect the demographic response to wages (Eq. 2) but not the Eq. 1, derived from the production function. In this case, the response of wages to the exogenous shift in population traces out the effect of the *b* coefficient of Eq. 1.

Clark (2007) and Scheidel (2010) use this principle to construct or infer population numbers. Scheidel observes that the second-century Antonine and sixth-century Justinian plagues of the Roman Empire were associated with higher Egyptian real wages, from which he infers that population must have fallen substantially on these occasions. Clark (2007) reconstructs English population back to 1200 with a peak of six million around 1300 from Eq. 1. An example is the inference from the rise in agricultural wages to population scarcity created by the mortality crises of the fourteenth century.

Harvest failures might lower real wages in a preindustrial society for reasons outside the model; they are u_t in Eq. 1, an exogenous shock. If so, they could be used to identify the demographic response (Eq. 2). Higher food prices lower the value of the real wage exogenously, and the effect on population change might be assessed, perhaps especially through changes in marriage rates. In this case, the lagged responses are a major potential problem because by the time population responds, other shocks with which the harvest failure may be confused will have struck the economy.

Subject to this qualification, if wages fall and population rises, the dominant shock must be demographic (Eq. 2). If both wages and population fall, the dominant shock is technological, such as harvest failure. But this classification only

[13]Williamson (2005) also discusses the corollary that the position of blacks deteriorated after 1970 because of competition from immigrants.

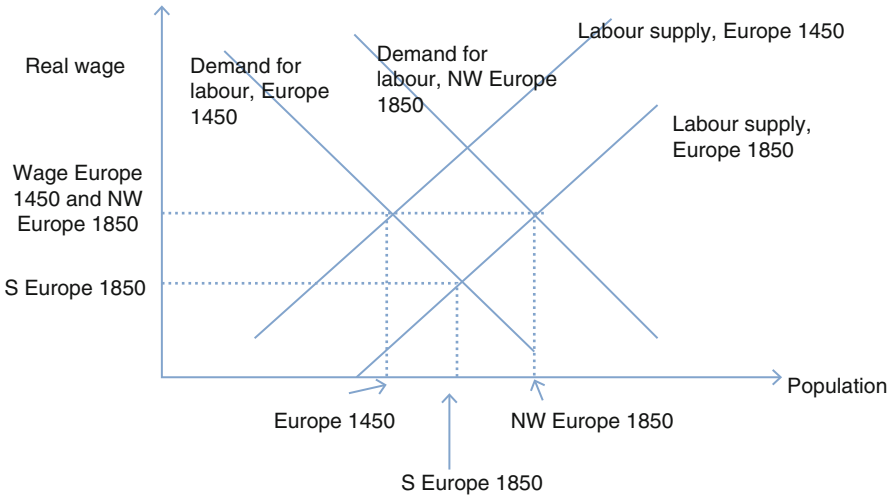


Fig. 3 Allen's (2001) interpretation of European divergence 1450–1850

distinguishes the dominant shock, not how much of the change in population and in wages is due to which type of shock.

An illustration of the identification problem arises in the interpretation of early modern European wage movements with the two-equation model. Real wages in the cities of northwest Europe tended to increase, or at least did not fall, by 1850 compared with 1450. But in Southern and Eastern Europe, they did decline, with few exceptions. Allen's (2001) interpretation, shown in Fig. 3, is that this was due to more vigorous economic development in northwest Europe – in Britain and the Netherlands especially. In principle, the divergence could have been due to a stronger exogenous growth of population in Southern Europe (a greater rightward shift of "labor supply" in the figure) with similar rates of economic development expanding the demand for labor. Apart from the qualitative evidence to the contrary, the stronger growth of population in the Netherlands and Britain compared with that in Spain or Italy identifies the greater shift as in the demand for labor. But we cannot conclude that there was no technical progress in Southern Europe, only that there was less than in the northwest.

With the more recent demographic-economic interactions, more variables are available for identification. For migration modelling, it is important to note that population is a stock and migration is a flow that adds to the stock in the same way as do births. Taylor and Williamson (1997) quantify how international migration in the Atlantic economy altered wages between 1870 and 1910. They estimate the demand for labor (Eq. 1) and utilize the many studies that provide parameter estimates for calibration. They assume that migration was exogenous to national populations or labor forces; it shifted Eq. 2, thereby identifying Eq. 1. Wage changes are then found from the change in population or labor force due to migration over the period.

Time Series Analyses

The econometric alternatives for estimating identified economic-demographic models are generally concerned with time series analysis. In particular, the study of preventive and positive shocks has utilized vector autoregressions (VAR) and impulse response functions. VAR analysis arose from the difficulties of identification, of finding any truly exogenous variables. Past values of every variable in the model are assumed to be potential influences on current values of these variables. In the two-equation model, the following equations would be estimated, although with more lags than written:

$$w_t = a + bP_{t-1} + cw_{t-1} + u_t$$
$$P_t = g + dw_{t-1} + fP_{t-1} + v_t$$

This system of equations can be thought of as encompassing several structural models in an unrestricted way; for example, the original equations only allowed past wages to influence population, whereas now, past population does as well.

The residuals u_t and v_t represent the unexplained movements in population and wages, reflecting the influence of exogenous shocks (i.e., shocks that arise outside the assumed model). These residuals are now an aggregation of the various exogenous shocks affecting the endogenous variables in the underlying structural model. Therefore, no economic interpretation can be derived from the residuals without transforming the equations. If movements of the endogenous variables within the VAR system reflect the effects of exogenous shocks or innovations, the VAR can be used to examine these shocks. Different shocks and their effects may be disentangled by placing identifying restrictions on the VAR.

Then a derived impulse response functions can give the response of birth and death rates (and therefore of population) to an impulse in wages. If there is a reaction of one variable to an impulse in wages, we may call the latter a "cause" of the former. This type of causality can be studied by tracing out the effect of an exogenous shock or innovation in wages on some or all the other variables.

A problematic assumption in this type of impulse response analysis is that a shock occurs only in one variable at a time. Such an assumption may be reasonable if the shocks in different variables are independent. Eckstein et al. (1984) in an early demographic-economic VAR study of Sweden 1749–1860 employ weather variables for this purpose. Shocks to weather affect other variables but are not themselves affected by shocks to wages or demographic variables.

Using the Wrigley and Schofield demographic series and the Phelps-Brown-Hopkins wage data, Nicolini (2007) finds that in England, contrary to the Malthusian model, positive checks disappeared during the seventeenth century and preventive checks vanished before 1740. In Germany, Pfister et al. (2012) from 1500 obtain a strong negative relationship between population and the real wage until the middle of the seventeenth century. On English data, Moller and Sharp (2008) estimate highly significant preventive checks working through marriages but agree with

Nicolini that positive checks were insignificant. They suggest that growing population enhanced income by increasing the size of the market.

Crafts and Mills (2009) establish that wages ceased to be "Malthusian" at the end of the eighteenth century, after which they grew strongly. They maintain that the preventive check cannot be found after the mid-seventeenth century, as in Germany, but unlike Moller and Sharp, they do not use a total marriages variable. Demographic growth was permitted by an expanding demand for labor of about 0.5% per annum. Crafts and Mills find no indication of positive feedback between population size and technological progress, contrary to the assumption of unified growth models (Galor and Weil 2000). For Sweden, Eckstein et al. (1984) incorporate more variables into their VAR system, albeit for a shorter period than with the English data. They include infant mortality and a crop index as well as weather variables. As Malthus postulated, Eckstein et al. estimate that a positive innovation in the general crop index, or in the real wage, increases fertility for several years and decreases infant and non-infant death rates over the same period.

For earlier periods, some of the key variables indicated by theoretical discussion are not available as continuous time series. An example is human capital accumulation that drives technical progress, a vital component of many models (Boucekkine et al. 2003; Cervelatti and Sunde 2005; Lagerlof 2003; Galor and Weil 2000; Foreman-Peck 2011). One approach to this problem is the Kalman filter (Lee and Anderson 2002). This can be illustrated with the (initially) two-equation Malthusian model – which has been shown to generate an equilibrium around a steady-state population and real wage. Technical progress or human capital accumulation in effect shifts the a parameter of the demand for labor equation in the Malthusian model. Lee and Anderson model the behavior of the disturbance terms u_t and v_t and the rates of shift of the parameters a and c. For expositional purposes, the simplest approach is to introduce only one more equation which explains unobserved H, human capital, say, by a disturbance term (summarized below by ε_t) reflecting other unmeasured factors.

$$w_t = a + H_t - bP_t + u_t$$
$$P_t = c + dw_{t-1} + v_t$$
$$H_t = f_t + gH_{t-1+\varepsilon_t}$$

All variables are endogenous in this system; it can be solved for each endogenous variable only in terms of lags and disturbance terms, because there are no exogenous variables. The second step is to use this system for prediction, beginning in the base period $t = 0$, assuming $u_0 = v_0 = \varepsilon_0 = 0$. Starting values of all the six parameters ($a - g$) must be postulated. The third step is to update the prediction in the light of the "new" data. Estimated w_1 and P_1 are compared with the actual values of w_1 and P_1 and new values of the six parameters chosen that maximize their likelihood.

This process is repeated for each period. When best estimates of all the model parameters are obtained, the effects of the human capital can be inferred from the values of f and g and the assumed unit effect on real wages. Lee and Anderson's

(2002) parameter estimates are very similar to those from the Crafts and Mills' VAR (as noted above). Foreman-Peck and Zhou (2018) use a related approach to demonstrate a jump in the English female marriage age at the end of the fourteenth century and the effect on human capital accumulation over the ensuing centuries. Their method, unlike the Kalman filter, does not depend on distributional assumptions for the disturbance terms, but only requires that variables have first and second moments.[14] But most importantly for their purposes, it enables them to estimate the economic-demographic model with different length time series, to use all the available information; the real wage series goes back to 1209, which provides many more observations than the mortality and birth rate series beginning only in 1541, than the even shorter celibacy and marriage age series.

Conclusion

Malthus's model of the preindustrial economy in which increases in productivity raise population, but higher population drives down wages, appears to be a good description of much of demographic/economic history. The Western European marriage pattern – the late age of female first marriage – promised to retard the driving down of living standards by restricting fertility. Otherwise, the positive check of mortality, induced by disease, war, or malnutrition, constrained population in regions that had been long settled by arable and pastoral farmers, despite high birth rates. Living standards then largely depended upon how recently mortality shocks had occurred. Cycles in the Malthusian economy may have been due to lags such as those in the response of fertility to wages, but shocks were perhaps more likely to be the drivers.

Since there is good evidence that population, or natural increase, responded to higher wages, it is likely that subsidy policies like those attributed to the English 1601 Poor Law would encourage fertility of those subsidized, as Malthus feared. There are indications that this happened and that the "social safety net" reduced premature deaths by breaking the link between harvest failure and mortality.

The demographic transition and the transition from Malthusian economies to modern economic growth have attracted many cliometric models, but as yet no consensus about the process has been achieved. Population expanded most rapidly in the most dynamic European economies, so fertility restriction was not obviously the key. Yet, the association of the Western European marriage pattern with economic development, combined with Malthus's emphasis on the vital necessity to balance population against resources, suggests there should be some connection. By the end of the nineteenth century, there was a European pattern that towards the east,

[14]Coefficients are not updated observation by observation as they are with the Kalman filter and the Kalman gain. Instead the recursive system goes all the way through the observations to get one vector of coefficients, minimizing the square of the distance between actual and forecast values of the variables.

mortality rates were higher and age at marriage lower than in the west. Mortality explained fertility statistically, and fertility explained age at marriage, across English counties and across Europe. Time series behavior was rather different because of lags in responses. French fertility control began at the end of the eighteenth century, earlier than elsewhere in Europe. Other European populations increased rapidly for perhaps a century, before fertility fell, with rising child costs and greater opportunity costs of time (in which children were intensive).

Interactions between the economy and migration have been modeled with cliometric structures closely related to those of natural increase and the economy. Similar problems arise with identification, of distinguishing cause from effect from contingent association. The historical focus of recent literature has, however, been different, even when bulk of the data has been available for the same period. Whereas the natural increase literature has been concerned with equilibrium and growth over many centuries, the bulk of the cliometric migration literature has focused on the great migrations from Europe of the later nineteenth century. These studies have yielded clear-cut and conventional results compared with the longer period studies; wages were driven up by emigration from Europe and reduced in the economies receiving immigrants. Policies of migrant restrictions therefore must have influenced wages similarly, if they were effective. Over all economies involved, if migration was an improvement in resource allocation, total output will have increased, but it is unclear how these gains were distributed between source and host economies.

References

Acemoglu D, Johnson S, Robinson J (2005) The rise of Europe: Atlantic trade, institutional change and economic growth. Am Econ Rev 95:546–579

A'Hearn B, Baten J, Crayen D (2009) Quantifying quantitative literacy: age heaping and the history of human capital. J Econ Hist 69(3):783–808

Allen RC (2001) The great divergence in European wages and prices from the middle ages to the first world war. Explor Econ Hist 38:411–447

Allen RC (2003) Progress and poverty in early modern Europe. Econ Hist Rev 56(3):403–443

Allen RC (2009) How prosperous were the Romans? Evidence from Diocletian's price edict (A.D. 301). In: Bowman A, Wilson A (eds) Quantifying the Roman economy: methods and problems. Oxford University Press, Oxford

Alvarez-Nogal C, Prados de la Escosura L (2013) The rise and fall of Spain (1270–1850). Econ Hist Rev 66(1):1–37

Ashraf Q, Galor O (2011) Dynamics and stagnation in the Malthusian Epoch. Am Econ Rev 101(5):2003–2041

Baten J, Van Zanden JL (2008) Book production and the onset of modern economic growth. J Econ Growth 13(3):217–235

Bengtsson T (1993) A re-interpretation of population trends and cycles in England, France and Sweden, 1751–1860. Hist Mesure 8:93–115

Beveridge W (1939) Prices and wages in England from the twelfth to the nineteenth century. Frank Cass, London, 1965

Bhattacharya J, Chakraborty S (2017) Contraception and the demographic transition. Econ J 127(606):2263–2301

Boucekkine R, de la Croix D, Licandro O (2003) Early mortality declines at the dawn of modern growth. Scand J Econ 105(3):401–418
Boyer GR (1989) Malthus was right after all: poor relief and birth rates in southeastern England. J Polit Econ 97:93–114
Broadberry S, Campbell MS, Klein A, Overton M, van Leuween B (2015) British economic growth 1270–1870. Cambridge University Press, Cambridge
Brown J, Guinnane TW (2002) Fertility transition in a rural, Catholic population: Bavaria, 1880–1910. Popul Stud 56(1):35–49
Brown JC, Guinnane TW (2007) Regions and time in the European fertility transition: problems in the Princeton project's statistical methodology. Econ Hist Rev 60(3):574–595
Carlsson G (1966) The decline of fertility: innovation or adjustment process. Popul Stud 20:149–174
Carmichael SG, De Pleijt A, Van Zanden JL, De Moor T (2016) The European marriage pattern and its measurement. J Econ Hist 76(1):196–204
Cervellati M, Sunde U (2005) Human capital formation, life expectancy, and the process of development. Am Econ Rev 95(5):1653–1672
Chesnais JC (1992) The demographic transition: stages, patterns and economic implications; a longitudinal study of sixty-seven countries covering the period 1720–1984. Clarendon, Oxford
Clark G (2005) The condition of the working class in England, 1209 to 2004. J Polit Econ 113(520):1307–1340
Clark G (2007) The long march of history: farm wages, population, and economic growth, England 1209–1869. Econ Hist Rev 60(1):97–135
Coale AJ, Watkins SC (eds) (1986) The decline of fertility in Europe. Princeton University Press, Princeton
Crafts NFR (1984) A time series study of fertility in England and Wales, 1877–1938. J Eur Econ Hist 13(4):571–590
Crafts NFR, Ireland NJ (1976) A simulation of the impact of changes in the age at marriage before and during the advent of industrialization in England. Popul Stud 30:495–510
Crafts NFR, Mills TC (2009) From Malthus to Solow: how did the Malthusian economy really evolve? J Macroecon 31:68–93
De Moor T, Van Zanden JL (2010) Girl power: the European marriage pattern and labour markets in the North Sea region in the late medieval and early modern period. Econ Hist Rev 63(1):1–33
De Vries J (1984) European urbanization, 1500–1800. Methuen/Harvard University Press, London/Cambridge, MA
Dennison T, Ogilvie S (2014) Does the European marriage pattern explain economic growth? J Econ Hist 74:651–693
Diebolt C, Perrin F (2013a) From stagnation to sustained growth: the role of female empowerment AFC working paper nr 4
Diebolt C, Perrin F (2013b) From stagnation to sustained growth: the role of female empowerment. Am Econ Rev Papers Proc 103(3):545–549
Dodds B (2004) Estimating arable output using Durham priory tithe receipts, 1341–1450. Econ Hist Rev 57(2):245–285
Doepke M (2005) Child mortality and fertility decline: does the Barro-Becker model fit the facts? J Popul Econ 18:337–366
Eckstein Z, Schultz TP, Wolpin KI (1984) Short-run fluctuations in fertility and mortality in pre-industrial Sweden. Eur Econ Rev 26(3):295–317
Flinn MW (1981) The European demographic system 1500–1820. Harvester, Brighton
Foreman-Peck J (1991) Railways and late Victorian economic growth. In: Foreman-Peck J (ed) New perspectives on the late Victorian economy: essays in quantitative economic history. Cambridge University Press, Cambridge, pp 1860–1914
Foreman-Peck J (2011) The Western European marriage pattern and economic development. Explor Econ Hist 48(2):292–309

Foreman-Peck J, Zhou P (2018) Late marriage as a contributor to the industrial revolution in England. Econ Hist Rev 71(4):1073–1099
Galor O (2012) The demographic transition: causes and consequences. Cliometrica 6:1–28
Galor O, Moav O (2002) Natural selection and the origin of economic growth. Quart J Econ 117(4):1133–1191
Galor O, Weil DN (2000) Population, technology and growth: from the Malthusian Regime to the demographic transition and beyond. Am Econ Rev 90(4):806–828
Hajnal J (1965) European marriage patterns in perspective. In: Glass DV, Eversley DEC (eds) Population in history: essays in historical demography. Edward Arnold, London
Hatton TJ (2010) The cliometrics of international migration: a survey. J Econ Surv 24(5):941–969
Hatton TJ, Williamson JG (1998) The age of mass migration: causes and economic impact. Oxford University Press, New York
Hatton TJ, Williamson JG (2006) Global migration and the world economy: two centuries of policy and performance. MIT Press, Cambridge, MA
Herlihy D (1965) Population plague and social change in rural Pistoia, 1201–1430. Econ Hist Rev 18(2):225–244
Historicalstatistics.org. The population in Sweden within present borders 4000 BC-2004 AD www.historicalstatistics.org/htmldata6/index/html. Accessed 29 May 2014
Humphries J, Weisdorf J (2015) The wages of women in England 1260–1850. J Econ Hist 75:405–447
Huzel JP (1980) The demographic impact of the Old Poor Law: more reflexions on Malthus. Econ Hist Rev 33(3):367–381
Kelly M, O Grada C (2014) Living standards and mortality since the middle ages. Econ Hist Rev 67(2):358–381
Kydland FE, Prescott C (1982) Time to build and aggregate fluctuations. Econometrica 50(6):1345–1370
Lagerlof N-P (2003) Mortality and early growth in England, France and Sweden. Scand J Econ 105(3):419–439
Lee RD (1973) Population in preindustrial England: an econometric analysis. Quart J Econ 87:581–607
Lee RD (1993) Accidental and systematic change in population history: homeostasis in a stochastic setting. Explor Econ Hist 30:1–3
Lee RD, Anderson M (2002) Malthus in state space: macroeconomic-demographic relations in English history, 1540 to 1870. J Popul Econ 15:195–220
Lesthaeghe RJ (1977) The decline of Belgian fertility 1800–1970. Princeton University Press, Princeton
Lewis WA (1954) Economic development with unlimited supplies of labour. Manch School 22:139–151
Lewis PJ (1960) Building cycles: a regional model and its national setting. Econ J 70(279):519–535
Ljungberg J (1997) The impact of the great emigration on the Swedish economy. Scand Econ Hist Rev 44:159–189
Malthus TR (1798/1830/1970) An essay on the principle of population. Pelican, London
Møller NF, Sharp P (2008) Malthus in cointegration space: a new look at living standards and population in pre-industrial England. Discussion papers, Department of Economics, University of Copenhagen, pp. 08–16
Nicolini EA (2007) Was Malthus right? A VAR analysis of economic and demographic interactions in pre-industrial England. Eur Rev Econ Hist 11(1):99–121
O'Rourke KH, Williamson JG (1999) Globalization and history: the evolution of a nineteenth -century Atlantic economy. MIT Press, Cambridge, MA
Pfister U, Fertig G (2010) The population history of Germany: research strategy and preliminary results. Max Planck Institute for demographic research working paper no 35
Pfister U, Riedel J, Uebele M (2012) Real wages and the origins of modern economic growth in Germany, 16th to 19th centuries, EHES working paper nr 17

Phelps Brown EH, Hopkins SV (1955) Seven centuries of building wages. Economica 22:195–206
Phelps-Brown EH, Hopkins SV (1956) Seven centuries of the prices of consumables compared with builders' wage rates. Economica 23:296–314
Phelps-Brown EH, Hopkins SV (1981) A perspective of wages and prices. Methuen, New York
Ruggles S (1992) Migration, marriage and mortality: correcting sources of bias in English family reconstitutions. Popul Stud 46:507–522
Scheidel W (2010) Roman real wages in context. Princeton/Stanford working papers in classics
Söderberg J (2010) Long-term trends in real wages of labourers, chapter 9 of historical-monetary-and-financial-statistics-for-Sweden-exchange-rates-prices-and-wages. Riksbank, Stockholm, pp 1277–2008
Solar P (1995) Poor relief and English economic development before the industrial revolution. Econ Hist Rev 48(1):1–22
Squicciarini MP, Voigtländer N (2015) Human capital and industrialization: evidence from the age of enlightenment. Quart J Econ 130(4):1825–1883
Stone D (2014) The impact of drought in early fourteenth century England. Econ Hist Rev 67(2):435–462
Taylor AM, Williamson JG (1997) Convergence in the age of mass migration. Euro Rev Econ Hist 1:27–63
Thomas B (1954/1973) Migration and economic growth: a study of Great Britain and the Atlantic economy. Cambridge University Press, Cambridge
Thorold Rogers E (1866) A history of agriculture and prices in England. Clarendon, Oxford
Tzannatos Z, Symons J (1989) An economic approach to fertility in Britain since 1860. J Pop Econ 2:121–138
Voigtländer N, Voth H-J (2013a) How the west 'invented' fertility restriction. Am Econ Rev 103(2013):2227–2264
Voigtländer N, Voth H-J (2013b) The three horsemen of riches: plague, war, and urbanization in early modern Europe. Rev Econ Stud 80(2):774–811
Weir DR (1984) Life under pressure: France and England, 1670–1870. J Econ Hist 44:34–65
Weir DR (1994) New estimates of nuptiality and marital fertility for France, 1740–1911. Popul Stud 48:307–331
Wilke J (2004) From parish register to the "historical table": the Prussian population statistics in the 17th and 18th centuries. Hist Family 9:63–79
Williamson JG (2005) The political economy of world mass migration: comparing two global centuries. American Enterprise Institute, Washington, DC
Wrigley EA, Schofield RS (1981/1989) The population history of England, 1541–1871: a reconstruction. Arnold, London
Wrigley EA, Davies RS, Oeppen JE, Schofield RS (1997) English population history from family reconstitution 1580–1837. Cambridge University Press, New York

The Golden Age of European Economic Growth

A Cliometric Perspective

Nicholas Crafts

Contents

Abstract

This chapter surveys cliometric research on economic growth in Western Europe from 1950 to the early 1970s, the so-called "Golden Age." Several hypotheses to explain the very rapid growth of that period are examined including those proposed by Abramovitz, Eichengreen, Janossy, and Kindleberger. Cross-country variation in growth performance is highlighted and explanations for it are

N. Crafts (✉)
CAGE, University of Warwick, Coventry, UK
e-mail: N.Crafts@warwick.ac.uk

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_35

explored. Further insights into the Golden Age are obtained by considering the reasons for the subsequent growth slowdown. It is concluded that research in this area has made substantial progress in the last 30 years informed by ideas from new growth economics.

Keywords
Catch-up growth · Growth regressions · Relative economic decline · Social capability · Technological congruence

Introduction

The focal point of this chapter is the study of Western Europe's "Golden Age" of economic growth, which lasted from the early 1950s through the mid-1970s. The retrospective analysis of this period by quantitative economic historians only really began in the late-1980s although, obviously, this built upon the earlier work of applied economists. At that point, this historical research received a massive stimulus from the revival of interest in growth economics, which produced potentially appealing new theoretical ideas together with a much more empirical focus which was initially noteworthy in particular for its stress on issues relating to catch-up and convergence.

Hitherto, formal growth theory had been dominated by the neoclassical growth model in which long-run productivity growth was a result of exogenous technological change and changes in the investment rate only affected the level of output per person rather than its growth rate. The key feature of the endogenous growth models, which appeared during the 1980s, was that they made long-run growth rates a result of investment decisions (relating to a broad concept of capital) based on microeconomic foundations. Two different types of models were developed, namely, AK models of growth in which diminishing returns to (broad) capital accumulation were assumed away, and endogenous innovation models in which the rate of technological progress is a result of profit-seeking investments. Two well-known variants of the former type were Romer (1986), based on constant returns to physical capital, and Lucas (1988), where endogenous growth could be the result of human and physical capital accumulation combined, with the former generating externalities. Two well-known variants of the latter type were the quality-ladders approach of Grossman and Helpman (1991) and the Schumpeterian growth model of Aghion and Howitt (1992). The relevance of both these types of models is that well-designed institutions and supply-side policy can have positive growth-rate effects through their effects on incentives to invest and to innovate, rather than just levels effects as in the neoclassical growth model.

Empirical analysis of economic growth was also changing in the 1980s as new data sets became available, notably an early version of what became very well-known estimates of long-run real income levels for OECD economies (Maddison 1982), and a much improved version of the Penn World Tables (Summers and Heston 1984). These permitted more sophisticated international comparisons of

performance and underpinned the development of a huge growth regressions literature. Very early contributions to this by Baumol (1986) and De Long (1988) debated whether the historical record showed a general experience of convergence of income levels among advanced countries. Notions such as "conditional β-convergence" (Barro 1991) and "social capability" for catch-up and convergence (Abramovitz 1986) emerged as attempts were made to examine the roles that institutions and policies had played in growth outcomes.

Already in the 1980s, there were some excellent economic-history-textbook accounts of postwar European economic growth using a traditional approach, for example, Van der Wee (1986). The historiography also featured well-known interpretations of the Golden Age including Kindleberger (1967) and Olson (1982). Against this background, the cliometric contribution came in three main ways. First, it provided superior international and inter-temporal comparisons of growth performance based on much improved measurement techniques and a better articulated conceptual framework. Second, it pursued statistical testing of hypotheses that had previously been evaluated quite informally. Third, it developed more sophisticated ways of thinking about the role of institutions and policies in underpinning the Golden Age and explaining cross-country differences in growth outcomes.

Two (linked) questions are central to this topic, namely, "what explains rapid economic growth in Western Europe during the Golden Age?" and "why did Western European growth slow down so markedly after the Golden Age?" The first of these questions has produced a substantial body of work without by any means arriving at a complete consensus, but the second is still relatively underresearched by cliometricians. In each case, important insights can be gained from considering the variance of growth performance across countries and the correlates of relative success and failure.

Growth Performance

The basic data on growth are set out in Table 1 which is based on the original work of Angus Maddison, in particular with regard to establishing levels of real GDP measured at purchasing power parity. Several points should be noted. First, for European countries the growth rates of real GDP per person (Y/P) and of labor productivity (Y/HW) were much faster in the Golden Age than subsequently. The medians for 1950–1973 were 3.62% and 4.58% per year, respectively, compared with 1.76 and 2.54 during 1973–1995. Second, median European growth rates of Y/P and Y/HW were well above those of the United States in the Golden Age, but from 1973 to 1995 while labor productivity continued to grow faster in Europe than the United States, real GDP per person did not. Catch-up in income levels was rapid during the Golden Age but not thereafter. Third, in each period there is a clear inverse correlation among European countries between initial levels and subsequent growth rates of GDP per person so that in this special case the data exhibit unconditional β-convergence, as is confirmed by the regressions in Crafts and

Table 1 Initial levels and subsequent rates of growth of real GDP per person and per hour worked ($1990GK and % per year)

(a) 1950–1973					
	Y/P, 1950	**Y/P Growth, 1950–1973**		**Y/HW, 1950**	**Y/HW Growth, 1950–1973**
Switzerland	9064	3.08	Switzerland	9.14	3.04
Denmark	6943	3.08	United Kingdom	7.87	3.47
United Kingdom	6939	2.42	Denmark	7.33	2.71
Sweden	6769	3.21	Netherlands	6.88	4.14
Netherlands	5996	3.69	Sweden	6.84	4.18
Belgium	5462	3.54	Belgium	6.64	4.22
Norway	5430	3.24	Norway	5.42	4.53
France	5186	4.02	France	4.92	5.30
West Germany	4281	5.02	West Germany	4.29	5.91
Finland	4253	4.25	Italy	4.16	5.67
Austria	3706	4.94	Finland	4.09	4.63
Italy	3502	4.95	Austria	3.66	5.78
Ireland	3453	3.03	Ireland	3.29	4.06
Spain	2189	5.60	Portugal	2.58	5.97
Portugal	2086	5.45	Spain	2.48	6.07
Greece	1915	6.21	Greece	2.22	6.43
United States	9561	2.45	United States	11.93	2.57
(b) 1973–1995					
	Y/P, 1973	**Y/P Growth, 1973–1995**		**Y/HW, 1973**	**Y/HW Growth, 1973–1995**
Switzerland	18,204	0.59	Switzerland	18.22	0.91
Sweden	14,018	1.11	Sweden	17.55	1.30
Denmark	13,945	1.74	Netherlands	17.50	1.96
West Germany	13,152	1.76	Belgium	17.20	2.60
Netherlands	13,081	1.65	France	16.13	2.67
France	12,824	1.64	Denmark	16.08	2.63
Belgium	12,170	1.87	West Germany	16.02	2.86
United Kingdom	12,025	1.75	Norway	15.06	3.16
Norway	11,324	2.96	Italy	14.74	2.30
Austria	11,235	2.19	United Kingdom	14.55	2.12
Finland	11,085	1.72	Austria	13.28	2.48
Italy	10,634	2.22	Finland	11.62	3.06
Spain	7661	2.48	Portugal	9.77	1.53

(*continued*)

Table 1 (continued)

Greece	7655	1.37	Spain	9.61	3.76
Portugal	7063	2.29	Greece	9.31	1.28
Ireland	6867	2.82	Ireland	8.22	3.37
United States	16,689	1.81	United States	21.40	1.27
(c) 1995–2007					
	Y/P, 1995	**Y/P Growth, 1995–2007**		**Y/HW, 1995**	**Y/HW Growth, 1995–2007**
Norway	21,591	2.20	Belgium	30.26	1.46
Switzerland	20,660	1.84	Norway	29.84	1.66
Denmark	20,350	1.83	France	28.79	1.75
Netherlands	18,697	2.48	Denmark	28.48	1.24
France	18,318	1.75	Netherlands	26.86	1.66
Belgium	18,270	2.08	Italy	24.28	0.49
Austria	18,096	2.21	Germany	24.11	1.70
United Kingdom	17,955	2.54	Sweden	23.28	2.63
Sweden	17,848	2.94	United Kingdom	23.06	2.13
Italy	17,228	1.18	Austria	22.75	1.84
Germany	17,127	1.56	Finland	22.55	2.66
Finland	16,112	3.66	Switzerland	22.13	1.62
Spain	13,132	2.67	Spain	21.69	0.30
Ireland	12,662	5.37	Ireland	17.06	3.64
Portugal	11,614	1.92	Portugal	13.63	1.42
Greece	10,321	3.67	Greece	12.30	2.65
United States	24,712	1.81	United States	28.21	2.21

Source: The Conference Board (2016)
Note: post-1973 Ireland based on GNP

Toniolo (2008).[1] The convergence rate was a little above 2% per year during the Golden Age but slowed to about 1.5% subsequently. Fourth, during the Golden Age, both the United Kingdom and Ireland underperformed in the sense that their growth was significantly slower than in countries with higher initial levels of income and productivity. For the UK, a prima facie verdict of "growth failure" in these years is reinforced by the evidence of being overtaken so that by 1973 levels of real GDP per person and of labor productivity were below those in many other European countries.

Table 2 benchmarks the sources of growth in labor productivity using a standard neoclassical growth accounting framework implemented in identical fashion for each country and permitting comparisons not only between Western European countries but also between the latter part of the Golden Age and the subsequent

[1]Their results show that this did not apply prior to 1950.

Table 2 Contributions to labor productivity growth, 1960–1990 (% per year)

	1960–1970				1970–1990			
	H/L	K/L	TFP	Y/L	H/L	K/L	TFP	Y/L
Austria	0.18	2.39	2.90	5.47	0.22	1.32	1.00	2.54
Belgium	0.42	1.36	2.33	4.11	0.18	0.96	1.38	2.52
Denmark	0.13	2.15	1.25	3.53	0.24	0.82	0.02	1.08
Finland	0.37	1.66	2.64	4.67	0.62	0.98	0.90	2.50
France	0.29	2.02	2.62	4.93	0.36	1.28	0.84	2.48
West Germany	0.23	2.10	2.03	4.36	0.40	0.79	0.69	1.88
Greece	0.26	3.63	4.45	8.34	0.50	1.24	0.06	1.80
Ireland	0.22	1.78	2.21	4.21	0.38	1.47	1.18	3.03
Italy	0.36	2.39	3.50	6.25	0.32	0.98	1.22	2.52
Netherlands	0.74	1.43	0.89	3.06	0.25	0.72	0.65	1.62
Norway	0.48	1.18	1.80	3.46	0.70	0.90	0.84	2.44
Portugal	0.35	2.05	3.99	6.39	0.44	0.90	1.01	2.35
Spain	0.38	2.45	3.73	6.56	0.37	1.54	1.13	3.04
Sweden	0.19	1.34	2.40	3.93	0.36	0.67	0.27	1.30
Switzerland	0.40	1.40	1.37	3.17	0.30	0.72	−0.38	0.64
UK	0.17	1.45	1.24	2.86	0.32	0.83	0.74	1.89

Source: Bosworth and Collins (2003)

Note: Estimates are for the whole economy and labor productivity is measured on a per worker basis. Growth accounting is based on assuming a Cobb-Douglas production function in which $Y = AK^{\alpha}(HL)^{1-\alpha}$ where H is the average educational standard of the labor force, A is TFP, K is capital and L is labor. This allows the following expression for the sources of labor productivity growth to be derived: $\Delta\ln(Y/L) = \alpha[\Delta\ln(K/L)] + (1-\alpha)[\Delta\ln(H/L)] + \Delta\ln A$

period of slowdown. The Golden Age era of rapid catch-up was indeed a period when both capital-deepening and TFP growth contributed greatly to labor productivity growth. Nevertheless, in most cases, TFP growth made the larger contribution and in countries with very rapid labor productivity growth the differential with the slower growing countries was much more due to TFP growth than capital deepening.

When TFP growth is as rapid as it was for Europe during the Golden Age, it can be expected that there is a substantial component from reductions in inefficiency, both allocative and productive. Maddison (1987) in a somewhat speculative exercise concluded that much of the Solow residual was typically attributable to some combination of labor quality, improved allocation of resources, changes in the utilization of factors of production, reductions in technology gaps and economies of scale, leaving only a modest share "unexplained" – and perhaps reflecting disembodied technical change. Maddison's list of the components of rapid TFP growth in the European Golden Age is broadly in line with conventional economic histories, but precise quantification is, of course, very difficult and there is no consensus on the details.[2]

[2]It should be noted that the results of a data envelopment analysis also give strong support to the claim that TFP growth during the European Golden Age was boosted considerably by improvements over time in the efficiency of factor use (Jerzmanowski 2007).

Table 2 reveals that the slowdown in labor productivity growth in Western Europe after the Golden Age reflected declines in both capital deepening and TFP growth in every country but that the latter was generally more important. The unweighted average decrease between 1960–1970 and 1970–1990 was 1.00 percentage points per year for the capital-deepening contribution but 1.75 percentage points per year for TFP growth, which was largely the result of the evaporation of transitory components mentioned above. The dramatic decline of TFP growth in Southern European countries reflects this point.

What Explains the Golden Age of European Growth?

This section of the chapter considers attempts by cliometricians to explain why growth was so much faster in these earlier postwar years than either before or since. Much of this work has entailed testing hypotheses put forward in earlier vintages of economic history. The focus will be primarily on research with a cross-country perspective, but at the end its implications for understanding the UK growth failure will be reviewed.

The Janossy Hypothesis

The conventional idea of the so-called Janossy Hypothesis, which is based on an interpretation of Janossy (1969) proposed by Dumke (1990), is shown in Fig. 1.[3] The basic idea is that after World War II the level of output in European economies was unusually low (at C) and that a period of super growth followed in which output first bounced back to its prewar level (CD) after which growth continued at a higher than normal rate (DE) until the normal growth rate resumed (at E). Obviously, countries varied in terms of how far the point C was below trend, and therefore in the scope for reconstruction and super growth, with West Germany often seen as an extreme observation and thus a good candidate for a *wirtschaftswunder*. This analysis raises two questions, which the cliometric literature has addressed: first, how much did reconstruction contribute to growth? and second, is it correct to see European countries returning to a pre-existing trend growth path?

Since a pioneering paper by Dumke (1990), the issue of the reconstruction contribution has been tackled by running what might be called augmented unconditional convergence regressions. In this approach, growth of real GDP per person is regressed against the initial level of income per person, a measure of the shortfall of the actual output level compared with an estimate of the level of trend output in the late 1940s (typically called "GAP") to capture the potential for reconstruction effects

[3]The graph embodies a trend-stationary view of the world. Arguably, Janossy himself believed in segmented trends where labour productivity growth in each period was underpinned by investment in human capital, see Vonyo (2008).

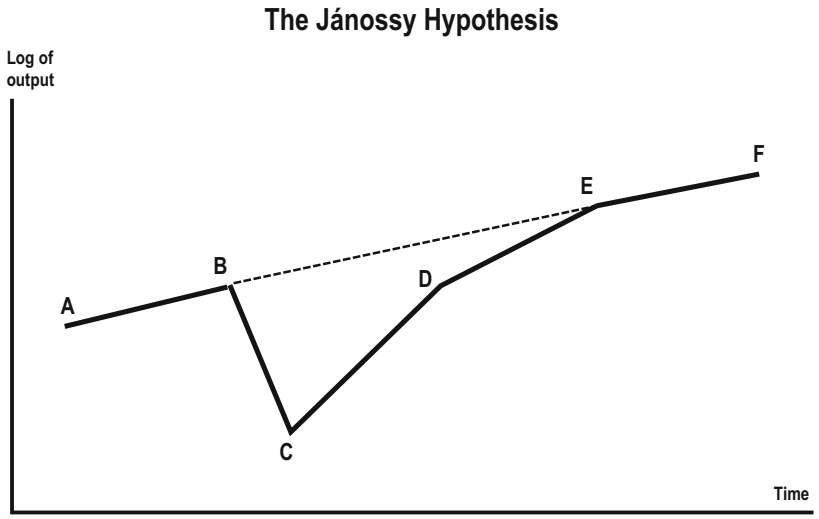


Fig. 1 AF, long run trend line; AB, pre-war output levels; BC, war induced shocks to output; CE, total length of the reconstruction period; CD, recovery of pre-war levels; DE, re-attaining the trend line; EF, output levels after reconstruction; Source: Dumke (1990)

and the share of the labor force in agriculture for a sample of European or OECD economies. Dumke (1990) estimated an equation of this type to explain average growth rates in cross-section for 1950–1980. He found that reconstruction was generally statistically but not economically significant and accounted for only about 5% of growth over the period. For West Germany, however, reconstruction did account for a large part of the difference between its growth and the average growth outcome and was complementary to conventional β-convergence.

A subsequent paper by Eichengreen and Ritschl (2009) echoed these findings for West Germany but argued that the output gap at the end of the war and reconstruction effect in the 1950s was somewhat bigger than Dumke had thought. They emphasized that this explained much of the extremely rapid TFP growth of that decade.[4] Temin (2002) estimated a similar equation for European countries for decadal growth during the 1950s through the 1990s but focused only on average performance. He found that reconstruction had a statistically significant positive effect only for 1950–1960 but not thereafter. Vonyo (2008) developed this approach further; his innovations included adding labor force growth as an independent variable and adopting a panel-data estimation setting. His results are summarized in Table 3. He found that reconstruction effects, which lingered through to the 1980s, were on average economically significant throughout the Golden Age with a very strong

[4]Eichengreen and Ritschl (2009) estimate that TFP growth in West Germany was 5.39% per year during 1950–1960 and it seems reasonable to infer from their discussion that as much as 3.5 percentage points per year may have been due to reconstruction.

Table 3 Contribution of reconstruction to growth of real GDP per person (% per year)

	1950s	1960s	1970s	1980s
Austria	2.18	1.60	1.02	0.44
Belgium	0.73	0.53	0.34	0.15
Denmark	0.77	0.56	0.36	0.16
Finland	1.39	1.02	0.65	0.28
France	1.22	0.89	0.57	0.25
Germany	3.15	2.31	1.47	0.64
Italy	1.48	1.09	0.69	0.30
Netherlands	0.71	0.52	0.33	0.14
Norway	0.39	0.29	0.18	0.08
Sweden	−0.03	−0.02	−0.01	−0.01
Switzerland	−0.86	−0.63	−0.40	−0.17
United Kingdom	0.51	0.38	0.24	0.10
Unweighted Average	0.97	0.71	0.45	0.20

Source: Vonyo (2008)

Note: Derived from panel estimation of two equations in one of which growth is regressed on initial income, labor force growth, GAP*Time, and a country fixed effect and in the other the fixed effect is regressed on GAP. GAP is the proportional difference between potential and actual output in 1948 where potential output is extrapolated from the 1938 level using the average annual growth rate from 1920 to 1938. The reconstruction contribution is obtained by adding the estimated impact of GAP in the two equations

impact in West Germany, but his results are notable also for underlining that in Sweden and Switzerland there was no reconstruction bonus.

Clearly, there is some support among cliometricians for Janossy's emphasis on reconstruction as an important factor in early postwar growth in addition to regular catch-up growth. It is also apparent that its impact varied greatly across countries. An implication of this is that international comparisons of growth performance during the Golden Age should attempt to normalize for differential scope for reconstruction, especially if they involve West Germany. The claim that the effects of reconstruction on growth still mattered in the 1980s should be treated with caution given the different results presented by Temin and Vonyo.

The second issue, relating to trend stationarity in long-run European growth, was investigated using time-series econometrics by Crafts and Mills (1996). They took the "pure-Janossy" hypothesis to mean that the return to normal growth would entail going back to the pre-1914 trend growth rate for real GDP per person by the end of the Golden Age. They found that the growth process could be modelled as segmented trend-stationary for all the European countries in their sample. On this basis they were able to reject the "pure-Janossy" hypothesis except in the case of Denmark. In the cases of Spain, Sweden, and Switzerland, a "modified-Janossy" hypothesis that after the Golden Age they had returned to the pre-1914 trend growth rate (but on a path at a higher income level) could not be rejected. This implies that all European countries except Denmark had higher real GDP per person at the end of the Golden Age than would have been predicted by extrapolating the pre-1914 trend.

This is very much what a reader of Abramovitz (1986) or Maddison (1982) would have expected given their emphasis on the unprecedented nature of rapid catch-up growth after 1950.

Macroeconomic Stability

It is widely recognized that the Bretton Woods era, which coincided with the Golden Age, was a period when macroeconomic fluctuations were relatively gentle (cf. Table 4), and it has been argued that this provided a highly favorable context for rapid postwar growth (Boltho 1982). Clearly, it was important that there was no repeat of the disastrous policy errors associated with the interwar gold standard which led to a lost decade for the gold bloc countries in the 1930s. The absence of banking crises in an era of capital controls and tight regulation of banks would also appear to be a positive feature of these years, as has been underlined by recent experience. Finally, the OPEC oil-price shocks only materialized in the 1970s. That said, some important caveats should be noted.

First, the hypothesis that volatility can harm growth has been somewhat controversial. One reason for this is that models were produced that had the opposite prediction, namely, that larger and more frequent business-cycle fluctuations can raise growth, for example, since the opportunity cost of productivity enhancing activities falls in recessions (Aghion and Saint-Paul 1998) and conflicting results have been found in empirical work. Nevertheless, the preponderance of evidence points fairly clearly to a negative effect on investment and growth of unexpected volatility, uncertainty, or the variance of innovations to a forecasting equation for growth (Bloom 2014; Rafferty 2005; Ramey and Ramey 1995), and this implies that a benign macroeconomic environment could have been beneficial for long-run growth outcomes in the Golden Age.

Second, the reconstruction of the world economy after World War II was based on the Bretton Woods Compromise (Rodrik 2000), which severely restricted international capital mobility while seeking to liberalize trade. This implied a new macroeconomic trilemma choice in which priority was given to scope for independent monetary policy to manage the domestic economy rather than the efficient allocation of capital. In the1950s and 1960s, average current account positions as a fraction of GDP were at an all-time low (1.3% in 1960–1973), and in cross-sectional Feldstein-Horioka regressions to predict investment/GDP the coefficient on savings/GDP is

Table 4 Real GDP growth: average for G7 countries in four macroeconomic eras (% per year)

	G7 Mean	G7 Standard deviation
1881–1913 (Gold Standard)	1.5	3.7
1919–1938	1.2	6.8
1946–1970 (Bretton Woods)	4.2	2.7
1974–1989	2.2	2.3

Source: Bordo (1993)

0.92 during 1946–1972 (Obstfeld and Taylor 2004). This imposed a domestic savings constraint on growth.

Third, the successful prevention of banking crises surely entailed lower growth in "normal years." Financial regulation reduced leverage in bank balance sheets and surely implied a higher cost of capital since Modigliani-Miller offsets are known to be imperfect (Miles et al. 2013). Capital controls also raised the cost of capital and adversely affected growth (Voth 2003). The period was characterized by high costs of adjusting capital stocks to the optimal level and also by high values of Tobin's *Q* in both West Germany and the United Kingdom (Crafts and Mills 2005). Overall, this suggests that investment and growth were constrained by the regulatory environment, but at present there is no study that quantifies the overall impact.

Structural Change

Some, but not all, European economies had substantial low-productivity agricultural sectors at the start of the Golden Age; for example, in 1950, 49% and 42% of employment was in agriculture in Spain and Italy, respectively, but only 5% in the UK. Kindleberger (1967) argued that in some cases fast growth could be understood in terms of a Lewis dual-economy model with elastic supplies of labor available to the industrial sector. Temin (2002) found that a statistically significant effect from adding the initial share of the labor force in agriculture to unconditional-convergence growth regressions for the 1960s and 1970s but not thereafter. He argued simply that this reflected the productivity gain from correcting a misallocation of resources.

The orthodox "shift-share" way to measure the contribution of structural change in employment to labor productivity growth is to calculate it as aggregate labor productivity growth minus a weighted average of intra-sectoral productivity growth rates based on initial employment shares, as, for example, in Maddison (1987). This assumes, however, that the intrasectoral productivity growth rates are unaffected by the labor transfer. This will not be the case, however, if there was surplus labor in agriculture (Kindleberger 1967) and/or the increased size of the manufacturing sector allowed economies of scale to be realized (Kaldor 1966) since one or both of the intrasectoral productivity growth rates will be increased. The problem facing cliometricians has been to evaluate this additional contribution.

Broadberry (1998) proposed a way to address the first of these issues. He suggested that in a declining sector, such as agriculture, a modified shift-share calculation should be used in which the actual should be replaced by a counterfactual labor productivity growth rate obtained by actual output growth minus national labor force growth. Table 5 reports the results of this method and compares them with those obtained using the orthodox method for a number of European economies. Not surprisingly, the contribution of structural change is much larger using Broadberry's method and is indeed quite large for Italy and Spain. Nevertheless, it must be accepted that although some correction for surplus labor probably is justified, this is no more than a rough and ready approach.

Table 5 Contribution of structural change to labor productivity growth, 1950–1973 (% per year)

	Orthodox measure	Broadberry measure
Denmark	0.24	1.10
UK	−0.12	0.31
Sweden	0.00	0.60
Netherlands	−0.31	0.29
France	0.00	0.52
West Germany	0.18	0.77
Italy	0.83	1.77
Spain	0.80	1.77

Source: Crafts and Toniolo (2008) who derived the estimates from data in van Ark (1996) based on Broadberry's methodology using a three sector (agriculture, industry, services) de-composition where agriculture is deemed to be the declining sector

Note: The orthodox approach considers the contribution of structural change equals $\Delta A_O/A_O - \Sigma \Delta A_i/A_i * A_i/A_O * S_i$ where A is labor productivity, S is share of employment, and subscripts o and i stand for the whole economy and sector i, respectively (Nordhaus, 1972). Broadberry (1998) modified this so that labor productivity growth in the case of declining sectors was measured using the overall national rate of labor force growth not the sectoral rate

The evidence on economies of scale in manufacturing is also far from satisfactory. Kaldor based his views on a belief in Verdoorn's Law which he interpreted as implying a positive relationship in manufacturing between the rate of growth of labor productivity and the rate of growth of employment. This could reflect various sources of economies of scale including learning by doing. On the whole, the literature has been rather skeptical of the validity of Verdoorn's Law and, in any case, most economists would look to better-specified estimating equations to try to infer the presence of scale economies.[5] The evidence from econometric investigations for European industries in the 1970s and 1980s is in fact somewhat mixed (Caballero and Lyons 1990; Henriksen et al. 2001) and might suggest that experience varies.

The best developed example of a structural change account of the Golden Age relates to Italy which experienced a major shift of labor from agriculture to industry and was able to expand internationally tradable manufacturing by sustaining a substantially undervalued exchange rate for about two decades. Rossi and Toniolo (1996) using a modified growth accounting technique pioneered by Morrison (1988) found evidence of significant economies of scale during the postwar period. Based on an approach suggested by Rodrik (2008), Di Nino et al. (2013) estimated that undervaluation contributed 0.6–1.2% per year to GDP growth in an economy where wage growth was lower than labor productivity growth in tradables in the context of a strong domestic flow of migrant labor. Putting these components together gives an account of Italian growth quite similar to that which a reader of Kaldor and Kindleberger might expect.

[5]Magacho and McCombie (2017) is a recent review which suggests that the proposition will be rejected by mainstream growth economists. Maddison (1987) attributed only a small part of the Solow residual during the Golden Age to scale economies, which he assumed amounted to 3% of GDP growth, but this was no more than a guess.

The Marshall Plan and the European Economic Community

The reconstruction of the European economy entailed not only investment but policy reforms. The Golden Age was notable not only for the Bretton Woods agreement but also for the Marshall Plan and the European Common Market. Both of the latter promoted European economic integration and it is this aspect which is the main focus of this section.

In terms of short-run static effects, trade liberalization can improve allocative efficiency and/or productive efficiency, i.e., given existing costs, factors of production are deployed more efficiently or production costs are lowered. Insofar as freer trade increases competition in product markets (through actual or potential entry), it may have both effects as market power is reduced and price-cost margins fall, while managers of firms are pressured to reduce costs to the minimum feasible (principal-agent problems are reduced). In terms of long-run dynamic effects, according to endogenous growth models, it is possible that the growth rate will rise as a result of economic integration. In a basic AK model, if investment (or more generally the rate of growth of the capital stock) responds positively, there is no tendency for diminishing returns to erode this initial effect so there is a "permanent" impact on growth. Perhaps more plausibly, if a larger market and/or more competition in product markets ensues from economic integration, this may raise the rate of innovation and total factor productivity (TFP) growth. Even so, in a perhaps more realistic (semi-endogenous) growth model, the trade-liberalization impact on the growth rate would be a transitory phenomenon reflecting a move to a higher level of output rather than faster trend growth.

Growth regressions can be used to estimate the effect of European economic integration on income growth. Here the most useful paper is Badinger (2005), which made an index of the level of integration for each EU15 country from 1950 to 2000, and in a panel-regression setting with suitable controls examined its relationship with growth and with investment. The integration index, which took account both of GATT liberalization and European trade agreements, shows that 55% of the protectionism of 1950 was eliminated between 1958 and 1975, a figure which then rose steadily to 87% by 2000. The results of the regressions were that changes in integration were positive for growth but that the level of integration had no effect and that changes in integration had somewhere between half and three quarters of their impact through investment with the remainder coming from changes in TFP. Across the EU15 as a whole, GDP was estimated to be 26% higher than if there had been no economic integration after 1950, with a narrow range from 21.6% for Sweden to 28.9% for Portugal. The peak effect on the level of income resulting from the rapid liberalization prior to 1975 would have raised the growth rate over the period by about 1% per year – impressive, but only about a quarter of the western European growth rate in a period of rapid catch-up growth (Crafts and Toniolo 2008). The implication of the results in Badinger (2005) is that European economic integration has had a sizeable impact on the level of income but has not had a permanent effect on the rate of growth. This amounts to rejecting the endogenous growth hypothesis and is line with investigations of the

impact of recent trade liberalizations using difference-in-difference approaches (Estevadeordal and Taylor 2013).[6]

The Marshall Plan was a major program of aid which transferred $12.5 billion from the United States to Western Europe during the years 1948–1951 and provided inflows to recipient countries which typically averaged about 2% of GDP per year.[7] It helped reduce the costs of the Bretton-Woods Compromise by speeding up European integration via trade liberalization and the conditionality that it entailed also promoted pro-market reforms that were conducive to growth. As De Long and Eichengreen (1993) stressed, rather than being a handout, the Marshall Plan was a "structural adjustment program" along the lines of the Washington Consensus – "the most successful ever" – which succeeded by raising productivity growth and steered Europe away from becoming Argentina, i.e., away from an inward-looking development strategy based on protectionism.

In particular, each country signed a bilateral treaty with the United States, which committed them to follow policies of financial stability and trade liberalization while the Organization for European Economic Co-operation (OEEC) provided "conditional aid" to back an intra-West European multilateral payments agreement; in 1950, recipients had to become members of the European Payments Union (EPU). This lasted until 1958 by which time intra-European trade was 2.3 times that of 1950 and a gravity-model analysis confirms that the EPU had a large positive effect on trade levels.[8]

The Marshall Plan worked by tipping the balance in favor of pro-market structural reforms that raised productivity growth rather than through a direct stimulus.[9] This analysis can be developed by thinking in terms of a 3-gap model (Bacha 1990). This takes into account that aid can have positive growth effects through relaxing savings, foreign-exchange, or fiscal constraints. Eichengreen and Uzan (1992) provided an analysis of this type. The bottom line is that the direct effect of an average inflow of 2% of GDP would have raised the growth rate by 0.3 percentage points during the years 1948–1951.

The establishment of the European Economic Community increased trade considerably. In 1958 the EEC was formed by the original six countries following the signing of the Treaty of Rome in 1957.[10] The signatories pledged to lay the foundations of "ever closer union" among the peoples of Europe and Article

[6]Estevadeordal and Taylor point to reductions in tariffs on capital goods with consequent increases in the capital to labor ratio as central to the positive levels effect of trade liberalization; for a similar argument applied to Golden Age Europe, see Cubel and Sanchis (2009).

[7]More details of how it worked can be found in Crafts (2013).

[8]See Eichengreen (1993) which also contains a detailed description of the logic and mechanics of EPU.

[9]Exclusion from the Marshall Plan and EPU postponed but did not necessarily preclude liberalization, as is shown by the 1959 reform in Spain which raised the growth rate by 1 percentage point per year for the next decade and a half (Prados de la Escosura et al. 2011).

[10]The six founder members were Belgium, France, Germany, Italy, Luxembourg and Netherlands.

2 committed members to form a customs union, to establish a common market and to harmonize policies. The EEC customs union was achieved in 1968, but the common market took much longer and awaited the Single European Act, which addressed nontariff barriers to trade, liberalized trade in services and ended capital controls and was (less than fully) implemented from 1992. Even so, trade costs between the six countries fell relatively rapidly (cf. Table 6). Using a gravity model, Bayoumi and Eichengreen (1995) estimated that intra-EEC trade among the original six members was increased by 3.2% per year between 1956 and 1973 implying that EEC membership may have raised income levels by 4–8% by 1970 (Eichengreen and Boltho 2008) based on the elasticity between trade volumes and income estimated by Frankel and Romer (1999).

Social Capability and Technological Congruence

These concepts are central to the explanation given by Abramovitz and David (1996) for the Golden Age. They emphasize that catch-up growth is not automatic but depends on "Social capability" and "technological congruence." Their approach is one of conditional convergence and their argument is that lack of social capability and technological congruence had held Europe back prior to World War II, but these constraints on growth were much reduced by the 1950s and 1960s (Nelson and Wright 1992). Technological congruence refers to the cost effectiveness of the leader's technology in the follower countries which can be undermined by differences in factor prices or market size. Social capability refers to the ability to assimilate new technology. Absorptive capacity is underpinned by education, skills, and economic competences including organizational effectiveness, appropriate business models, and training. Beyond this, however, social capability turns on institutions, economic policies, and the incentive structures that they imply, which affect the profitability of innovation and investment.

Table 6 Trade costs

	Germany-France	Germany-Italy	Spain-France	UK-France	UK-Italy	UK-Norway
1929	0.99	1.10	1.18	1.00	1.22	0.87
1938	1.33	1.12	2.26	1.21	1.54	0.98
1950	1.12	1.27	1.55	1.22	1.36	0.98
1960	0.91	1.01	1.52	1.22	1.25	0.91
1970	0.73	0.79	1.24	1.10	1.21	0.90
1980	0.55	0.61	0.89	0.74	0.86	0.69
1990	0.53	0.56	0.74	0.70	0.84	0.77
2000	0.61	0.66	0.70	0.75	0.90	0.88

Source: Data underlying Jacks et al. (2011) generously provided by Dennis Novy
Note: Trade costs are inferred using a gravity model and comprise both policy and non-policy barriers to trade; 1929–1938 estimates are not strictly comparable with those for 1950–2000; estimates that include Spain are for 1939 not 1938

There is strong evidence of a significant increase in the rate of technology transfer into Europe after World War II. Madsen (2007) estimated a model in which TFP is impacted by imports of knowledge embodied in high technology goods and concluded that this played a major part in the reduction of TFP gaps. Comin and Hobijn (2010) investigated diffusion processes and showed that adoption lags for new technologies became much shorter. These studies do not, however, account for the relative importance of increased social capability and technological congruence.

The best evidence on technological congruence was provided by Jerzmanowski (2007). He devised a method to decompose TFP gaps into a part due to inefficiency (distance from the production function) and a part due to technology (inferior production function due to inappropriateness of American technology). His results (Table 7) show that TFP gaps were still quite large in 1960 but that they were primarily due to shortfalls in efficiency rather than technology gaps. Subsequently, big reductions in these efficiency gaps by 1985 accounted for a good part of European catch-up. This tends to confirm both the importance of improvements in resource allocation and that American technology was congruent for early post-war Western Europe.

When he proposed the concept of social capability, Abramovitz (1986) stated that the problem was that no-one knew just what it meant or how to measure it. Since then, a good deal of progress has been made. Giorcelli (2017) provides an important example of improved absorptive capacity in Italy as a result of managerial training under the auspices of the Marshall Plan. She is able to compare firm performance where managers participated in sponsored visits to learn about American management practices with a control group that was excluded from the program. Productivity in the former grew by 52.3% relative to the latter over 15 years. This paper

Table 7 Decomposition of 1960 level of TFP into efficiency and technology components (USA = 1.00)

	TFP	Efficiency	Technology
Austria	0.60	0.64	0.94
Belgium	0.65	0.64	1.01
Denmark	0.69	0.68	1.01
Finland	0.62	0.60	1.04
France	0.72	0.71	1.01
Greece	0.49	0.57	0.86
Ireland	0.51	0.55	0.93
Italy	0.67	0.71	0.94
Netherlands	0.77	0.74	1.04
Norway	0.54	0.63	0.86
Portugal	0.57	0.66	0.87
Spain	0.64	0.74	0.86
Sweden	0.73	0.72	1.01
Switzerland	1.05	1.00	1.05

Source: Jerzmanowski (2007)
Note: TFP = Efficiency*Technology

indicates that management quality was important to realizing the potential of new technology and, especially in the context of research by Bloom and Van Reenen (2007), suggests that improvements in management might have contributed to enhanced social capability in postwar Europe. This is a promising area for more research.

Turning to institutions and incentive structures, when Abramovitz first put forward the idea of social capability, the fashionable thesis was that of institutional sclerosis as set out by Olson (1982). Put simply, this proposed that in the absence of shocks such as wars and invasions over time democracies tend to experience a proliferation of rent-seeking interest groups, which act as a constraint on growth. The logic of this argument would be that World War II and its aftermath significantly relaxed this constraint for some countries, notably, West Germany. This hypothesis has generated a large literature by economists and political scientists exploring cross-country evidence which has, on balance, been supportive of Olson (Heckelman 2007). European economic historians have, however, been distinctly skeptical. In the case of the prime example, Eichengreen and Ritschl (2009) summarized a literature emphasizing institutional continuity rather than change. Comin and Hobijn (2011) developed a neat test based on changes in the adoption lags for new technologies after World War II for technologies with and without a competing predecessor. They found that for the latter these fell by 13 years, whereas for the former they increased by 5 years. The Olson hypothesis predicts the opposite.

Another approach to assessing the impact of institutional and policy settings on growth has been to use indices of economic freedom. Countries that are economically free have secure property rights and effective enforcement of contracts, stable prices, low barriers to trade, and resources mainly allocated through the market. Because of data limitations, analyses incorporating an economic freedom variable in a growth-regression specification have examined post-1975 experience. On balance, papers that pay attention to issues of endogeneity (De Haan and Sturm 2000; Kacprzyk 2016) give some support to the hypothesis that either the level of or increases in economic freedom promote growth. Recently, Prados de la Escosura (2016) has developed a Historical Index of Economic Liberty that will allow the hypothesis to be tested for earlier periods. Table 8 shows a general tendency to increased economic liberty in Western Europe; this was the case comparing 1960–1964 with 1935–1939 for all countries except Greece and Norway. It seems possible that these data can be used to provide some further support for the Abramovitz and David (1996) claim that improved social capability was an important ingredient in the recipe for the Golden Age.

High Investment/Wage Restraint Cooperative equilibrium

The most striking hypothesis to explain enhanced social capability in post-war Western Europe is that of Eichengreen (1996) who argued that high investment rates which allowed successful exploitation of catch-up opportunities were facilitated by successful social contracts which sustained wage moderation by workers in

Table 8 Economic liberty (0–10)

	1935–1939	1950–1954	1960–1964	1970–1974
Austria	6.2	6.5	8.5	8.6
Belgium	7.4	8.0	8.4	8.0
Denmark	8.1	7.9	8.5	8.5
Finland	7.2	6.6	8.1	8.2
France	6.7	7.1	7.3	8.0
Germany	5.9	7.9	9.2	8.8
Greece	7.3	6.1	7.2	5.8
Ireland	7.7	8.1	8.6	8.2
Italy	5.9	7.2	8.1	7.8
Netherlands	8.2	7.6	8.5	8.5
Norway	8.3	7.5	8.0	7.7
Portugal	6.7	6.7	6.9	6.5
Spain	3.0	5.4	6.2	6.3
Sweden	8.4	8.0	9.1	8.5
Switzerland	8.0	8.9	9.0	8.9
United Kingdom	7.8	7.9	8.2	8.1

Source: Prados de la Escosura (2016)

return for high investment by firms. To achieve a cooperative equilibrium of this kind, it is necessary to solve a time inconsistency problem. In the event of wage restraint, capitalists can go back on their promise to invest and raise dividends instead. On the other hand, if investment is increased, then workers can abandon wage restraint and seek to appropriate the profits. "Corporatist" arrangements provided institutions to monitor capitalists' compliance and centralized wage bargaining that protected high-investment firms and prevented free-riding by subsets of workers. A game theoretic analysis suggests that the central foundation of a cooperative equilibrium with high investment and wage moderation is that both sides are patient take a long-term view of the payoffs to their decisions and have reason to expect those payoffs to be substantial (Cameron and Wallace 2002).

Table 9 records the evolution of systems of industrial relations. Compared with 1925, in 1950 there were more countries classified as "corporatist" or "neo-corporatist," characterized by coordination in wage bargaining, and fewer with decentralized collective bargaining. In a growth-regression study, Gilmore (2009) found that coordinated wage bargaining may have had quite strong positive effects on investment and growth prior to 1975, but these were not found in subsequent periods. These results are supportive of Eichengreen's hypothesis. But this was not the only route to rapid catch-up growth as was apparent in the earlier discussion of the Lewis dual-economy model favored by Kindleberger. In Italy, for example, elastic supplies of labor to the industrial sector restrained wage increases and underpinned high investment at least until the late 1960s (Crafts and Magnani 2013).

If patience and optimism about large future rewards, rather than coordinated bargaining alone, are required to achieve the high investment with wage restraint cooperative equilibrium, then it may be quite fragile. As Cameron and Wallace

Table 9 A classification of industrial relations systems

	1925	1950	1963	1975
Corporatism	A,D,E	A,B,DK,N, NL,S	A,B,DK,N, NL,S	A,B,CH,D,DK,N, NL,S
Collective bargaining				
Neo-corporatist	CH	CH,D	CH,D	FIN
Decentralized	B,DK,N,NL,S, UK	IRL,UK	IRL,UK	IRL,IT,UK
Contestation	F, FIN, IRL	FIN,F,IT	FIN,F,IT	E,F,P
Authoritarian	IT,P	E,P	E,P	

Source: Crouch (1993)
Note: A = Austria, B = Belgium, D = Germany, DK = Denmark, E = Spain, FIN = Finland, F = France, IRL = Ireland, IT = Italy, N = Norway, NL = Netherlands, P = Portugal, S = Sweden

(2002) point out, the economic environment of the Golden Age was conducive to this outcome, whereas the turbulence of the 1970s was not. The cooperative equilibrium is more likely in a world of restricted capital mobility, fixed exchange rates, weak bargaining power for workers, and lots of scope for productivity growth. This describes the 1950s much better than the 1970s.

Relative Economic Decline in the UK

A brief look at the UK during the Golden Age is instructive. As was noted earlier, growth was relatively slow compared with other European countries. Some of this reflects less scope for catch-up including having an already small agricultural sector in 1950. But there was more to it than that notwithstanding the claim to the contrary by Temin (2002). So, while the Janossy and Kindleberger hypotheses are relevant to the British case and would, of course, predict that economic growth would be slower than West Germany and Italy, for example (cf. Tables 3 and 5), there is nothing in these models that would predict the UK's slide down the European ranking of labor productivity from 2nd in 1950 to 10th in 1973. In this context, being overtaken connotes avoidable failure even though simplistic growth regressions might seem to say otherwise. A case-study approach allows the cliometrician further insights, especially with regard to the role played by social capability.

Britain did not achieve the transformation of industrial relations that happened elsewhere in Europe which implied a considerable growth penalty. When it is not possible to write binding contracts, either the absence of unions or strong corporatist trade unionism would have been preferable to the idiosyncratic British system. This can readily be understood in terms of the Eichengreen model or an extension of it to incorporate endogenous innovation. In Britain, it was generally not possible to make the corporatist deals to underpin investment and innovation because bargaining took place with multiple unions or with shop stewards representing subsets of a firm's workforce who could not internalize the benefits of wage restraint. This exposed

sunk-cost investments to a "hold-up" problem.[11] In the terminology of Hall and Soskice (2001), the UK was a "liberal market economy," whereas a "co-ordinated market economy" was the foundation of the Eichengreen model.

Failure to successfully reform industrial relations was a major shortcoming of British governments from the 1950s through the 1970s. Throughout this period, there were continual efforts to persuade organized labor through an informal social contract to accept wage moderation in the interests not only of encouraging investment, but even more to allow low levels of unemployment without inflation at a time when politicians believed that this was crucial to electoral success after the interwar trauma. At worst, this was tantamount to allowing a de facto trade union "veto" on economic reforms and certainly obstructed industrial-relations reform. In any event, British supply-side policy, shaped by the "post-war consensus," was unhelpful towards growth in several respects.[12] These included a tax system characterized by very high marginal rates, described by Tanzi (1969) as the least conducive to growth of any of the OECD countries in his study; missing out on benefits from trade liberalization by retaining 1930s protectionism into the 1960s (Oulton 1976); a misdirected technology policy that focused on invention rather than diffusion (Ergas 1987); an industrial policy that ineffectively subsidized physical investment (Sumner 1999) and slowed down structural change by protecting ailing industries through subsidies (Wren 1996); and tariffs (Greenaway and Milner 1994).

A key feature of the Golden-Age British economy was the weakness of competition in product markets, which had developed in the 1930s and intensified subsequently. Competition policy was largely ineffective, protectionism continued through the 1960s, and market power was substantial. The evidence shows that weak competition interacted with the institutions, notably the systems of industrial relations and corporate governance, to undermine British productivity performance during the Golden Age (Crafts 2012). The rents resulting from weak competition were shared with trade unions partly through effort bargains that entailed overstaffing as was revealed in the 1980s when competition subsequently intensified (Machin and Wadhwani 1989). Nickell et al. (1997) estimated that for firms without a dominant external shareholder to control managers (the norm for big British firms at this time), an increase in supernormal profits from 5% to 15% of value added would reduce total factor productivity growth by 1 percentage point.

All this confirms that the emphasis of Abramovitz (1986) on social capability is well placed. It also goes beyond what has been possible with the growth regression

[11] In the endogenous innovation framework the "hold-up" arises when after a successful innovation workers use their bargaining power to extract a share of the profits. This reduces the incentive to innovate and thus the rate of growth. The more unions are involved in the bargaining, the more profits are reduced. The problem can be eliminated if a binding contract prevents renegotiation or there is no union or if a cooperative equilibrium is achieved with a single union. For a formal model and empirical evidence, see Bean and Crafts (1996).

[12] The concept of the "post-war consensus" should be understood as the set of policies regarded as feasible by senior politicians and civil servants given presumed political constraints (Kavanagh and Morris 1994). This implied a high degree of policy convergence but did not connote ideological convergence between the Conservative and Labour parties (Hickson 2004).

approach discussed earlier partly by flagging up the importance of competition and its interaction with institutions, but also by suggesting that the past places important constraints on institutional reforms and policy choices. The case-study approach allows a deeper analysis of the underlying reasons for success and failure in growth performance.

What Explains the Big Slowdown After the Golden Age?

After the early 1970s, growth slowed down markedly right across Europe. The end of the Golden Age had a number of unavoidable aspects including the exhaustion of transitory components of fast growth such as postwar reconstruction, reduced opportunities to redeploy labor out of agriculture, narrowing of the technology gap, and diminishing returns to investment. Moreover, the United States itself experienced a productivity growth slowdown. All-in-all, the scope for catch-up growth was considerably reduced although by no means eliminated. There were big reductions in the contributions of capital deepening and, especially, TFP growth to labor productivity growth (cf. Table 2). Although there were unavoidable reasons why productivity growth slowed down and European countries generally continued to narrow the productivity gap with the United States until the 1990s, it is clear that productivity performance could have been better after the Golden Age. Explaining this undue slowdown in productivity growth is a worthwhile future project for cliometricians in any case, but in addition, this later experience can provide further insights into Golden-Age growth.

Incomplete Catch-Up

The process of catch-up growth typically entails a series of ongoing reforms with the danger that at some point the political economy of the next step in modernization becomes too difficult. As modern growth economics stresses (Aghion and Howitt 2006), the institutions and policy choices that can galvanize a far-from-frontier economy differ in many ways from what is appropriate for a close-to-frontier economy. In particular, in the latter case, stronger competition in product markets and high-quality education become more important. A strong capacity for creative destruction matters more as countries become more advanced (Acemoglu et al. 2006). As new technologies come along, institutions and policies may need to be reformed (Abramovitz 1986). Yet, making the requisite adjustments may be problematic – history matters – and achieved only slowly and incompletely such that catch-up growth falters. Arguably, European countries needed reforms after the Golden Age to be well positioned for a later stage of growth but were slow to make this transition (Eichengreen 2007).

An important aspect of the changing nature of productivity growth as countries get closer to the frontier is that the weight of manufacturing in the economy declines while that of services increases. This typically contributes directly to a slowdown in growth as catch-up proceeds since the rate of productivity advance in manufacturing

Table 10 Contributions to labor productivity growth in the market sector, 1950–2005 (% per year)

	EU			USA		
	1950–1973	1980–1995	1995–2005	1950–1973	1980–1995	1995–2005
Labor productivity growth						
Manufacturing	5.0	3.2	2.0	2.7	2.1	2.9
Market services	3.1	1.6	1.0	2.2	1.6	3.2
ICT production	N/A	4.9	6.5	N/A	5.9	10.0
Contribution to total labor productivity growth						
Manufacturing	1.81	0.89	0.48	0.98	0.52	0.58
Market services	0.85	0.58	0.44	0.96	0.69	1.57
ICT production	N/A	0.33	0.42	N/A	0.52	0.81

Sources: Timmer et al. (2010); van Ark (1996)
Notes: Contributions are (%value added*productivity growth) for 1980–2005 and (%employment*-productivity growth) for 1950–1973. "EU" is aggregate of eight countries (Denmark, France, Germany, Italy, Netherlands, Spain, Sweden, and UK) in 1950–1973 and an aggregate of 10 countries (Austria, Belgium, Denmark, Finland, France, Germany, Italy, Netherlands, Spain, and UK) in 1980–2005. ICT production is included in manufacturing and market services in 1950–1973

generally exceeds that in services. It also has the implication that supply-side policies and institutions have to support a different type of economy. These issues are reflected in Table 10. European countries experienced a substantial decrease in the productivity-growth contribution of manufacturing after the Golden Age, as might be expected. In addition, however, the contribution from services was disappointing and, at the end of the century, Europe was unable to emulate the marked increase in productivity growth in market services achieved by the United States.

Social Capability in Different Technological Eras

In the 1950s and 1960s, Italy prospered and the UK floundered in the era of Fordist manufacturing, but in the ICT revolution at the end of the twentieth century, the opposite was true. This illustrates both that social capability depends on circumstances and that continued success in diffusion of new technologies may require reform. Each of these issues is ripe for cliometric research.

The acceleration in American productivity growth after 1995 was underpinned by ICT. For most countries, the main impact of ICT on economic growth comes through its diffusion as a new form of capital equipment rather than through total factor productivity (TFP) growth in the production of ICT equipment. The implication is that Europe had a great opportunity to significantly increase its productivity growth. However, growth-accounting estimates of the contribution of ICT-capital deepening to the growth of labor productivity show that European countries were generally much less successful than the United States in exploiting the potential of this new general purpose technology (Timmer et al. 2010). Interestingly, it was also true that American multinationals operating in Europe achieved significantly higher productivity from investments in ICT and used more ICT capital than other multinationals

or domestic firms, and that a major reason for this was their management practices (Bloom et al. 2012). In other words, the absorptive capacity of American-owned firms was bigger in the context of ICT.

Restrictive regulation of labor and product markets and, in some cases, shortfalls in human capital explain Europe's sluggish take up of ICT (Cette and Lopez 2012). Restrictive product market regulation deterred investment in ICT capital directly (Conway et al. 2006) and the indirect effect of regulation through raising costs was relatively pronounced in sectors that use ICT intensively. There was a strong correlation between product market regulation and the contribution of ICT-using services (notably in distribution) to overall productivity growth (Nicoletti and Scarpetta 2005). Notably, employment protection has been shown to deter investment in ICT equipment (Gust and Marquez 2004) because it increases the costs of reorganizing working practices and upgrading the labor force, which are central to realizing the productivity potential of ICT (Brynjolfsson and Hitt 2003). Since these forms of regulation have weakened over time, the story is not that European regulation became more stringent, but rather that existing regulation became more costly (and social capability decreased) in the context of a new technological era.

Italy has experienced major obstacles to the rapid diffusion of ICT for which it was not well positioned. The effective assimilation of this new technology has been hindered by the small size of firms, oppressive regulation, and shortfalls in human capital by comparison with the European leaders in the take up of ICT, as micro-economic studies confirm (Crafts and Magnani 2013). For the UK, success in ICT diffusion was an unintended and unexpected consequence of economic reforms in the Thatcher period. For the UK, the 1980s' de-regulation of services that are intensive in the use of ICT (notably finance and retailing), which reduced barriers to entry, and reform of industrial relations was important for its relatively successful response to the new technology. The move away from the postwar consensus removed any impetus towards employment protection, and rapid expansion of higher education proved timely (Crafts 2015).

Supply-Side Policy

There are two ways in which supply-side policy may bear some responsibility for the post-Golden Age slowdown. Reforms may have had an adverse effect on growth or there may have been failures to upgrade policies to sustain catch-up growth better. The literature has found examples of both problems.

Europe's route to catch-up growth brought with it increasing demands for social protection. Partly, this came simply as a result of raising income levels, but to a large extent it resulted from greater openness as European integration and globalization advanced (Lindert 2004). The median European economy spent 21.1% of GDP on social transfers in 1980 compared with 10.5% in 1960 and 1.2% in 1930. Managing these demands without undermining growth was an important challenge; insofar as they were financed by "distortionary" taxation, this became a drag on growth. Financing this expansion of government outlays by a different tax mix would have

been considerably better for growth (Johansson et al. 2008). The similar estimates of Kneller et al. (1999) indicate that the average 10 percentage point increase in the share of direct tax revenues in GDP between 1965 and 1995 could have entailed a fall in the growth rate of about 1 percentage point. In some countries, there were also marked increases in employment protection levels through the troubled 1970s. High levels of employment protection slow down the process of creative destruction and the labor force adjustment that it entails. The difference in employment protection between France and the United States could account for a difference of 0.5 percentage points per year in labor productivity growth in the 1980s and 1990s according to the estimates in Caballero et al. (2004).

From the 1970s through the 1990s, the impetus to economic growth from European integration continued, notably, through enlargements which expanded membership to 15 countries by 1995 and the inauguration of the European Single Market in 1992. An analysis based on a synthetic counterfactuals method suggests that the impact of EU accession on economic growth varied considerably across countries but was generally positive and in some cases provided a significant boost to growth (cf. Table 11). However, the impact of the Single Market has disappointed probably because nontariff barriers to trade have been lowered by less than might have been hoped. For example, Ilzkovitz et al. (2007) estimated GDP had been raised by 2.2% by 2006, well below the 4.8–6.4% projected ex-ante by the European Commission. Establishing a true Single Market in services could probably double this impact by reducing barriers to entry, but governments have had considerable discretion to maintain these barriers (Badinger and Maydell 2009). A recent estimate is that implementation of the Services Directive had raised EU GDP by about 0.8%, whereas full implementation would have tripled this (Monteagudo et al. 2012).

It is also relevant to look at the progress that European countries made in the upgrading needed as they moved closer to the frontier, in particular with regard to education and competition, the areas stressed by Aghion and Howitt (2006). A measure of cognitive skills shown, based on test scores, correlates strongly with growth performance (Hanushek and Woessmann 2012), and it is striking that the EU15 average for the late twentieth century is about 40 points less than Japan and

Table 11 Post-accession differences between level of actual and synthetic GDP per person (%)

	After 5 years	After 10 years
Denmark (1973	10.3	14.3
Ireland (1973)	5.2	9.4
United Kingdom (1973)	4.8	8.6
Greece (1981)	−11.6	−17.3
Portugal (1986)	11.7	16.5
Spain (1986)	9.3	13.7
Austria (1995)	4.5	6.4
Finland (1995)	2.2	4.0
Sweden (1995)	0.8	2.4

Source: Campos et al. (2014)
Note: accession dates in parenthesis

Table 12 Competition policy indicator (0–1)

	1995	2005
France	0.38	0.42
Germany	0.49	0.52
Italy	0.41	0.44
Netherlands	0.42	0.53
Spain	0.36	0.42
Sweden	0.69	0.66
United Kingdom	0.31	0.60
USA	0.59	0.62

Source: Buccirossi et al. (2013)
Note: First year for Netherlands is 1998 and for Spain is 2000. The index takes account of the independence of the competition authority and its resources, scope of investigative powers, and sanctions

South Korea which, according to the growth regressions for 1960–2000 presented in the paper, would mean growth was lower by about 0.5 percentage points per year.[13] The implication is that quality of schooling is an area where Europe could have done better. Woessmann et al. (2007) show that the variance in outcomes in terms of cognitive skills is largely explained by the way the schooling system is organized rather than educational spending.

Slowness to relax strict product market regulation (PMR) has raised mark-ups and lowered entry rates, thus reducing competitive pressure on managers with adverse impacts on both investment and innovation (Griffith and Harrison 2004; Griffith et al. 2010), and reduced European TFP growth relative to the United States in the late twentieth century by around 0.75 percentage points on average based on the estimates in Nicoletti and Scarpetta (2005). Similarly, in many European countries, competition policy remained much weaker than in the United States (Table 12). The econometric analysis in Buccirossi et al. (2013) found that this held back TFP growth; a reduction from a competition policy score of 0.69–0.36 (Sweden to Spain in 1995) is estimated to cut TFP growth by about 0.3 percentage points per year.

The Celtic Tiger

Ireland was, of course, a striking exception to the general picture of disappointing growth; far from slowing down, growth accelerated dramatically from the late 1980s to the end of the century. Table 13 reports growth data from the 1990s. It should be noted that GNP, which is a better measure of Irish performance, grew more slowly than GDP and that labor productivity grew less rapidly than GNP per person because hours worked increased considerably. Nevertheless, this is a very impressive performance reminiscent of other European countries in the Golden Age.

[13]Some countries have improved test scores quite rapidly, notably, Finland while elsewhere, for example, Italy, there is retrogression.

Table 13 Economic growth in Ireland, 1990–2000 (% per year)

Real GDP	7.5	Real GDP/Hour Worked	4.7
Real GNP	6.8	Real GNP/Hour Worked	4.0
Population	0.8	TFP	2.5
Hours Worked	2.8	Real GNP/Person	6.0

Source: Crafts (2014)
Note: TFP is on a GNP basis

At the start of the Celtic Tiger period, Ireland had considerable scope for catch-up growth. In 1987 labor productivity was 48% of the US level and only just above 50% of the European leaders. Unemployment was 17.2% of the labor force. Ireland had underperformed in the Golden Age partly because it was slow to abandon protectionism, had a malfunctioning labor market, and went through a period of macroeconomic disarray prior to a successful stabilization in the late 1980s. As we might expect, the rapid growth of the 1990s depended on favorable supply-side policies and good institutions. It was also predicated on the continuing globalization that characterized the late twentieth century.

At one level, the Celtic Tiger is just a story of delayed catch-up based on better policies and improved Social capability, but at the same time, there were key features which made Irish growth a special case. In this period Ireland had, by European standards, relatively low direct taxation and employment protection together with high quality education and had become a very open economy as a member of the EU (Crafts 2005). These features made Ireland an exceptionally attractive location for foreign direct investment (FDI) which was central to Ireland's development strategy. A major result of FDI was a very large ICT production sector which accounted for a much higher share of gross output than in any other EU country, including Finland, and contributed a little over 2 percentage points per year to TFP growth during the 1990s (van Ark et al. 2003).

Ireland's success in attracting FDI was based largely on its idiosyncratic corporate tax regime. It is clear from the literature that the semi-elasticity of FDI with respect to the corporate tax rate is quite high, perhaps of the order of −2.5 or even −3.5 (OECD 2007). At the start of the Celtic Tiger period, the Irish tax rate for manufacturing FDI was easily the lowest in Europe, and a study by Gropp and Kostial (2000) suggested that the stock of American manufacturing investment in Ireland was about 70% higher than if Ireland had had a tax rate equivalent to the next lowest in the EU. As trade costs fell, the impact of low taxes on FDI appears to have been accentuated significantly and their relative importance for location compared with proximity to demand increased (Romalis 2007).

Rapid employment growth in the 1990s came from a combination of large reductions in unemployment, which had fallen to 4.6% by 2000; a change in net migration flows that saw the tradition outmigration turn into net inflows that amounted to 67,000 between 1987 and 2000; and rising labor force participation, especially of women. The period also saw a large reduction in the NAIRU underpinned by wage moderation under the auspices of social partnership (Baccaro and Simoni 2007) and increases in human capital per worker (Bergin and Kearney 2004).

In the context of favorable shocks to labor demand, an unusually elastic labor supply prolonged the boom (Barry 2002).

Although it is not usually seen in this light, Ireland's success story has echoes of themes in the wider literature on the Golden Age. Three points stand out here. First, as in Germany during the 1950s, rapid growth was underpinned by putting unused resources back to work – in this case by substantially increasing the employment rate. Second, there was an elastic labor supply based on improvements in the workings of the domestic labor market and the historic switch from net emigration to net immigration. This encouraged wage moderation and helped sustain investment (Barry and Devereux 2006) and can be seen as having some similarity to the Lewis-type model favored by Kindleberger. Third, wage restraint was also achieved under the auspices of a social contract in which the quid pro quo was cuts in personal taxation. The standard and top rates of income tax fell by about 10 percentage points between the late 1980s and the late 1990s during which time tax cuts amounted to about a third of the growth in real take-home pay (Barry 2002). Here we see a permutation on the Eichengreen co-operative equilibrium.

Insights for the Golden Age

This research on the slowdown period highlights the potential importance of conditions favorable to rapid growth in the Golden Age which have not yet attained much prominence in cliometric work at least partly because they are hard to quantify for the early postwar years. These include the following: First, the importance of the quality of education and development of cognitive skills rather than just years of schooling in the accumulation of human capital should be underlined: Second, as the ICT revolution has highlighted, absorptive capacity of firms is a key determinant of the rate of diffusion of new technologies. In turn, this may be strongly influenced by the quality of management: Third, competition affects productivity performance, which not only means that competition policy deserves some attention but also may be an important aspect of the impact of European economic integration.

Conclusions

This chapter has considered two questions, namely, "what explains rapid economic growth in Western Europe during the Golden Age?" and "why did Western European growth slow down so markedly after the Golden Age?" It has focused mainly on the first of these questions on which cliometricians have made important contributions.

A number of (not mutually exclusive) hypotheses have support in the cliometrics literature. They establish a basis for understanding the rapidity of growth in the boom years, but the details vary in different countries. Reconstruction made a strong contribution to growth in some countries especially in the 1950s, as the Janossy hypothesis suggests, but while this had a big impact in West Germany it was irrelevant for Switzerland. In accordance with Kindleberger's analysis, the transfer

of labor out of agriculture had a significant impact on growth where it was feasible. For a country like Italy, this was a key component of Golden-Age growth, but it was of little consequence for the Netherlands. There is evidence also to support Eichengreen's model of a cooperative equilibrium featuring wage restraint in return for high investment based on coordinated wage bargaining in a sizeable subset of countries, but while this benefited Sweden it did not materialize in Ireland. Economic integration had a substantial impact on income levels, especially for those countries which signed the Treaty of Rome. The UK, a relatively slow-growing economy, did not benefit from any of these growth stimulants.

The notion of "social capability" has proved somewhat elusive in quantitative research, as Abramovitz expected. The Olson hypothesis of the absence of institutional sclerosis as a positive for some fast-growing countries has been viewed with skepticism, but it is surely the case that supply-side policy did matter and this seems to be borne out quite strongly by the examples of Ireland and the UK in the Golden Age during which both experienced serious policy failures and underperformed. Recent research has returned to the issue of absorptive capacity as a key aspect of technology transfer and this seems a priority area for future research.

The Keynesian hypothesis advanced by Boltho that macroeconomic stability was a key condition underlying the Golden Age has been neglected in cliometric research, although everyone would agree that the absence of financial crises or a depression was a big plus. In this context, a potentially important development is recent interest in, and improved methods for, measuring uncertainty as an important influence on investment. This methodology based on textual analysis could be a useful way to reconsider the role of the macroeconomic environment and policy framework. This also seems a good candidate for research effort.

The starting point for thinking about the post-Golden Age growth slowdown is to note that it did not simply reflect the exhaustion of transitory components, important though that was. The key lesson is that economic policy does matter and that many European countries were slow to make necessary reforms. The shining exception to this generalization was Ireland, which enjoyed fast growth in the late twentieth century on the basis of successful policy reforms which delivered an idiosyncratic and belated version of the Golden Age.

Cross-References

► Cliometrics of Growth

References

Abramovitz M (1986) Catching up, forging ahead, and falling behind. J Econ Hist 46:385–406

Abramovitz M, David PA (1996) Convergence and delayed catch-up: productivity leadership and the waning of American exceptionalism. In: Landau R, Taylor T, Wright G (eds) The mosaic of economic growth. Stanford University Press, Stanford, pp 21–62

Acemoglu D, Aghion P, Zilibotti F (2006) Distance to frontier, selection, and economic growth. J Eur Econ Assoc 4:37–74
Aghion P, Howitt P (1992) A model of growth through creative destruction. Econometrica 60:323–351
Aghion P, Howitt P (2006) Appropriate growth theory: a unifying framework. J Eur Econ Assoc 4:269–314
Aghion P, Saint-Paul G (1998) On the virtue of bad times: an analysis of the interaction between economic fluctuations and productivity growth. Macroecon Dyn 2:322–344
Baccaro L, Simoni M (2007) Centralized wage bargaining and the 'Celtic Tiger' phenomenon. Ind Relat 46:426–469
Bacha EL (1990) A three-gap model of foreign transfers and the GDP growth rate in developing countries. J Dev Econ 32:279–296
Badinger H (2005) Growth effects of economic integration: evidence from the EU member states. Rev World Econ 141:50–78
Badinger H, Maydell N (2009) Legal and economic issues in completing the EU internal market for services: an interdisciplinary perspective. J Common Mark Stud 47:693–717
Barro RJ (1991) Economic growth in a cross-section of countries. Q J Econ 106:407–443
Barry F (2002) The Celtic Tiger Era: delayed convergence or regional boom? Q Econ Commentary 21:84–91
Barry F, Devereux MB (2006) A theoretical growth model for Ireland. Econ Soc Rev 37:245–262
Baumol WJ (1986) Productivity growth, convergence and welfare: what the long-run data show. Am Econ Rev 76:1072–1085
Bayoumi T, Eichengreen B (1995) Is regionalism simply a diversion? Evidence from the evolution of the EC and EFTA. NBER working paper no. 5283
Bean C, Crafts N (1996) British economic growth since 1945: relative economic decline … and renaissance? In: Crafts N, Toniolo G (eds) Economic growth in Europe Since 1945. Cambridge University Press, Cambridge, pp 131–172
Bergin A, Kearney I (2004) Human capital, the labour market and productivity growth in Ireland. ESRI working paper no. 158
Bloom N (2014) Fluctuations in uncertainty. J Econ Perspect 28(2):153–176
Bloom N, Van Reenen J (2007) Measuring and explaining management practices across firms and countries. Q J Econ 122:1351–1408
Bloom N, Sadun R, Van Reenen J (2012) Americans do IT better: US multinationals and the productivity miracle. Am Econ Rev 102:167–201
Boltho A (1982) Introduction. In: Boltho A (ed) The European economy: growth and crisis. Oxford University Press, Oxford, pp 9–37
Bordo M (1993) The Bretton Woods international monetary system: a historical overview. In: Bordo M, Eichengreen B (eds) A retrospective on the Bretton Woods system: lessons for international monetary reform. University of Chicago Press, Chicago, pp 3–108
Bosworth BP, Collins SM (2003) The empirics of growth: an update. Brook Pap Econ Act 2:113–206
Broadberry SN (1998) How did the United States and Germany overtake Britain? A sectoral analysis of comparative productivity levels, 1870–1990. J Econ Hist 58:375–407
Brynjolfsson E, Hitt L (2003) Computing productivity: firm-level evidence. Rev Econ Stat 85:793–808
Buccirossi P, Clari L, Duso T, Spagnolo G, Vitale C (2013) Competition policy and economic growth: an empirical assessment. Rev Econ Stat 95:1324–1336
Caballero RJ, Lyons RK (1990) Internal versus external economies in European Industry. Eur Econ Rev 34:805–830
Caballero R, Cowan K, Engel E, Micco A (2004) Effective labor regulation and microeconomic flexibility. NBER working paper no. 10744
Cameron G, Wallace C (2002) Macroeconomic performance in the Bretton Woods Era and after. Oxf Rev Econ Policy 18:479–494

Campos NF, Coricelli F, Moretti L (2014) Economic growth and political integration: estimating the benefits from membership of the European Union using the synthetic counterfactuals method. CEPR discussion paper no. 9968
Cette G, Lopez J (2012) ICT demand behaviour: an international comparison. Econ Innov New Technol 21:397–410
Comin D, Hobijn B (2010) An exploration of technology diffusion. Am Econ Rev 100:2031–2059
Comin D, Hobijn B (2011) Technology diffusion and postwar growth. NBER Macroecon Annu 2010 25:209–259
Conway P, de Rosa D, Nicoletti G, SteinerF (2006) Regulation, competition and productivity convergence. OECD economics department working paper no. 509
Crafts N (2005) Interpreting Ireland's economic growth. Background paper for UNIDO industrial development report 2005
Crafts N (2012) British relative economic decline revisited: the role of competition. Explor Econ Hist 49:17–29
Crafts N (2013) The Marshall Plan. In: Parker R, Whaples R (eds) Routledge handbook of major events in economic history. Routledge, London, pp 203–213
Crafts N (2014) Ireland's medium term growth prospects: a phoenix rising? Econ Soc Rev 45:87–112
Crafts N (2015) Economic growth: onwards and upwards? Oxf Rev Econ Policy 31:217–241
Crafts N, Magnani M (2013) The golden age and the second globalization in Italy. In: Toniolo G (ed) The Oxford handbook of the Italian economy since unification. Oxford University Press, Oxford, pp 69–107
Crafts N, Mills TC (1996) Europe's golden age: an econometric investigation of changing trend rates of growth. In: Van Ark B, Crafts N (eds) Quantitative aspects of postwar European economic growth. Cambridge University Press, Cambridge, pp 415–431
Crafts N, Mills TC (2005) TFP growth in British and German manufacturing, 1950–1996. Econ J 115:649–670
Crafts N, Toniolo G (2008) European economic growth, 1950−2005: an overview. CEPR discussion paper no 6863
Crouch C (1993) Industrial relations and European state traditions. Clarendon Press, Oxford
Cubel A, Sanchis MT (2009) Investment and growth in Europe during the Golden Age. Eur Rev Econ Hist 13:219–249
De Haan J, Sturm J-E (2000) On the relationship between economic freedom and economic growth. Eur J Polit Econ 16:215–241
De Long JB (1988) Productivity growth, convergence and welfare: comment. Am Econ Rev 78:1138–1154
De Long JB, Eichengreen B (1993) The Marshall Plan: history's most successful adjustment program. In: Dornbusch R, Nolling W, Layard R (eds) Postwar economic reconstruction and lessons for the east today. MIT Press, Cambridge, MA, pp 189–230
Di Nino V, Eichengreen B, Sbracia M (2013) Real exchange rates, trade, and growth. In: Toniolo G (ed) The Oxford handbook of the Italian economy since unification. Oxford University Press, Oxford, pp 351–377
Dumke RH (1990) Reassessing the *Wirtschaftswunder*: reconstruction and post-war growth in West Germany in an international context. Oxf Bull Econ Stat 52:451–492
Eichengreen B (1993) Reconstructing Europe's trade and payments. Manchester University Press, Manchester
Eichengreen B (1996) Institutions and economic growth: Europe after World War II. In: Crafts N, Toniolo G (eds) Economic growth in Europe Since 1945. Cambridge University Press, Cambridge, pp 38–72
Eichengreen B (2007) The European economy since 1945. Princeton University Press, Princeton
Eichengreen B, Boltho A (2008) The economic impact of European integration. CEPR discussion paper no. 6820
Eichengreen B, Ritschl A (2009) Understanding West German economic growth in the 1950s. Cliometrica 3:191–219

Eichengreen B, Uzan M (1992) The Marshall Plan: economic effects and implications for Eastern Europe and the former USSR. Econ Policy 7(14):13–75
Ergas H (1987) Does technology policy matter? In: Guile BR, Brooks H (eds) Technology and global industry. National Academy Press, Washington, DC, pp 191–245
Estevadeordal A, Taylor A (2013) Is the Washington consensus dead? Growth, openness and the great liberalization, 1970s–2000s. Rev Econ Stat 95:1669–1690
Frankel JA, Romer D (1999) Does trade cause growth? Am Econ Rev 89:379–399
Gilmore O (2009) Corporatism and growth: testing the Eichengreen hypothesis. MSc. Dissertation, University of Warwick
Giorcelli M (2017) The long-term effects of management and technology transfers. Mimeo, UCLA
Greenaway D, Milner C (1994) Determinants of the inter-industry structure of protection in the UK. Oxf Bull Econ Stat 53:265–279
Griffith R, Harrison R (2004) The link between product market regulation and macroeconomic performance. European Commission economic papers no. 209
Griffith R, Harrison R, Simpson H (2010) Product market reform and innovation in the EU. Scand J Econ 112:389–415
Gropp R, Kostial K (2000) The disappearing tax base: Is FDI eroding corporate income taxes? IMF working paper no. 00/173
Grossman G, Helpman E (1991) Quality ladders in the theory of growth. Rev Econ Stud 58:43–61
Gust C, Marquez J (2004) International comparisons of productivity growth: the role of information technology and regulatory practices. Labour Econ 11:33–58
Hall PA, Soskice D (2001) An introduction to varieties of capitalism. In: Hall PA, Soskice D (eds) Varieties of capitalism. Oxford University Press, Oxford, pp 1–68
Hanushek EA, Woessmann L (2012) Do better schools lead to more growth?: cognitive skills, economic outcomes, and education. J Econ Growth 17:267–321
Heckelman JC (2007) Explaining the rain: the rise and decline of nations after 25 years. South Econ J 74:18–33
Henriksen E, Midelfart Knarvik KH, Steen F (2001) Economies of scale in European manufacturing revisited. CEPR discussion paper no. 2896
Hickson K (2004) The postwar consensus revisited. Political Q 75:142–154
Ilzkovitz F, Dierx A, Kovacs V, Sousa N (2007) Steps towards a deeper economic integration: the internal market in the 21st century. European Economy Economic Papers no. 271
Jacks DS, Meissner CM, Novy D (2011) Trade booms, trade busts, and trade costs. J Int Econ 83:185–201
Janossy F (1969) The end of the economic miracle. IASP, White Plains
Jerzmanowski M (2007) Total factor productivity differences: appropriate technology vs. efficiency. Eur Econ Rev 51:2080–2110
Johansson A, Heady C, Arnold J, Brys B, Vartia L (2008) Taxation and economic growth. OECD Economics Department working paper no. 620
Kacprzyk A (2016) Economic freedom-growth Nexus in European Union countries. Appl Econ Lett 23:494–497
Kaldor N (1966) Causes of the slow rate of growth of the United Kingdom. Cambridge University Press, Cambridge
Kavanagh D, Morris P (1994) The rise and fall of consensus politics. In: Kavanagh D, Morris P (eds) Consensus politics from Attlee to Major. Blackwell, Oxford
Kindleberger CP (1967) Europe's postwar growth: the role of labor supply. Harvard University Press, Cambridge, MA
Kneller R, Bleaney M, Gemmell N (1999) Fiscal policy and growth: evidence from OECD countries. J Public Econ 74:171–190
Lindert PH (2004) Growing public. Cambridge University Press, Cambridge
Lucas RE (1988) On the mechanics of economic development. J Monet Econ 22:3–42
Machin S, Wadhwani S (1989) The effects of unions on organisational change, investment and employment: evidence from WIRS Data. London School of Economics Centre for Labour economics discussion paper no. 355

Maddison A (1982) Phases of capitalist development. Oxford University Press, Oxford
Maddison A (1987) Growth and slowdown in advanced capitalist economies: techniques of quantitative assessment. J Econ Lit 25:649–698
Madsen JB (2007) Technology spillover through trade and TFP convergence: 135 years of evidence for the OECD countries. J Int Econ 72:464–480
Magacho GR, McCombie JSL (2017) Verdoorn's law and productivity dynamics: an empirical investigation into demand and supply approaches. J Post-Keynesian Econ 40:600–621
Miles D, Yang J, Marcheggiano G (2013) Optimal bank capital. Econ J 123:1–37
Monteagudo J, Rutkovski A, Lorenzani D (2012) The economic impact of the services directive: a first assessment following implementation. European Economy Economic Papers no. 456
Morrison CJ (1988) Unraveling the productivity growth slowdown in the US, Canada and Japan: the effects of sub-equilibrium, scale economies and mark-ups. Rev Econ Stat 74:381–393
Nelson RR, Wright G (1992) The rise and fall of American technological leadership: the postwar era in historical perspective. J Econ Lit 30:1931–1964
Nickell SJ, Nicolitsas D, Dryden N (1997) What makes firms perform well? Eur Econ Rev 41:783–796
Nicoletti, G. and Scarpetta, S. (2005), Regulation and economic performance: product market reforms and productivity in the OECD. OECD Economics Department working paper no. 460
Obstfeld M, Taylor AM (2004) Global capital markets: integration, crisis and growth. Cambridge University Press, Cambridge
OECD (2007) Tax effects on foreign direct investment. OECD, Paris
Olson M (1982) The rise and decline of nations. Yale University Press, New Haven
Oulton N (1976) Effective protection of British industry. In: Corden WM, Fels G (eds) Public assistance to industry. Macmillan, London, pp 46–90
Prados de la Escosura L (2016) Economic freedom in the long run: evidence from OECD countries, 1850–2007. Econ Hist Rev 69:435–468
Prados de la Escosura L, Roses J, Sanz Villaroya I (2011) Economic reforms and growth in Franco's Spain. Rev Hist Econ 30:45–89
Rafferty M (2005) The effects of expected and unexpected volatility on long-run growth: evidence from 18 developed countries. South Econ J 71:582–591
Ramey G, Ramey VA (1995) Cross-country evidence on the link between volatility and growth. Am Econ Rev 85:1138–1151
Rodrik D (2000) How far will international economic integration go? J Econ Perspect 14(1):177–186
Rodrik D (2008) The real exchange rate and growth. Brook Pap Econ Act (Fall):365–412
Romalis J (2007) Capital taxes, trade costs, and the Irish miracle. J Eur Econ Assoc 5:459–469
Romer PM (1986) Increasing returns and long-run growth. J Polit Econ 94:1002–1037
Rossi N, Toniolo G (1996) Italy. In: Crafts N, Toniolo G (eds) Economic growth in Europe since 1945. Cambridge University Press, Cambridge, pp 427–454
Summers R, Heston A (1984) Improved international comparisons of real product and its composition, 1950–1980. Rev Income Wealth 30:207–262
Sumner M (1999) Long-run effects of investment incentives. In: Driver C, Temple J (eds) Investment, growth and employment: perspectives for policy. Routledge, London, pp 292–300
Tanzi V (1969) The individual income tax and economic growth. Johns Hopkins University Press, Baltimore
Temin P (2002) The Golden Age of European growth reconsidered. Eur Rev Econ Hist 6:3–22
The Conference Board (2016) The Conference Board Total Economy Database, May 2016. http://www.conference-board.org/data/economydatabase/
Timmer M, Inklaar R, O'Mahony M, van Ark B (2010) Economic growth in Europe: a comparative economic industry perspective. Cambridge University Press, Cambridge
Van Ark B (1996) Sectoral growth accounting and structural change in postwar Europe. In: Van Ark B, Crafts N (eds) Quantitative aspects of postwar European economic growth. Cambridge University Press, Cambridge, pp 84–164

van Ark B, Melka J, Mulder N, Timmer M, Ypma G (2003) ICT investments and growth accounts for the European Union. Groningen Growth and Development Centre Research Memorandum GD−56

Van der Wee H (1986) Prosperity and upheaval: the world economy, 1945–1980. Penguin Books, Harmondsworth

Vonyo T (2008) Post-war reconstruction and the Golden Age of economic growth. Eur Rev Econ Hist 12:221–241

Voth H-J (2003) Convertibility, currency controls, and the cost of capital in Western Europe, 1950–1999. Int J Financ Econ 8:255–276

Woessmann L, Ludemann E, Schutz M, West MR (2007) School accountability, autonomy, choice and the level of student achievement: international evidence from PISA 2003. OECD education working paper no. 13

Wren C (1996) Industrial subsidies: the UK experience. Macmillan, London

GDP and Convergence in Modern Times

Emanuele Felice

Contents

Abstract

In this chapter, I discuss historical estimates of GDP at both the national and the regional level and their application for assessing economic performance in modern times. Having been invented in (and conceived for) industrial capitalist societies, GDP has stronger informative power in those contexts where industry and services, and market exchange, retain the lion's share of production. In modern times, when comparing the series available for different countries, there are three major methodological problems to be acknowledged and possibly addressed: the dissimilarity of the quantity series and related proxies, deflation through purchasing power parities distant in time, and the differences in the base year used to construct GDP constant price (Laspeyres) indices (the latter issue may be less widely recognized, but it may have a remarkable impact). The way

Financial support from the Spanish Ministry of Economy and Competitiveness, project HAR2013-47182-C02-01, and the Generalitat de Catalunya, project 2014 SGR 591, is gratefully acknowledged.

E. Felice (✉)
Dipartimento di Scienze Filosofiche, Pedagogiche ed Economico-Quantitative, Università "G. D'Annunzio" Chieti-Pescara, Pescara, Italy
e-mail: emanuele.felice@gmail.com

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_5

the estimates are constructed also has a bearing upon the statistical tools and models we should use to interpret them; owing to the lack of reliable long-run series, cross-sectional techniques are often preferable to time series analysis; provided we have reliable estimates, growth accounting – decomposing GDP growth into productivity and industry mix effects – may provide important clues about the choice between theoretical approaches; not least for the quality of our data, cross-country convergence models based on conditioning variables should always be supplemented by historical information from qualitative sources and case studies. More generally, cliometricians should prove themselves capable of adapting their models to different historical contexts and relativizing findings to the limits of their estimates.

Keywords
GDP · Convergence · Purchasing power parity · Neoclassical school · Endogenous growth · New economic geography

Introduction

To the extent that economics should use facts to verify theories, history is precious, being the fieldwork where empirical information can be found. Of course, information must be reliable: potential mistakes but also methodological differences can affect the results to the point that data cannot serve the purpose, all the more so in international comparisons. When we deal with historical GDP estimates – the primary indicator of any macroeconomic reasoning – what may appear less obvious is that in order to evaluate their soundness, we must rely not only upon historical knowledge but also on some basic expertise in quantitative techniques: economists may pick up a misleading series if they overlook the historical context, but non-quantitative historians can also accept the wrong figures if they are unable to assess the validity of the techniques used to produce them.

In this respect, quantitative economic historians – admittedly, a more comprehensive definition for cliometricians – are vital to both economics and more traditional history. From their historian backgrounds, they can provide a useful contribution to the former, insofar as they warn against a superficial approach to historical information (and estimates) based on the inattentive use of datasets and aprioristic assumptions about the past that do not meet the facts. They may even be able to contribute models that effectively account for historical change. Using their quantitative expertise, cliometricians may also help traditional historians understand why, and under which conditions, various models and estimates are useful descriptions of the past and tenable explanations for growth. In short, they can identify instances in which our historical interpretation should change according to the results proposed by quantitative history and economics. Such a double-sided task is not an easy one, because it implies that a good quantitative economic historian must have proficiency in both economics and history. However, the efforts have their rewards, as they may endow us with some of the most powerful instruments to understand the past.

GDP stands out among these instruments. It is virtually impossible for anyone studying economic growth to avoid using GDP estimates. Hence, it is important to understand how the series are constructed and what assumptions undergird the most popular growth models. However, it is also crucial to recognize that the choice of model and the interpretation of its results are informed and affected by the procedure employed to produce the figures. This chapter is dedicated to explaining and developing these issues. It reviews the procedures and uses of historical GDP estimates in modern times, roughly from the second half of the nineteenth century onward, at both the national and the regional level. In doing so, I highlight the main problems that can arise in terms of comparability between different estimates and make a case for improving explanatory models with an understanding of both the historical context and the GDP estimation procedures.

GDP: Concept, Limits, and Success

The production approach of calculating GDP considers it to be the sum of the final values of all the saleable goods and services produced within an economic system (a country or a region) over a certain period of time. Values are measured at market price, and they are final in the sense that they are net of the costs of intermediate goods and inputs to avoid duplication. According to the expenditure approach, GDP is the sum of consumption, investment, government spending, and net exports (exports minus imports). Finally, according to the income approach, GDP is the sum of all the incomes earned in that economic system (e.g., Lequiller and Blades 2006).[1] So many dimensions, into a single number: this is probably the ultimate reason behind its success. For instance, when divided by the number of inhabitants, total GDP corresponds to average income[2]; and when divided only by employment, it equals average per worker productivity. Production and expenditure, income, and productivity: the basics of any economic discourse cannot be addressed nowadays without GDP.

Less widely known is the fact that the most important measure of economic performance is a recent invention, at least from a historical perspective. It was born in the United States during the Great Depression in order to monitor the impact of the 1929 crisis and the time and pace of recovery (Carson 1975). It was then elaborated in the National Bureau of Economic Research, a private institute of empirically oriented scholars directed by Wesley Clair Mitchell, one of the leading figures in institutional economics (Schumpeter 1950). Further, it should be credited mostly to the work of Simon Kuznets: under his authorship, the first official estimates were

[1]For a country, GDP includes the incomes earned by the individuals not officially living in that country. Gross national production (GNP) includes instead the incomes earned abroad by the citizens of that country.

[2]To be consistent with the definition of the previous footnote, GDP should be divided by the population de facto (present population) and GNP by the resident population.

published in 1934, with reference to the US economy from 1929 to 1932 (Kuznets 1934). After World War II, in a western world governed by Keynesian policies (thus paying particular attention to cyclical fluctuations) and one strongly influenced by the economic and political power of the United States, GDP (and GNP)[3] turned into official statistics in Europe[4] and then throughout the world (although planned economies used a different system of national accounts). However, the origins of GDP should not be forgotten, at least from the point of view of cliometricians and economic historians, since they are essential in order to grasp the three basic features of the measure we are dealing with. First, GDP was conceived in an empirically oriented environment, as a sort of practical shortcut to solve the complex problem of how to monitor the economy, and thus it had strong theoretical limitations and even some related methodological contradictions. Second, it was born into an advanced industrial economy with the aim of measuring *that* economy, where industry (manufacturing) and services had by far the *magna pars* of national income to the detriment of agriculture (and mining) and where most of the production was sold and bought in the market. Third, it was created at a later stage in the history not only of the modern world but also of industrial capitalism as we have come to know it: it did not exist during the Industrial Revolution or in the first globalization era or at the time of World War I, not to mention medieval or ancient times.

There is now a vast literature on the theoretical limitations of GDP, which is of interest not only to economic historians and economists but also to social scientists and to an extent policymakers and the general public (Felice 2016). Nevertheless, some confusion on this should be sorted out. Some of the limitations of GDP are neither theoretical nor the result of a methodological contradiction. For instance, GDP is neither a measure of well-being nor the standard of living: it excludes the nonmonetary dimensions of well-being (from clean air to free time to the quality of affective life) while including other items that do not contribute directly to well-being but at best prevent it from falling (such as the expenditures on defense or on the administration of justice), and it does not consider the impact of the distribution of income on personal utilities. But there is no contradiction on this: GDP simply was not born for this purpose. GDP cannot be a measure of "human development" – at least as intended in the capability approach by Sen (1985) that was developed half a century after the creation of GDP – since it does not allow for other fundamental dimensions of human development, namely, education and longevity.[5] But again,

[3]The United States used GNP instead of GDP as late as 1991. By that time, virtually all the other countries had already adopted GDP.

[4]The first official estimates for the United Kingdom were made in 1941 by Richard Stone and James Meade. The former also was the main contributor to developing a standardized system that since 1952 was implemented in OEEC (Organization for European Economic Cooperation) countries (Stone 1956, 1961).

[5]However, many others are equally excluded: take, for instance, political and civil freedoms. Nussbaum (2000) increases up to ten the number of basic capabilities: (1) life; (2) bodily health; (3) bodily integrity; (4) sense, imagination, and thought; (5) emotion; (6) practical reason; (7) affiliation; (8) other species; (9) play; and (10) control over one's environment.

GDP was never designed to be a comprehensive measure of all the desired goals a human being can nurture, and so there is no contradiction or theoretical limitation in this. Rather, limitations are in those who regard GDP as the ultimate icon of human fulfillment. But even then, it is only fair to acknowledge that there is still no agreement about alternative measures to GDP that would better monitor non-monetary dimensions. Even the Human Development Index, which is gaining consensus among economic historians (Crafts 1997, 2002; Prados de la Escosura 2013, 2015), is far from undisputed for what concerns its formula, weights, and components (Prados de la Escosura 2010; Ravallion 2012a, b), let alone its theoretical foundations. This may be the fundamental reason why GDP, although *it is not* a measure of well-being and human development, was and still is often *considered to be one* or at least a measure of economic progress, broadly defined.

Similar arguments can be raised to oppose another well-known accusation brought against GDP: it excludes unpaid work (Waring 1988). This can have paradoxical effects, such as the often quoted textbook insight that having grandparents take care of children, instead of hiring domestic help, may cause a fall in GDP. But we need to remember that GDP was conceived when policymakers needed to contrast official unemployment, not unofficial employment. Less known but particularly telling is instead what happens with the mining sector, which actually represents a theoretical limitation (and even a methodological contradiction). At the time GDP was invented, the US census didn't ask firms owning their mines to declare the value of their reserves (Fenoaltea 2008). As a consequence, GDP does not compute the net value of production or value added (total mining production minus an estimate of the depletion of natural resources) but only the value of outputs. In other words, the more you consume your reserves, the more GDP (artificially) increases. The mining sector is important by itself, of course, but also for being part of a major problem. GDP has serious theoretical limitations in dealing with the environment. Not only does it not account for air and water pollution or land contamination, but indeed all these phenomena can even indirectly increase GDP, as long as they lead to the creation of specific counter-pollution activities in the market economy. This is probably the most worrying issue, which in the future may negate the ability of GDP to measure economic progress, at least until it is modified to account for some costs of pollution and the consumption of the planet's resources.[6] Of course, at the time GDP was invented, the concern for the environment was practically unknown in the United States or anywhere else.

The second and third characteristics of GDP should be of particular concern to cliometricians and economic historians. GDP was born in order to monitor advanced industrial economies, where most of the production comes from industry and services. In these sectors, there are two factors of production, labor (L) and capital (K), meaning that the standard growth model starts from the following production

[6]In this direction, some progress has recently been made, but with little or no heed, thus far, in the systems of national accounts: see Boyd and Banzhaf (2007) and Ferreira et al. (2008).

function: $Y = f(L, K)$. A widely accepted specification of this function is the Cobb-Douglas form

$$Y = A \times L^{\alpha} \times K^{\beta} \tag{1}$$

and in particular the one with $\alpha + \beta = 1$ (i.e., with constant returns to scale)

$$Y = A \times L^{\alpha} \times K^{1-\alpha}. \tag{2}$$

In both Eqs. 1 and 2, α and β (or $1 - \alpha$) are the output elasticities of labor and capital, respectively, and in Eq. 2, assuming perfect competition, α and $\beta = 1 - \alpha$ also are their respective shares of output (Douglas 1976). A stands for total factor productivity (TFP), a factor measuring the efficiency with which capital and labor are employed in production: this captures both the technological change not incorporated in capital and the gains of efficiency in production processes due to the reallocation of activities from one sector to another (Solow 1957). Provided that we find values for α, or for α and β, and that we reconstruct the amount of labor (number of workers or, better, number of hours of work) and the value of capital (the physical capital stock, in turn composed of machinery, infrastructure, and equipment; means of transport; nonresidential construction; housing), the growth rate of GDP (Y) can be decomposed into the contributions of increases in labor (L) and in capital (K) and of improvements in their combination (A). And even if we don't have values for α and β, whose historical estimates are usually far from undisputed, the formula clearly indicates that capital deepening (K) and TFP growth (A) bring about an increase in GDP per worker (Y/L). According to the simple equation $Y/P = Y/L \times L/P$, GDP per worker is in turn one of the two determinants of GDP per capita (Y/P), the other being the percentage of workers in the total population (L/P). In short, this means that technological progress (in its broader sense) leads to a rise in GDP per worker and hence GDP per capita. Thus it follows that, other things being equal, countries with higher GDP are more technologically advanced.

These conclusions do not necessarily hold in a preindustrial world where agriculture maintains a significant share of the total output. The agricultural production function includes land as a third factor of production. Furthermore, similar to the problem with mining, GDP does not compute land as a cost (again, in part as a consequence of the specific context in which it was created): in agriculture, when passing from gross saleable production to value added, a figurative sum to account for the extension of the land used to produce agricultural goods is *not* detracted, as if land was an inexhaustible resource. All of this means that a rise in GDP, either per worker or per capita, can be due not only to technological progress but also to an extension of the land cultivated. In turn, this implies that in the preindustrial world, we can have countries with high GDP – or with high standards of living – that are not technologically advanced. They may be rich simply thanks to a favorable relation between land and population (because they have high land per capita), but that land can be inefficiently used:

they would have low *per hectare* GDP (land productivity), but since they may rely upon a lot of land, relatively high *per worker* (and thus per capita) GDP. Obviously, in this situation, the standard coefficients of the Cobb-Douglas function do not hold. In addition, the assumption of perfect competition may be incorrect, at the very least because a significant proportion of preindustrial societies are not even market economies. These considerations make the use of GDP for eras and contexts radically different from ours, namely, for those preceding the Industrial Revolution, particularly problematic. At the very least, the interpretation we give to those GDP figures should be more cautious and not a mere replication of the interpretative framework we have assumed for the last stretch of human history. Because of such limitations, in turn I am limiting the present study to the use of GDP in modern times.

Even so, however, things are far from simple. And here we come to the third characteristic of GDP than any cliometrician or economic historian (but also any shrewd economist) should always have in mind. As discussed, the first official statistics of national income were produced in the United States in the 1930s. They progressively spread across the world only after World War II. For the previous periods, quantitative historians or applied statisticians – or "chipprephiles," as Maddison (1994) once named himself – had to reconstruct their own historical series of GDP by making the best out of several different sources and hypotheses.[7] When they were lucky, they could benefit from data on production, prices, labor force, and wages, but these data were not always comprehensive or exhaustive and often not even available. We may draw a line roughly at the mid-nineteenth century. For earlier epochs, available sources are scant, and GDP estimates often come from a handful of figures on urbanization and demography, related assumptions on the share of nonagricultural sectors (as most of Maddison's figures for the years before 1820), plus a few reliable series on prices and wages for a limited number of countries, and maybe some information about public revenues and tax collection. We cannot help warning once again against a too relaxed use of these shaky figures. Gregory Clark (2009, p. 1156) has efficaciously defined Maddison's pre-1820 estimates "as real as the relics peddled around Europe in the Middle Ages."[8] However, a more in-depth discussion of these issues would go beyond the scope of this chapter.

For the years after the mid-nineteenth century, which also coincide with our period of concern, historical data are much more abundant and solid: they usually include production series that are complete, or nearly so, and at times also extended price series, plus reliable and highly detailed data on wages and employment in some benchmark years (those of official censuses). This is true

[7]They could, of course, take advantage of a long tradition of income and macroeconomic estimates, dating back to the seventeenth century (for an outline, see Maddison 2007, pp. 393–401).

[8]However, some improvements on this are now on the way (Bolt and van Zanden 2014).

for Europe, at least, where following the Enlightenment and Napoleonic wars in the course of the nineteenth-century modern bureaucratic states replaced *ancien régime* governments. For other parts of the world, the colonial administrations notwithstanding, unless we are willing to use indirect procedures (such as import – export charts), more often than not we must wait until the second half of the twentieth century, when we are in the realm of the official GDP statistics.

In other words, historical GDP reconstructions are the result of ad hoc efforts by individual scholars[9] who had to make the best possible use of the available incomplete sources. The available information typically changes from one country to another, and even within the same country, it changes across years and economic sectors (e.g., Prados de la Escosura 2016). As a consequence, even for modern times, (country and regional) GDP series are often the product of different methodologies and hypotheses, and this has significant bearings on the results. Cliometricians need to be aware of the methodologies (and limitations) behind the GDP series they are using. The following section is intended to offer an outline of the main methodological problems we encounter when dealing with, and working on, historical GDP estimates in modern times.

Reconstructing GDP: Methods and Problems

In order to be able to assess the soundness of GDP figures, transparency is of course a preliminary condition: sources and methods must always be adequately described, ideally up to the point that results must be replicable. This may seem obvious, but actually it is not. For example, Italy's official historical series of GDP (beginning in 1861), one of the first in the world to be produced (Istat 1957), was a pioneering effort that also came to be famed for its lack of transparency in sources and methods, which did not help remedy the faults discovered by subsequent scholars (Federico 2003; Fenoaltea 2003; Felice and Carreras 2012). The original series has finally been replaced with a new one reconstructed almost entirely by economic historians (e.g., Baffigi 2013), more than half a century after it was originally published. Every country has its issues in this regard, and it would be impossible to review them all. The good news is that the standards have changed, and now an established rule of the scientific community is that GDP estimates must be transparent and replicable, which they are, for the most part. Maddison's magnum opus (1995, 2001, 2006), which presents GDP figures spanning the past 2000 years for most countries, also accomplishes this rule: although some of his assumptions for the nineteenth century

[9]For modern times, outstanding examples are Feinstein (1972) for the United Kingdom and Prados de la Escosura (2003) for Spain.

are questionable – or may simply look too crude[10] – an outline of the procedure is always provided, with further reference to the primary and secondary sources used; new contributions from the literature are also discussed and at times properly integrated.[11] The *Maddison project* was created[12] in 2010, the same year that Maddison died. Its aim is to revise and improve Maddison's original dataset as new information becomes available. The first results have already been produced, and they incorporate a great deal of the new statistical evidence and historical estimates that had become available in the meantime (Bolt and van Zanden 2014).[13] Other scholars are at work on comparative estimates for shorter periods of time or with a sectoral focus, producing data that can usefully complement and integrate those of Maddison. For its scope and accuracy, it is worth citing Williamson's (2011) *Project on industrialization in the poor periphery*, which, after reviewing and harmonizing a number of primary and secondary sources, presents estimates of industrial output for the period 1870–1939 at constant prices for the European eastern and southern periphery (12 countries), Latin America (7 countries), Asia (7 countries), the Middle East (Egypt and the Ottoman empire), and Africa (South Africa), plus three leaders (Germany, the United Kingdom, and the United States). As these works progress, it is possible to imagine a future in which we may be able to take advantage of an international GDP dataset whose problems of reliability and comparability will have been progressively reduced and perhaps even become negligible.

However, reaching such a goal will not be an easy task, and it is only fair to acknowledge that we are still far from it: information is lacking, and research is sparse not only for minor countries but also for the most important ones whose data

[10]Just a handful of examples: for Switzerland, per capita GDP growth from 1820 to 1951 is assumed equal to average for France and Germany (Maddison 2006, p. 409); for Italy, a "guesstimate" for 1820 is created, "assuming that GDP per capita grew at the same pace from 1820–1861 as from 1861–90" (Maddison 1991, p. 234; Maddison 2006, p. 408); but for this country, see Malanima (2006, 2011), his 2011 article having been incorporated in the updated version of Maddison's database (Bolt and van Zanden 2014). For Albania, per capita GDP from 1870 to 1950 was assumed to move in the same proportion as the average for Bulgaria, Romania, Yugoslavia, Hungary (!), Czechoslovakia (!), and Poland (!); but what is more worrisome, this same average should work also for the entire Russian empire (Soviet Union territories) from 1820 to 1870 and for Greece from 1820 to 1913 (Maddison 2006, pp. 407, 469–471).

[11]See, for example, the review of Good and Ma's (1999) proxy measures of per capita GDP for six Eastern European countries (Bulgaria, Czechoslovakia, Hungary, Poland, Romania, Yugoslavia) plus Austria, which are derived by regression by using three indicators (letters posted per capita, crude birth rate, and the share of nonagricultural employment in the labor force) and are accepted by Maddison only for some countries (Bulgaria, Poland, Rumania, Yugoslavia), owing to the lack of any other information (Maddison 2006, pp. 403–404, 471–472). For a comprehensive picture of Maddison's amendments to his previous (2001) estimates, see Maddison (2006, p. 624).

[12]The project consists of a small working party of four established economic historians and a larger advisory board composed of 22 scholars from around the world. See the website of the project: http://www.ggdc.net/maddison/maddison-project/home.htm.

[13]Despite the title of the article ("Re-estimating Growth Before 1820"), updated estimates referring to the last two centuries also are included.

surely look more robust. Moreover, even when we have *reliable* estimates, it is not assured that these are *comparable* between countries.

Indeed, comparability probably looms as the biggest challenge. At least three problems need to be recognized: one is about quantities, while the other two are about prices. However, at this point, before entering into further details, it may be useful to provide an outline of how GDP series are normally produced. As a general rule, since price data are not usually available throughout the period, but only for some reference years, GDP series are estimated at constant prices: a base year is taken (for which there are current-price GDP estimates) and that *current-price benchmark* becomes the year of the *constant-price series*. In order to do so, for each i sector and t year, it is assumed that

$$\mathrm{GDP}^{\mathrm{ti}}/\mathrm{Q}^{\mathrm{ti}} = \mathrm{GDP}^{(\mathrm{t+1})\mathrm{i}}/\mathrm{Q}^{(\mathrm{t+1})\mathrm{i}}, \tag{3}$$

where Q is the elementary physical series. In other words, it is assumed that for each elementary series, the relation between GDP and quantity, that is, unitary GDP, does not change throughout the years of the series, with respect to the unitary GDP of the baseline year. From Eq. 3, we obtain the formula used to produce constant (base year)-price estimates as

$$\mathrm{GDP}^{\mathrm{ti}} = \left(\mathrm{Q}^{\mathrm{ti}}/\mathrm{Q}^{\mathrm{yi}}\right) \times \mathrm{GDP}^{\mathrm{yi}}, \tag{4}$$

where y stays for the baseline year.

From this formula, the problem with quantities is almost self-evident. Ideally, the elementary physical series of each country must be taken at a similar level of decomposition. This in turn should be as high as possible, because within each country, the physical series should be homogeneous. For instance, we should not estimate textiles via the total amount of textiles produced; rather, we must include separately at least each major fiber (silk, cotton, wool, linen), and, indeed, even within a major fiber, at least the main production processes (spinning, weaving) should be broken down. On the basis of textiles then, one could argue that for each country, it suffices to use the official series (of production and trade), which would then produce the finest comparable aggregate national series. But what about other sectors, such as mechanics, a sector with a non-negligible and growing impact on total GDP? In the long run, productions have changed enormously; even within a single subsector and a single production (e.g., automobiles), there are different types whose prices significantly vary from one model to another. And even the models could change: some disappear and we find them replaced by others, both backward and forward in time. As a consequence, in practice, for each country we must rely on a different methodology in order to produce the elementary physical series: not only the level of decomposition varies, but we also often resort to different proxies within the same sector or the same series (say, raw cotton instead of yarn cotton) with different hypotheses to cover the unknown productions (say, different elasticities between the other textiles, or the rest of cotton, and the chosen proxy). Even within

each country, there may be problems of comparability between different periods. For example, a remarkable degree of decomposition has been reached for the Italian industry in the liberal age, for which about 200 elementary series have been produced (Fenoaltea 2003). But it was not possible to maintain the same level of decomposition in the interwar years, when "only" 90 industrial series could be produced (Felice and Carreras 2012). Moreover, how can Italy be compared with other countries for which only the major industrial sectors can be estimated?

Procedures also vary because there is no common rule to firmly guide us. One rule could be "disaggregate as much as you can," but this inevitably results in many country-specific procedures, following differences in the systems of national statistics as well as the accidental availability of supplementary sources. Alternatively, it could be argued that if our goal is to compare the performances of countries, we should shift from the rule of disaggregating (which comes from a very national-centered estimating approach) to a "lowest common denominator" approach that would work for the highest number of countries. For example, we could decompose industry into a few major sectors, each one estimated through its aggregate total production (in quantities, say, tons, weighted with prices) or its most important product. However, not even this would solve the problem, simply because the most representative productions would also vary from one country to another, with possible distortions. To sum up, we must resign ourselves to the fact that having perfect cross-country comparability in elementary series which span long periods of time is all but a chimera. Once this limitation is accepted, we can look with more indulgence at the current state of the elementary series used to produce the available historical GDP estimates, that is, a disparate collection of what has been done in different countries during recent decades, by separate scholars concentrating on their own sources and problems, unworried by the need for a common aggregating methodology.

When dealing with constant-price series, however, comparability in prices may even be a more serious issue. In the choice of elementary price data, we encounter more or less the same problems briefly discussed above for physical quantities (although these are usually limited by the use of a few benchmarks instead of long series). However, there is also a further significant distortion due to the way in which relative prices vary over time. From Eq. 4, in fact, it is true that for $i = 1 \ldots n$ productions, total GDP (GDP^N) is

$$GDP^{tN} = \sum_{i=1}^{n} \left(\frac{Q^{ti}}{Q^{yi}} \right) \times GDP^{yi} \tag{5}$$

In Eq. 5, we can see that to each physical series, a GDP weight has been assigned, which is constant over time and corresponds to the GDP weight of that single production in the base year: this depends on the unitary GDP and the quantity produced, again in the base year. As Fenoaltea (2010, p. 91) efficaciously pointed out, such a bold assumption "is done … with a bad conscience but with good

precedent: all sorts of scholars, similarly constrained, have done the same." In short, Eq. 5 is a Laspeyres quantity index number, which uses the GDP weights of a base (fixed) year to convert the component quantities to comparable values and, at the same time, to weight them. Actually, most of the available GDP series are Laspeyres quantity indices.[14] Of course, unitary GDP is the result of the price system in use that year, i.e., the relative price of that single production compared with the others, at a specific point in time. The problem is that relative prices (and thus unitary GDPs) do not remain constant over time. It is well known that prices and quantities are usually negatively correlated, on the demand as well as on the supply side, especially in the presence of technological progress, which reduces the unitary costs of production. Over the course of decades, in fact, some sectors and productions (e.g., chemicals and mechanics in the West between the late nineteenth and the twentieth century) grow faster than others thanks to technological progress. As a consequence, the early-weight price series, those based on a price system early in time (say, 1870 in an 1870–1913 GDP series), assign a higher weight to the sectors growing faster (whose quantities increase and relative prices decrease), and therefore, they grow more rapidly in the long run. For the same reason, the late-weight indices (say, a 1913-price series) grow less. This has become known as the "Gerschenkron effect," since it was reasoned by Alexander Gerschenkron (1947), soon after World War II, when analyzing Soviet indices of industrial production. Today, it is also simply known as the "index number problem" (Feinstein and Thomas 2002, p. 513).

Of course, the Soviet Union in the interwar years was an extreme case of accelerated growth in heavy industrial sectors, and thus the distortion caused by the "Gerschenkron effect" was fundamental. However, it is worth stressing that the index number problem is also serious in countries that modernized at a slower pace. For example, Italy from 1911 to 1951 ranked more or less in the middle among OECD countries.[15] For Italy, three indices of industrial production at three different price bases are now available, all made up of the same elementary physical series (only the relative weights in the unitary GDP, which are 1911, 1938, or 1951, change). From 1911 to 1951, the 1911-price index of industrial value added more than triples from 100 to 362; the 1938-price index goes from 100 to 264; and the 1951-price index doubles from 100 to 210 (Felice and Carreras 2012, p. 447). It is clear that such a major distortion cannot be ignored when it comes to international comparisons. If large differences are observable in the *same* series (i.e., series constructed with the same methodology and proxies), which differ only in their benchmark years, then when it comes to comparing *different* series belonging to different countries, a minimum requirement is that their base years be the same or at least relatively close.

[14]For a detailed discussion of Laspeyres indices and their properties as well as of the other main indices used in time series (see Feinstein and Thomas (2002, pp. 507–525)).

[15]For updated international comparisons of Italy's GDP with the rest of the world, in 10-year intervals from the unification of the country (1861) until 2011, see Felice and Vecchi (2013, p. 28).

Nevertheless, this is barely the case. Actually, Maddison's GDP estimates put together a large collection of different price bases, following once again the national accounting systems of every country and the work of separate scholars. Even a brief examination of the price bases that Maddison reports to have used in order to produce constant price GDP series offers a discomforting picture: Austria, 1913 (for the 1820–1913 series) and 1937 (1913–1950); Belgium, 1913 (1913–1950); Denmark, 1929 (1820–1947); France, 1870 (1820–1870); Portugal, 1910 (1851–1910); Switzerland, 1913 (1913–1950); Australia, 1910/1911 (1861–1938/39); the United States, 1929 (1890–1929) and 1987 (1929–1950); and the Soviet Union, 1913 (1870–1928) and 1937 (1928–1950). And this is an incomplete list (Maddison 2006, pp. 403–409, 450–457, 471).[16] This means, for instance, that for 1913–1951, Switzerland is barely comparable with Austria, and the same is true for Belgium in comparison with Denmark, for the Soviet Union in comparison with the United States, and so on.

It is worth noting that the "Gerschenkron effect" produces a distortion not only for what concerns international comparisons but also in terms of intra-sectoral comparisons within the same country: the GDP sectoral shares of a series at constant prices tend to remain very close to those of the base year, for obvious reasons (only quantities vary). Both these distortions (between- and within-country) would not be present if we were able to estimate GDP at current prices for each year of the series – as is done today. In order to have "real" GDP figures, current-price GDP series could then be deflated by using a single common deflator instead of sector-specific deflators as in Eq. 5: wages, for instance (Fenoaltea 1976). In this way, we would have constant-price series, unbiased toward the GDP composition of the baseline year and comparable between countries. However, such a procedure is too data demanding, and in the end, it may also turn out to be a chimera, not least because the choice of the deflation system is far from undisputed (e.g., wages would ignore the share of GDP going to capital gains, while a consumer price index would ignore the price of investment goods). What can be reasonably done is to estimate as many current-price benchmark years as possible for every country. From these, short constant-price series can be created. Finally, a long-run constant-price series can be produced by connecting the shorter series through chain indices: ideally, a chain index rebased every year (a Divisia index) could be created. Alternatively, a Fisher Ideal index can be produced: the early-year and late-year indices can be combined through a geometric average, with weights inversely proportional to the distance between the year of the series and the price basis, according to the formula

$$y_{i_{ini}\,\text{min prices}}^{\frac{i\,\max - i}{i\,\max - i\,\min}} \times y_{i_{ini}\,\text{max prices}}^{\frac{i - i\,\min}{i\,\max - i\,\min}} \tag{6}$$

[16]For further details and more countries, reference must be made to the previous version of Maddison's work (1995, pp. 126–139) and to the country-specific sources cited by the author.

where i is the year of the series y, imin is the early benchmark, and imax is the late one.[17]

The third problem when comparing international GDP series comes with purchasing power parities (PPPs), which is not at all a minor issue (indeed, it is probably more easily recognizable than the Gerschenkron effect). With the ambitious goal of comparing not only income and production but also the standard of living, Maddison converted all his country estimates into Geary-Khamis PPP 1990 international dollars. It goes without saying that any purchasing power converter is different from the official exchange rate, since it allows for differences in the cost of living. The procedure is simple: (a) each national GDP series, expressed in constant prices at its own national currency, is converted into an index; (b) at the same time, for the baseline year 1990, each national GDP, expressed in its own national currency and at current prices, is converted into 1990 international dollars by using Geary-Khamis PPP deflators[18]; and (c) with the index in (a), a new national series in Geary-Khamis PPP 1990 international dollars is then created. By using this method, all series can be converted into a comparable unit of measurement without changing the growth rate of each national series. In order to estimate PPP converters, different multilateral measures (and methods) can be used, but it must be acknowledged that Geary-Khamis is a suitable one because it assigns each country a weight corresponding to the size of its GDP and considers the United States, the most important economy, as the *numeraire* country (i.e., the 1990 Geary-Khamis dollar has the same PPP as the US dollar has in the United States in 1990).[19] However, of course, both the country weights and the purchasing power differences are those measured in 1990. In fact, Maddison's entire magnificent edifice is based upon the situation recorded in 1990, as if the relative purchasing power of currencies (both domestic and international) was fixed, rather than changing over time, especially in the long run, as both the underlying forces (namely, the domestic and international flows of goods and services) that govern the movement of prices and the basket of goods and services used to construct the PPP converter change. This problem becomes more serious if

[17]For an application of Divisia and Fisher Ideal indices, see Crafts (1985) for England, Prados de la Escosura (2003) for Spain (Fisher Ideal index), and Felice and Carreras (2012) for Italy (Fisher Ideal index). In Prados de la Escosura (2003, pp. 46–47), an application of the Paasche index can also be found: the Paasche index (which uses a changing set of prices to value the quantities) is used to produce price series, which are then combined with the Laspeyres quantity index to estimate GDP at current prices.

[18]The Geary-Khamis purchasing power converters for most countries can be found in Maddison (2006, pp. 189 (OECD countries), 190 (five East European countries and USSR), 199 (Latin America), 219–220 (Asia), 228 (Africa)). The reference year was always 1990 only for OECD, East European countries, USSR, Japan, and China; for the others, it varies from 1975 to 1993.

[19]Other multilateral measures either give all countries the same weight (such the EKS system used by Eurostat for political reasons), are a shortcut approach based on reduced information (such as ESCWA used for 8 West Asian countries), or employ as a numeraire a currency different from the US dollar (such the ESCAP measure used for 14 East Asian countries, which takes as a reference the Hong Kong dollar) (see Maddison 2006, p. 172).

we go further backward in our extrapolation, thus distancing ourselves from the baseline year. As Prados de la Escosura (2007, p. 18) put it:

As growth occurs over time, the composition of output, consumption, and relative prices all vary, and the economic meaning of comparing real product per head based upon remote PPPs becomes entirely questionable. Hence, using a single PPP benchmark for long-run comparisons implies the hardly realistic assumption that no changes in relative prices (and hence, no technological change) takes place over time.

Even over a period of four decades, the distortions from the use of a baseline benchmark distant in time are large, "above 5% and often much higher, while showing a high dispersion" (Prados de la Escosura 2000, p. 4). For these reasons, the use of a number of PPP converters at different points in time, following at least the main historical ages, would be preferable; but constructing PPP converters is a highly demanding task in terms of time and resources (Ahmad 1988) and one undermined in terms of feasibility (and reliability) by data scarcity for the period before World War II. Given the lack of reliable PPP converters for distant periods, Maddison's approach, which was actually pioneered by Bairoch (1976), still represents a viable, if suboptimal alternative. It has been argued, for instance, that the distortion caused by comparing real products on the basis of long-run PPP projections can be larger than that generated by using current nominal exchange rates (Eichengreen 1986); thus, even simple exchange rates could turn out to be a more practical shortcut.

And yet there is indeed a superior shortcut, which is based on the reasonable assumption that price levels between a country and the rest of the world move according to some basic economic characteristics (e.g., the share of international trade, income, or population). By further developing the method originally envisaged by Kravis et al. (1978), Prados de la Escosura (2000) tested a number of variables against the 8 available PPP benchmarks (spanning 1950–1990) for 23 countries, through panel regressions. As a result, he proposed a structural relationship for each country between its price level (defined as the ratio between PPP and exchange rates), on the one hand (y, dependent variable), and its nominal GDP per capita plus an additional set of explanatory variables (ratio of commodity exports and imports to GDP, population, area, a periphery dummy indicating if the country's nominal income represents half or less the US income), on the other ($x_{1,\ 2,\ 3,\ 4}$ and dummy independent variables).[20] By applying the estimated parameters to the independent variables recorded in past times for the same countries (when available) as a second step, Prados de la Escosura could calculate additional PPP benchmarks, spanning 1820–1938 (and for some countries, previously uncovered, up to 1990), and then propose comparisons of real per capita GDP at current historical PPPs. The author is aware of the limitations of his method that "even for the same group of countries" is based on "the application of a structural relationship derived from advanced western

[20]In that article, an excellent discussion of the literature about these issues and the different shortcut methods is also provided (pp. 2–8).

economies over the past 50 years to earlier and different historical contexts."[21] Nevertheless, the potential error is minor compared with that residing in Maddison's approach.[22] The latter retropolates a PPP without any adjustment; in Prados de la Escosura, we still have retropolation, but with adjustments for changes in the underlying economic structure based on an empirically tested relationship. Thus far, the results from Prados de la Escosura's method are available only for a limited number of countries. This may be the main reason why Maddison's data continue to be so widely used, even in papers published in top economics journals: they are the only available long-run GDP series for many countries (or, in any case, those more easy to pick up), their patent unreliability notwithstanding. Bad conscience, but good precedent. To clean our conscience or to make it feel even more guilty, it is fair to warn against this habit.

Convergence or Divergence? Measures and Models

Provided we have relatively sound estimates, we may then investigate the patterns of GDP growth in modern times. Did the country converge over time? A number of techniques are available to measure convergence, and different underlying theories are available to interpret the results. Techniques based on time series allow us to detect differences in cycles in trends and to identify country-specific break points. Unfortunately, they are more data demanding: any user should always check for the fact that the series at hand is not the result of some extrapolation or interpolation, as is often the case with historical estimates. Cross-sectional analyses allow us to test convergence when only a few benchmarks are estimated (possibly, each benchmark at its current prices) and therefore may result in more appealing long-run comparisons. Of course, they only consider the trend and for this can miss relevant information in between the two benchmarks.

Two concepts of convergence (Barro and Sala-i-Martin 1991) are generally accepted and used mostly – especially the second one – with benchmark data. σ-Convergence is a measure of dispersion in per capita GDP between different countries. The σ prefix comes from the standard deviation, which is used to quantify it. A simple test of σ-convergence is provided in Fig. 1. This figure displays the standard deviation of the logarithm of real per capita GDP for 20 countries, in selected benchmarks, from 1880 to 1990; the benchmarks are those for which the

[21] Prados de la Escosura (2000), p. 19

[22] As confirmed by the results. Just a couple of examples: in 1860, according to Maddison, Greece would have a per capita GDP higher than France (0.855 vs. 0.850), while according to Prados de la Escosura, France had a much higher GDP per capita (0.821 vs. 0.405) in 1860, 1870, and 1880. According to Maddison, Austria (at pre-World War I borders) would be above France, Germany, and Canada, while according to Prados de la Escosura, and much more plausibly, it would be below them (2000, pp. 24–25).

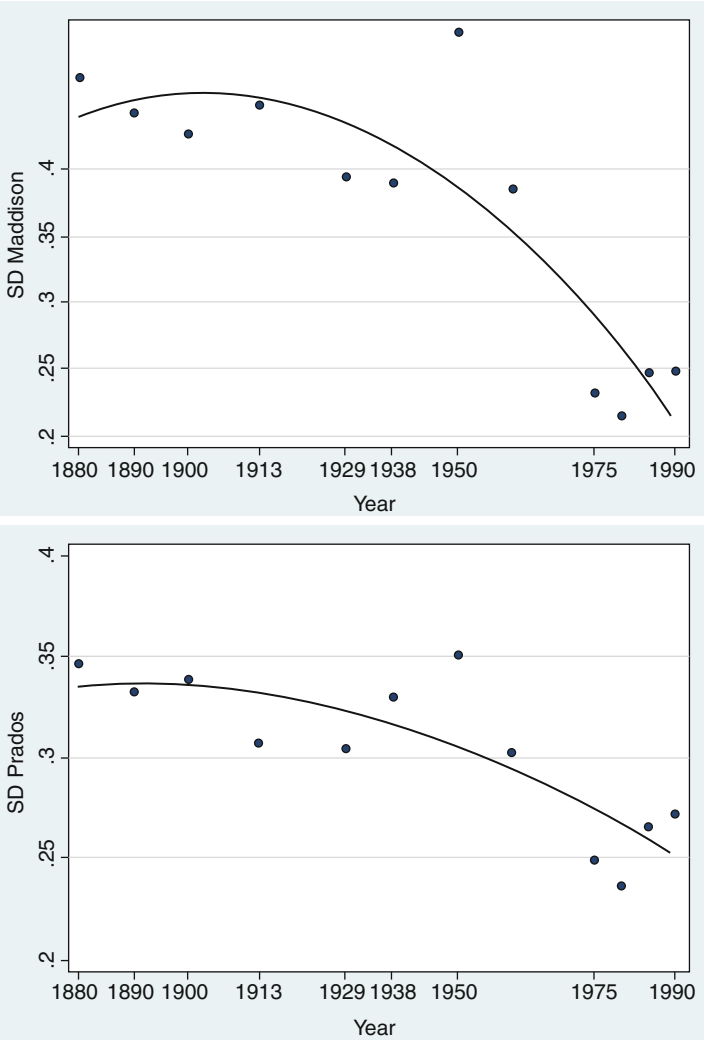


Fig. 1 σ-Convergence in GDP per capita from 1880 to 1990, according to different GDP estimates. (Sources and notes: see text)

estimates by both Maddison (upper quadrant) and Prados de la Escosura (lower quadrant) are available for an unchanged minimum number of 20 countries.[23]

[23]The countries are Argentina, Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Italy, Japan, the Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, the United Kingdom, and the United States. The benchmarks are 1880, 1890, 1900, 1913, 1929, 1939, 1950, 1960, 1975, 1980, 1985, and 1990. GDP per capita is expressed in 1990 international dollars, but in the case of Prados de la Escosura, the figures are rescaled with his current price PPPs (2000, pp. 24–31).

As can be seen, the results can differ significantly. For the same countries, σ-convergence is much stronger using Maddison's estimates than it is when using Prados de la Escosura's estimates. This should not come as a surprise, given that differences in nominal GDPs and differences in PPPs are usually positively correlated. Both authors record convergence in GDP per capita, and this means that differences in this variable are lower in later periods; for this very reason, Maddison's differences in PPPs (which are for 1990), when retropolated, may also be lower than the real (historical) ones (say, for the nineteenth century). The latter are those estimated and employed by Prados de la Escosura, who then makes use of higher differences in PPPs in the early periods. This means that in early periods, the cost of living was lower in poorer countries than that supposed by Maddison, and therefore poorer countries had at that time higher real GDP; as a consequence, they converge less.[24] However, it is also worth noting that both authors report similar trends: most of the convergence took place from 1950 to 1975, but then it came to a halt and even reversed.

The fact that poorer countries grow faster than richer ones is usually regarded as a precondition for a decrease in dispersion. Technically, this is known as β-convergence, which can be conditional or unconditional. The prefix in this case derives from the coefficient of the regression model used to measure it (Eq. 8). β-Convergence can be tested by regressing the growth rate of per capita income with its initial level; if there is a negative correlation, then countries with higher per capita GDP are growing less. It is worth noting, however, that when we have unconditional (or absolute) β-convergence, we may not necessarily also have σ-convergence. For example, the initial GDP of a country may pass from 0.6 to 1.6 (the average being 1), implying β-convergence but also an increase in dispersion (σ-divergence). The opposite, however, is not true, namely, if we have σ-convergence, we always have β-convergence. If a country goes from 1.6 to 0.6, we record both σ- and β-convergence. For the same countries as in Fig. 1, β-convergence is tested in Fig. 2, where the growth rate from 1880 to 1990 in real per capita GDP is regressed on the logarithm of initial income.

As expected, β-convergence is stronger in Maddison's than in Prados de la Escosura's estimates: in the former, R^2 is considerably higher, and, as a consequence, the standardized β-coefficient is also more elevated (−0.869 vs. −0.738). Figure 2 also provides information about the relative performances of individual countries, namely, which grew above average, *given their initial income*, and which grew below; the former position themselves above the fit line, whereas the latter are below. For example, in both cases, Argentina records a disappointing performance, while Japan is the big winner. The entire European northern periphery (Sweden, Finland, Norway, Denmark) has been growing above the average, and today it is no longer periphery. Instead, the southern periphery (Portugal, Spain, Greece), with the exception of Italy, is below the average.

[24]It should be reminded that all are expressed in logs. In absolute terms, the standard deviation of real GDP per capita increased in both Maddison and (more) in Prados de la Escosura.

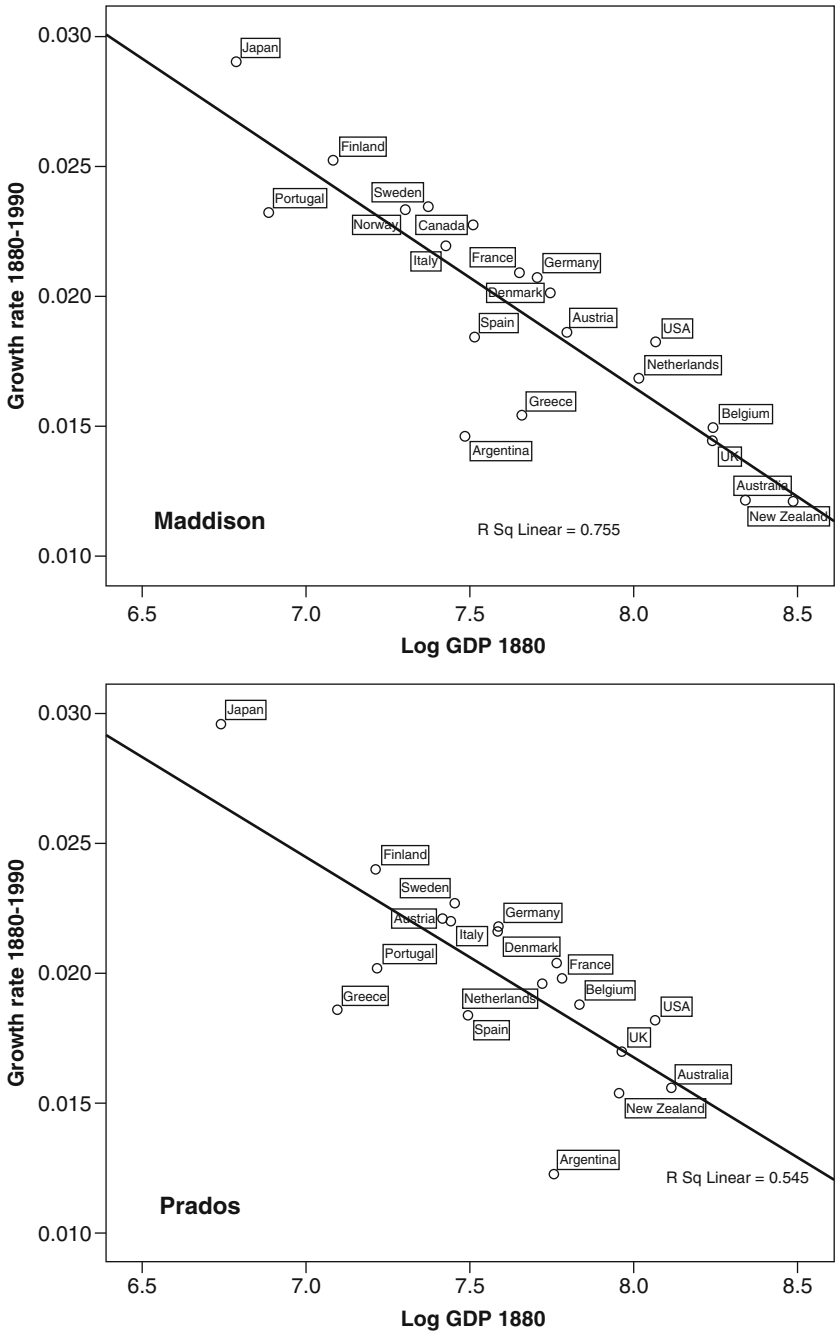


Fig. 2 β-Convergence in GDP per capita from 1880 to 1990, according to different GDP estimates. (Sources and notes: see text)

Why do some countries converge more than others? Economic theory is replete with elaborate models to explain the observed patterns. In the space of a few pages, it is impossible to review all of them thoroughly, but we may provide a sketch of the most important (and popular) ones. β-Convergence, both conditional and unconditional, can easily be incorporated in the neoclassical approach. This is based on the assumption of diminishing returns to capital or, in other terms, the downward slope of the savings curve. According to Solow (1956) and Swan (1956), in a closed economy where savings are equal to gross investments, the growth rate of capital stock would be

$$\gamma_k = s^* Af(k)/k - (\delta + n) \tag{7}$$

where s is the constant savings rate, ranging from 0 to 1, k is the capital stock per person, Af(k) is the production function in per capita terms, δ is the depreciation rate of the capital stock, and n is the exogenous rate of population growth. Thus, δ + n is the depreciation curve, a horizontal line, and s*Af(k)/k is the savings curve, a downward-sloping line. The argument for convergence holds that, given diminishing returns to capital, each addition to the capital stock generates higher returns when the capital stock is small. Of course, the capital stock determines per capita GDP, or income, via productivity. Thus, output and income should grow faster in countries or regions with smaller capital, i.e., with smaller income. It is worth stressing, however, that in order to satisfy this condition, the neoclassical model needs many collateral qualifications: the most important ones are that all economies must have a similar technology (considered in a broader sense to include taxation, property rights, and other institutional factors) as well as similar savings and population growth rates. These assumptions are anything but realistic in long-run cross-country comparisons. This is not a problem in itself, provided we always remind ourselves to use the models as they should be used: not as something true or false to be verified, in order to corroborate a theory, but as an analytical instrument useful to describe facts in a simplified way. In other words, we must always remind ourselves that theories are confirmed by facts, not by models, and that models rather serve us to draw the contours of the most relevant facts.

Using a Cobb-Douglas form of the production function, following Barro (1991), cross-country growth regressions may be expressed as

$$\gamma_i = \beta \log y_{i,0} + \psi X_i + \pi Z_i + \varepsilon_i \tag{8}$$

where γ_i is the growth rate of an i country, $y_{i,0}$ is its initial level of per capita GDP, X_i represents other growth determinants suggested by the Solow model apart from the initial level of income, and πZ_i represents those determinants not accounted for by the Solow model.

We have unconditional β-convergence (as seen in Fig. 2) when

$$\gamma_i = \beta \log y_{i,0} + \varepsilon_i \tag{9}$$

with the negative sign of the coefficient β.

Otherwise, we do not have *unconditional* convergence. We can still have *conditional* convergence, however, if after adding other variables to Eq. 8, the β coefficient becomes negative (Barro and Sala-i-Martin 1992). The basic idea behind conditional convergence is that differences in per capita incomes are not permanent only because of cross-country structural heterogeneity, that is, because the model does not satisfy collateral qualifications. This can be due to different resource endowments, institutions, and migration rates, as well as to human and social capital disparities, among others things. In the growth regressions, each of these factors can be a conditioning variable, coming either from within the Solow model variable group X i (i.e., human capital, institutions or social capital, if we consider technology in its broadest sense) or from outside the Solow model from the Z_i variable group (think of climate, but usually variables of this kind are much less common in the literature, while spanning an impressive range of categories). Once we have checked for the effects of structural heterogeneity, there can still be convergence; however, this is not convergence to a single common steady state, but rather the convergence of every country to its own steady state, given its own conditioning variables (i.e., *conditional* convergence). It has been called *convergence*, but truly this model does not measure convergence across regions or countries, since different regions or countries may have different steady states.

A major problem with this framework is the multiplicity of possible regressors, given that the conditioning variables that can be run are practically countless. Durlauf et al. (2005) classified about 150 independent variables used in growth regressions (in almost 300 articles) plus about 100 instrumental variables. In short, the number of possible regressors exceeds the number of cases, thus "rendering the all-inclusive regression computationally impossible" (Sala-i-Martin et al. 2004, p. 814). One reason for the multiplicity problem may lie in the analytical and theoretical weakness of the Cobb-Douglas function, which is valid only in the presence of a vast number of assumptions and has been verified only in a limited number of cases (namely, for the United States in the interwar years). There are two approaches to cope with the multiplicity problem: one is to take advantage of information from qualitative and case study research, while the alternative is to resort to econometrics in order to automatically sort out the irrelevant regressors. Bayesian models, which attach probabilities to each regressor, are an answer to the multiplicity problem safely within the second approach. Among these, the Bayesian Averaging of Classical Estimates (BACE) model, which makes use of the classical ordinary least squares (OLS) estimation, may be the most appealing technique. However, results from BACE models are far from convincing. To date, probably the most comprehensive exercise has been carried out by Sala-i-Martin et al. (2004), who proposed a BACE approach in order to sort 67 explanatory variables in cross-country regressions. Some of their findings look reasonable: for instance, they found primary school enrolment to be the second most important explanatory variable for 1960–1990 GDP growth rates. However, others don't. According to their model, the most significant explanatory variable was the dummy for East Asian countries. This outcome can be accepted only if we recognize that these regressions indicate a simple correlation; but if we are in search of an explanation (i.e., causation), what the model tells us is that South Korea grew because. . .it was South Korea. And there

are more problems with the results from that BACE model. For example, the socialist dummy is not correlated with (negative) growth. While the authors apparently do not note the tautology about the East Asian dummy, in the case of the socialist dummy, they discuss the unpersuasive result and specify that it "could be due to the fact that other variables, capturing political or economic instability such as the relative price of investment goods, real exchange rate distortions, the number of years an economy has been open, and life expectancy or regional dummies, capture most of the effect" (Sala-i-Martin et al. 2004, p. 829). Nevertheless, the ultimate determinant of most of these variables, as well as the particular political and economic features of those countries, was the socialist regime and the correlated planned economy: an econometric model concealing this evidence may lead to distorted interpretations of both history and the determinants of economic growth. These examples have been made to illustrate that the first approach must not be overlooked and, indeed, is often preferable. Historical knowledge, sensitivity to case studies, and country-specific characteristics should serve as a compass in order to choose among conditioning variables as well as be seen as an indispensable complement to any econometric analysis.

There is still the possibility that countries do not converge because initial conditions determine different outcomes in the long run, that is, the hypothesis that there are no decreasing returns to capital, for example, because the production function is not of a Cobb-Douglas form. A simple linear technology AK, instead of the neoclassical technology Af(k), would transform Eq. 7 into

$$\gamma_k = s^* A - (\delta + n) \tag{10}$$

where the savings curve is no longer downward sloping, but a horizontal line, just like the depreciation curve. Thus, two economies with different initial capital stocks would not converge even with all other conditions being equal. If technology or other parameters differ as well, these economies could still converge, but indeed they could also further diverge. They would converge if A or s are systematically higher in the poorer economy, if the depreciation line is systematically lower, or if other determinants of growth not included in the model are systematically higher as well; however, to quote Sala-i-Martin (1996, p. 1344), "there is no a priori reason why this should be the case." On the contrary, there is evidence that the savings curve is not even horizontal, but upward sloping. For example, because of economies of scale, increasing returns to capital have frequently been called into question to account for the rise in the United States during the second half of the nineteenth century or the rise of China in recent decades.

With the hypothesis of increasing returns to capital, we have entered the field of cumulative approaches. Following Myrdal (1957), this approach claims that growth is a spatially cumulative process that requires a minimum threshold of resources in order to start and thus may indeed increase cross-country disparities. Different schools refer to cumulative approaches. Among those worth mentioning are endogenous growth models (Romer 1986) that can still be regarded as a derivation from

the neoclassical approach and link economic growth to levels of human capital. Also important is new economic geography (NEG) (Krugman 1991), where the key determinants are either the economies of agglomeration (divergence) or the costs of congestion (convergence), and thus the size of the market plays a central role.

In practice, it is not easy to distinguish the increasing returns of endogenous growth models from the lack of collateral conditions of traditional (exogenous) neoclassical models. When there is no convergence, it may be difficult to conclude whether the traditional neoclassical model can still be valid with some qualifications to be satisfied or, on the contrary, that cumulative endogenous growth should be regarded as more suitable. Moreover, in historical analyses, crucial data, such as estimates of capital, are often lacking or unreliable. Furthermore, the models of increasing returns can easily be extended to predict convergence, such as in Eq. 10 by endogenizing the savings rate on the assumption that it would decrease with higher levels of capital (Sala-i-Martin 1996). Such a hypothesis is not at all implausible: think again of the opposite cases of China and the United States (the latter with higher capital but a lower savings rate). Thus, a unified long-term production function based on increasing returns could still be plausible in the case of convergence. On the other hand, some conditioning variables, such as the stocks of human (or even social) capital, can be seen alternatively as initial conditions in exogenous growth models, such as by decomposing K into physical and human capital (Mankiw et al. 1992).

Attempting to distinguish between the NEG approach, on the one hand, and the two neoclassical approaches, on the other, might be a more fruitful approach. Indeed, in terms of implications, it may be even more appealing. Broadly speaking, NEG models focus on the demand side. The resulting divergence in per capita GDP should be due to differences in "within-sector" productivity, brought about by economies of scale. The other two models, both the exogenous and the endogenous growth versions, are instead based on the supply side, namely, on imbalances in factor endowment. Divergence should refer to the "industry mix" effect, i.e., differences in the allocation of the working force between economic sectors. A simple algebraic calculation that decomposes GDP per capita into the product between GDP per worker (productivity) and workers per capita (employment rate), and then in turn decomposes the growth of productivity into "within-sector" productivity and the "industry mix" effects, may provide us with an (approximate) answer. This is also appealing in terms of interpretation given that, arguably, NEG growth can be explained by forces beyond human control (position, population density, and infrastructures that impact transportation costs, though they are at least in part the result of human decisions) in a larger portion than exogenous or endogenous neoclassical growth, which would typically include human capital, social capital or culture, and institutions as conditioning variables. Caution is warranted once again, since a significant proportion of the results may depend not only on the reliability of the estimates but also on the level of sectoral decomposition: within-sector productivity differences may be present between single industrial productions, but concealed at the aggregate level.

Besides these empirical difficulties, all three approaches seem to have theoretical limitations. Although the shift from divergence to convergence is usually allowed for and widely accepted, the economics literature has mostly neglected the possibility of a further reversal of fortune, namely, convergence to be followed by divergence again. This is common to all three models: there may be divergence at the beginning, because of either conditional exogenous variables, endogenous differences in factor endowments, or economies of scale, but then at a certain point convergence begins. Because differences in conditioning variables have been removed, factor endowments have converged, or congestion costs have exceeded economies of scale. Once progress is at work, it should go on until convergence is achieved. A renowned paper by Robert Lucas (2000) may be taken as paradigmatic of this frame of mind. Lucas argued that sooner or later a country will start industrial development and then converge. The problem is only to establish when, not if. However, once a region has embarked on economic growth, the process of convergence (in the long run) cannot be reversed. Nevertheless, how tenable is this argument? Many examples suggest that convergence may be stopped or even overturned. Take the cases of Western Europe and Japan (toward the United States) in the past two decades or of Southern Europe (toward Northern Europe) in recent years. The reason for the inadequacy of theoretical models can be explained by the fact that they are all static, in the broadest sense of the word: they are all based upon a single production function, which is supposed to be valid throughout the period of analysis. However, in reality – and especially over the long term – the shape of the production function may modify, following, for instance, technological progress. Think of human capital. Primary education was surely a fundamental ingredient of growth in the initial phases of the Industrial Revolution, while higher education may have made the difference in later phases. Other conditioning variables may change, too: natural resources may still have been important in the first Industrial Revolution, as testified by the geographical distribution of industries in nineteenth-century Europe. However, social capital has probably become more important in the current post-Fordist age, as far as it helps reduce transaction costs among a multiplicity of small firms. Some institutions (namely, authoritarian ones) may be effective in promoting growth at their early stages, but not at more advanced ones. Generally speaking, dynamic economics seem to be reconcilable with history better than static ones, since history too is essentially dynamic. But there is little or no use of dynamic models in the long-run analyses of GDP convergence.

A Further Step: From National to Regional Estimates (and Models)

In recent times, the reconstruction of GDP has been extended from the nation state to its regions and provinces. For these cases, the same caveats illustrated for national accounts apply, while methodological problems (and differences) are often even more serious due to the lack of data at the subnational level. A common methodological framework has been proposed and applied to produce comparable regional GDP figures for Europe (Rosés and Wolf 2014). The method elaborates on an idea

originally put forward by Geary and Stark (2002): at a sectoral, and hopefully sub-sectoral, level, national GDP is allocated by regional employment; the preliminary results are then corrected through regional nominal wages, which should approximate differences in per worker productivity; then, to have real GDP estimates, nominal figures should finally be rescaled by differences in the cost of living. Such a procedure is based on the assumption that capital gains are distributed along the lines of incomes from labor, namely, that the elasticity of substitution between capital and labor is equal to one. Moreover, the method is all the more effective the higher the degree of sector decomposition. For the reasons exposed in previous sections (namely, the Gerschenkron effect) , the national GDP to be allocated should be at current, rather than constant, prices. Another issue is that detailed figures on regional employment before World War II are available only from official censuses, which are usually taken at 10-year intervals. As a consequence, the production of regional GDP series is almost impossible. And even so, for some important sectors, census data may be misleading. For example, agricultural production may significantly vary from 1 year to the next, especially at the local level, without significant changes in the official labor force. At least in the primary sector, direct estimates (which are not impossible to find, even at the subnational level)[25] should be preferred.

The analytical tools are also similar to those briefly examined in the previous section, with only a few differences. First, at the subnational level, techniques based on benchmark estimates are de facto the only ones utilizable, at least for international comparisons. Since the subnational series of GDP for periods before World War II are often a product of interpolation,[26] with some possible exceptions at the sectoral level,[27] time series econometrics should be avoided. Second, when we measure σ-convergence, it may be useful to weight the regions with their population.[28] As long as we are interested in discussing the performance of national economic policies, we may treat different countries as statistical units with the same weights (thus giving the same importance to each national policy) and, at the same time, treat regions within a country with different real weights (thus measuring the overall dispersion of income within a national polity).

[25]See Federico (2003) for Italy.

[26]See, for example, the regional series for Italy estimated by Daniele and Malanima (2007), which have been produced by interpolating through the available regional benchmarks the national cycles of agriculture, industry, and services.

[27]For instance, the industrial production of Italy in the liberal age (1861–1913) (e.g., Ciccarelli and Fenoaltea 2009, 2014). In fact, time series techniques have been applied to the Italian regional construction movements during the liberal age (Ciccarelli et al. 2010). Even in this case, however, it must be pointed out that although the regional series by Ciccarelli and Fenoaltea running from 1861 to 1913 are indeed very accurate, they are estimated at constant 1911 prices, with possible distortions in interregional comparisons for the early years.

[28]Different population-weighted standard deviation measures are available and can be used, from the Williamson (1965) to the Theil (1967) index.

In addition, we can compare regions using an extension of the intercountry comparison models, with only a few qualifications. From a neoclassical perspective, the search for convergence within nation states should be simplified by the fact that here structural heterogeneity plays a minor role, given the usually common macroeconomic and institutional context. In fact, neoclassical scholars tend to be more optimistic about regional convergence than they are about convergence at the national level. For instance, Sala-i-Martin (1996) investigated unconditional β-convergence by applying the Solow-Swan growth model in five large European countries (Germany, the United Kingdom, France, Italy, Spain), plus Canada, Japan, and the United States, mostly for the years running from 1950 to 1990,[29] and found a similar rate of convergence, around 2% per year. Of course, this may not always be the case: at times structural heterogeneity can be hard to overcome, even within nation states, as is arguably the case for Italy, whose regional rate of β-convergence in the long run (1871–2001) has been found to be lower, barely 1% (Felice 2014)[30]; on the other hand, the forces of NEG should work better within a nation state provided there are no institutional barriers. The neoclassical approach of equalization in factor endowments and the increasing returns of the NEG have been both tested and compared for the Spanish regions by decomposing historical estimates (1860–1930) of regional per capita GDP in productivity and industry mix effects. The results suggest that they somehow reinforced each other, following the model by Epifani (2005), which combines both: from 1860 to 1930, the between-sector component was predominant, but NEG forces were gaining momentum in the last stretch, once industrialization had arrived in a considerable number of regions (Rosés et al. 2010). Furthermore, it should be considered that within nation states there may be regional development policies at work. Thanks to the common institutional framework, these can be more effective than development policies carried out at the international level (for least developed countries), and they may significantly change the pace of convergence, at least in specific periods.[31]

The descriptive model proposed by Williamson (1965) can be used to illustrate the observed patterns at the regional level. This is an extension of the Kuznets model (1955) of the evolution of personal income within a nation state. As for the personal income distribution, the relationship between national income and inequality would take a functional inverted U shape and a subsequent double movement: rising in the first phase, when industrialization begins and tends to concentrate in the strongest areas, then decreasing as industrialization spreads to the rest of the country. Williamson was mainly concerned about industrialization and structural change,

[29]Data for the United States run from 1880 to 1990, those for Canada from 1961 to 1991, and those for Spain from 1955 to 1987; data for Japan start in 1955.

[30]The results from panel models, for the years 1891–2001, are even lower: 0.5% (random effects GLS regression) (Felice 2011). The growth rate of convergence increases to 2% only once fixed effects are considered, that is, when we pass from unconditional to conditional convergence (Felice 2012).

[31]For Italy, the western country where the most impressive regional policy (in terms of expenditures as a share of GDP) was carried out (see again Felice (2010)).

and therefore his model focused on the supply side, and it is more easily reconcilable with the neoclassical approach (differences in conditioning variables would prevent industrialization spreading until they are removed). However, from a NEG perspective, the pattern would be similar (with rising inequalities due to economies of scale and then decreasing inequality due to congestion costs). There is some confirmation of the Williamson inverted U shape for the United States. The estimates suggest divergence between the late nineteenth and early twentieth century, as industrialization increased in the northeast and spread mainly to the northern and central regions. In the second half of the twentieth century, the southern and western states industrialized as well and thus converged (Kim 1998). When looking at Europe, we have confirmation for the Spain case, with divergence from 1860 to 1920 and then convergence from 1920 to 1980 (Martínez-Galarraga et al. 2014). For Italy, however, the inverted U shape is observed in the Center-North, but not when the southern regions are included (Felice 2014). Moreover, in many cases, regional convergence seems to have come to a halt in recent decades. This finding suggests that the long-run evolution of regional inequality may follow an N movement (divergence, then convergence, followed again by divergence) (Amos 1988), but on this issue, both empirical investigation and theoretical models have barely begun.

Concluding Remarks

The chapter reviewed the most common methods employed to produce historical GDP estimates at the national and the regional levels and the use of GDP to compare economic performances in the long run. The first point to be highlighted is that GDP estimates, even when relatively sound and well informed, are more suitable for measuring economic performance from the Industrial Revolution onward. GDP was born in the United States in the aftermath of the 1929 crisis, within an empirically oriented environment. It was designed for industrial advanced economies and may not correctly approximate material standards of living in preindustrial societies, where most production is from agriculture (for which the amount of land is a fundamental ingredient that GDP does not consider) and a non-negligible proportion is not even exchanged in the market (and thus is not included in GDP accounting). Moreover, for preindustrial societies, we often lack the minimum information required to produce reliable national accounts.

In modern times, when making cross-country comparisons, we should always make sure that the adoption of different estimating methodologies does not significantly affect the results. By themselves, national estimates can be reliable or made so given the available information, but this is not the point. For cross-country (and cross-regional) GDP comparisons, it is crucial that three basic conditions are satisfied. First, the decomposition level of the quantity series must be relatively homogeneous from one country to another. Even more important, and less generally acknowledged, the base year of constant-price series must be relatively close. Third, when considering real GDP figures, the PPPs used to compare countries must be as close as possible to the period of concern: as a consequence, the

renowned Maddison estimates at 1990 PPP international dollars may be not reliable for years before World War II, as illustrated by a contrast with the alternative PPPs proposed by Prados de la Escosura (2000).

Convergence tests may be significantly affected by the cumulative effect of these distortions. The way estimates are constructed also impacts upon the models used to interpret and describe the results. For instance, in international and (even more so) interregional comparisons, cross-sectional techniques are preferable to time series analysis, because the former are less data demanding even though they may be less informative. Provided we have reliable estimates, decomposing GDP growth into productivity and industry mix effects may yield important clues for distinguishing between the role of factor endowments and structural heterogeneity, on the one side, and market access, on the other. However, such clues should always be handled with care, for example, by searching for confirmation in the patterns of individual countries or regions. It also needs to be stressed that given the quality of the data, convergence models based on conditioning variables as well as more statistically refined ones such as the BACE techniques can be trustworthy only up to a certain point. They should always be supplemented by sound historical information, including qualitative sources and case studies, which should also help sort among the best conditioning variables to be tested, given the multiplicity of possible predictors.

In short, cliometricians should make an effort not to rely exclusively on statistical tools when searching for the determinants of growth but to complement them with historical expertise. They should also have a broad view of the available models, from exogenous to endogenous growth to NEG (and others that may or may not combine ideas from the three we have outlined), and be flexible enough to adapt both the models and the statistical techniques to the different historical settings and to the quality of their data.

References

Ahmad S (1988) International real income comparisons with reduced information. In: Salazar-Carrillo J, Prasada Rao DS (eds) World comparisons of incomes, prices, and product. North-Holland, Amsterdam, pp 75–92

Amos OM Jr (1988) Unbalanced regional growth and regional income inequality in the latter stages of development. Reg Sci Urban Econ 18(4):549–566

Baffigi A (2013) National accounts, 1861–2011. In: Toniolo G (ed) The Oxford handbook of the Italian economy since unification. Oxford University Press, Oxford, pp 157–186

Bairoch P (1976) Europe's gross national product: 1800–1975. J Eur Econ Hist 5(2):273–340

Barro RJ (1991) Economic growth in a cross section of countries. Q J Econ 106(2):407–443

Barro RJ, Sala-i-Martin X (1991) Convergence across states and regions. Brook Pap Econ Act 1:107–182

Barro RJ, Sala-i-Martin X (1992) Convergence. J Polit Econ 100(2):223–251

Bolt J, van Zanden JL (2014) The Maddison Project: collaborative research on historical national accounts. Econ Hist Rev 67:627–651. https://doi.org/10.1111/1468-0289.12032

Boyd J, Banzhaf S (2007) What are ecosystem services? The need for standardized environmental accounting units. Ecol Econ 63(2–3):616–626

Carson CS (1975) The history of the United States national income and product accounts: the development of an analytical tool. J Income Wealth 21(2):153–181
Ciccarelli C, Fenoaltea S (2009) La produzione industriale delle regioni d'Italia, 1861–1913: una ricostruzione quantitativa. 1. Le industrie non manifatturiere. Banca d'Italia, Roma
Ciccarelli C, Fenoaltea S (2014) La produzione industriale delle regioni d'Italia, 1861–1913: una ricostruzione quantitativa. 2. Le industrie estrattivo-manifatturiere. Banca d'Italia, Roma
Ciccarelli C, Fenoaltea S, Proietti T (2010) The effects of unification: markets, policy, and cyclical convergence in Italy, 1861–1913. Cliometrica 4(3):269–292
Clark G (2009) Review essay: Angus Maddison, contours of the world economy, 1-2030 AD: essays in macro-economic history. J Econ Hist 69(4):1156–1161
Crafts NFR (1985) British economic growth during the industrial revolution. Cambridge University Press, Cambridge, UK
Crafts NFR (1997) The human development index and changes in standards of living: some historical comparisons. Eur Rev Econ Hist 1(3):299–322
Crafts NFR (2002) The human development index, 1870–1999: some revised estimates. Eur Rev Econ Hist 6(3):395–405
Daniele V, Malanima P (2007) Il prodotto delle regioni e il divario Nord-Sud in Italia (1861–2004). Riv Polit Econ 67(3–4):267–315
Douglas PH (1976) The Cobb-Douglas production function once again: its history, its testing, and some new empirical values. J Polit Econ 84(5):903–916
Durlauf SN, Johnson PA, Temple JRW (2005) Growth econometrics. In: Aghion P, Durlauf SN (eds) Handbook of economic growth, vol 1A. Elsevier, Amsterdam, pp 555–677
Eichengreen B (1986) What have we learned from historical comparisons of income and productivity? In: O'Brien P (ed) International productivity comparisons and problems of measurement, 1750–1939. 9th international economic history congress, Session B6, Bern, pp 26–35
Epifani P (2005) Heckscher-Ohlin and agglomeration. Reg Sci Urban Econ 35(6):645–657
Federico G (2003) Le nuove stime della produzione agricola italiana, 1860–1910: primi risultati e implicazioni. Riv Storia Econ 19(3):359–381
Feinstein CH (1972) National income, expenditure and output for the United Kingdom, 1855–1965. Cambridge University Press, Cambridge, UK
Feinstein CH, Thomas M (2002) Making history count. A primer in quantitative methods for historians. Cambridge University Press, Cambridge, UK
Felice E (2010) Regional development: reviewing the Italian mosaic. J Mod Ital Stud 15(1):64–80
Felice E (2011) Regional value added in Italy, 1891–2001, and the foundation of a long-term picture. Econ Hist Rev 64(3):929–950
Felice E (2012) Regional convergence in Italy (1891–2001): testing human and social capital. Cliometrica 6(3):267–306
Felice E (2014) Regional income inequality in Italy over the long-run (1871–2001). Patterns and determinants. In: Rosés JR, Wolf N (eds) Europe's regions, 1900–2010. A new quantitative history of the economic development of Europe. Routledge, New York
Felice E (2016) The misty grail: the search for a comprehensive measure of development and the reasons of GDP primacy. Dev Chang 47(5):967–994
Felice E, Carreras A (2012) When did modernization begin? Italy's industrial growth reconsidered in light of new value-added series, 1911–1951. Explor Econ Hist 49(4):443–460
Felice E, Vecchi G (2013) Italy's growth and decline, 1861–2011. CEIS Tor Vergata. Res Pap Ser 11(13):293
Fenoaltea S (1976) Real value added and the measurement of industrial production. Ann Econ Soc Meas 5(1):113–139
Fenoaltea S (2003) Notes on the rate of industrial growth in Italy, 1861–1913. J Econ Hist 63(3):695–735
Fenoaltea S (2008) A proposito del PIL. Italianieuropei 8(1):165–169
Fenoaltea S (2010) The reconstruction of historical national accounts: the case of Italy. PSL Q Rev 63(252):77–96

Ferreira S, Hamilton K, Vincent J (2008) Comprehensive wealth and future consumption: accounting for population growth. World Bank Econ Rev 22(2):233–248
Geary F, Stark T (2002) Examining Ireland's post-famine economic growth performance. Econ J 112:919–935
Gerschenkron A (1947) The soviet indices of industrial production. Rev Econ Stat 29(4):217–226
Good D, Ma T (1999) The economic growth of central and eastern Europe in comparative perspective, 1870–1989. Eur Rev Econ Hist 3(2):103–137
Istat (1957) Indagine statistica sullo sviluppo del reddito nazionale dell'Italia dal 1861 al 1956. Ann Stat 8(9):1–271
Kim S (1998) Economic integration and convergence: U.S. regions, 1840–1987. J Econ Hist 58(3):659–683
Kravis IB, Heston A, Summers R (1978) Real per capita income for more than one hundred countries. Econ J 88:215–242
Krugman P (1991) Increasing returns and economic geography. J Polit Econ 99(3):483–499
Kuznets S (1934) National income 1929–1932. National Bureau of Economic Research, New York
Kuznets S (1955) Economic growth and income inequality. Am Econ Rev 45(1):1–28
Lequiller F, Blades D (2006) Understanding national accounts. OECD, Paris
Lucas R (2000) Some macroeconomics for the 21st century. J Econ Perspect 14(1):159–168
Maddison A (1991) A revised estimate of Italian economic growth, 1861–1989. Banca Nazionale Lav Q Rev 44(177):225–241
Maddison A (1994) Confessions of a chiffrephile. Banca Nazionale Lav Q Rev 47(189):123–165
Maddison A (1995) Monitoring the world economy 1820–1992. Development Centre Studies, OECD, Paris
Maddison A (2001) The world economy: a millennial perspective. Development Centre Studies, OECD, Paris
Maddison A (2006) The world economy. A millennial perspective and volume II: historical statistics, Development Centre Studies, vol I. OECD, Paris
Maddison A (2007) Contours of the world economy I-2030 AD. Oxford University Press, Oxford
Malanima P (2006) Alle origini della crescita in Italia 1820–1913. Riv Storia Econ 22(3):306–330
Malanima P (2011) The long decline of a leading economy: GDP in central and northern Italy, 1300–1913. Eur Rev Econ Hist 15(2):169–219
Mankiw NG, Romer D, Weil DN (1992) A contribution to the empirics of economic growth. Q J Econ 107(2):407–437
Martínez-Galarraga J, Rosés JR, Tirado D (2014) The evolution of regional income inequality in Spain, 1860–2000. In: Rosés JR, Wolf N (eds) Europe's regions, 1900–2010. A new quantitative history of the economic development of Europe. Routledge, New York
Myrdal G (1957) Economic theory and underdeveloped regions. Hutchinson, London
Nussbaum M (2000) Women and human development: the capabilities approach. Cambridge University Press, Cambridge, UK
Prados de la Escosura L (2000) International comparisons of real product, 1820–1990: an alternative data set. Explor Econ Hist 37(1):1–41
Prados de la Escosura L (2003) El progreso económico de España (1850–2000). Fundación BBVA, Bilbao
Prados de la Escosura L (2007) When did Latin America fall behind? In: Edwards S, Esquivel G, Márquez G (eds) The decline of Latin American economies: growth, institutions, and crises. University of Chicago Press, Chicago
Prados de la Escosura L (2010) Improving human development: a long-run view. J Econ Surv 24(5):841–894
Prados de la Escosura L (2013) Human development in Africa: a long-run perspective. Explor Econ Hist 50(2):179–204
Prados de la Escosura L (2015) World human development, 1870–2007. Rev Income Wealth 61(2):220–247

Prados de la Escosura L (2016) Mismeasuring long-run growth: the bias from splicing national accounts—the case of Spain. Cliometrica 10(3):251–275
Ravallion M (2012a) Mashup indices of development. World Bank Res Obs 27(1):1–32
Ravallion M (2012b) Troubling tradeoffs in the human development index. J Dev Econ 99 (2):201–209
Romer P (1986) Increasing returns and long run growth. J Polit Econ 94(5):1002–1037
Rosés JR, Wolf N (eds) (2014) Europe's regions, 1900–2010. A new quantitative history of the economic development of Europe. Routledge, New York
Rosés JR, Martínez-Galarraga J, Tirado DA (2010) The upswing of regional income inequality in Spain (1860–1930). Explor Econ Hist 47(2):244–257
Sala-i-Martin X (1996) Regional cohesion: evidence and theories of regional growth and convergence. Eur Econ Rev 40(6):1325–1352
Sala-i-Martin X, Doppelhofer G, Miller RI (2004) Determinants of long term growth: a Bayesian averaging of classical estimates (BACE) approach. Am Econ Rev 94(4):813–835
Schumpeter JA (1950) Wesley Clair Mitchell (1874–1948). Q J Econ 64(1):139–155
Sen AK (1985) Commodities and capabilities. Oxford University Press, Oxford
Solow RM (1956) A contribution to the theory of economic growth. Q J Econ 70(1):65–94
Solow RM (1957) Technical change and the aggregate production function. Rev Econ Stat 39(3):312–320
Stone R (1956) Quantity and price indexes in national accounts. OEEC, Paris
Stone R (1961) Input-output and national accounts. OEEC, Paris
Swan T (1956) Economic growth and capital accumulation. Econ Rec 32(2):334–361
Theil H (1967) Economics and information theory. North Holland, Amsterdam
Waring M (1988) If women counted. A new feminist economics. Macmillan, London
Williamson JG (1965) Regional inequality and the process of national development: a description of the pattern. Econ Dev Cult Chang 13(4):3–84
Williamson JG (2011) Industrial catching up in the poor periphery 1870–1975. CEPR discussion paper no 8335

Cliometric Approaches to International Trade

Markus Lampe and Paul Sharp

Contents

Abstract

This chapter gives a broad overview of the literature on the cliometrics of international trade and market integration. We start by motivating this by looking at the lessons from economic theory and, in particular, through the work which considers the effect of trade, openness, and trade policy on growth. Here theory, as well as empirical results, suggests no clear-cut relationship and points to the richness of historical experiences. We then turn to the issue of how to quantify trade and market integration. The former usually relies on customs records and the latter on the availability of prices in different markets. We then go one step back and look at the determinants of trade, usually tested within the framework of the gravity equation, and discuss what factors were behind periods of trade increases and declines and of market integration and disintegration. Finally, as one of the most important determinants of trade, and perhaps the most policy relevant, we include a separate section on trade policy: we both consider the difficulties of constructing a simple quantitative measure and look at what might explain it.

M. Lampe (✉)
Vienna University of Economics and Business, Vienna, Austria
e-mail: markus.lampe@wu.ac.at

P. Sharp
University of Southern Denmark, Odense M, Denmark
e-mail: pauls@sam.sdu.dk

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_8

Keywords
Cliometrics · International trade · Market integration · Tariffs

Why Look at Trade?

International trade can be considered a "biased" iceberg that stands out from the national economy and extends into foreign countries. As a topic in economic history, it has spawned a huge literature and with good reason. Adam Smith (1776) argued that trade would increase the "extent of the market," allowing for increased specialization and economic growth. David Ricardo (1817), inspired by the Methuen Treaty between Portugal and Britain which caused some specialization in port wine in the former and textiles in the latter, developed the concept of comparative advantage. He demonstrated, using the first mathematical model in economic theory, that since the opportunity costs of producing a good will differ in different countries, they can gain by trading and specializing according to their comparative advantages. Based on the trading patterns of the nineteenth century, which we will examine in more detail in the next section, Heckscher (1919) and Ohlin (1933) elaborated on the concept of comparative advantage, arguing that it was based on the relative endowments of different factors of production. More recently, new trade theory, particularly associated with the work of Paul Krugman (1979), has demonstrated how modern trade leads to trade in similar but differentiated goods, which is a gain for consumers, who have a love of diversity, although recent work for the case of Germany by Hungerland (2017) suggests that the welfare gains from increasing product variety were actually greater for the period before the First World War than today. Lastly, in as much as openness to trade leads to the spread of knowledge between countries, it can also lead to permanent gains in the growth of economies, rather than the one off gain from the exploitation of comparative advantages through a movement from autarky to free trade.

Economic history can also allow us to nuance the work of economic theorists, however. It has been pointed out that the UK and the USA both developed under protectionist regimes, and similar points have been made more recently on the emergence of the so-called tiger economies of Southeast Asia. Thus, even the father of the Washington Consensus, John Williamson (1990b), concluded that one exception from the general rule that free trade is always best is infant industry protection, whereby emerging industries are offered temporary protection so that they can enjoy the so-called dynamic comparative advantages which are not available at the initial stages of production. If these industries then allow for greater productivity growth than traditional sectors and they have spillover effects on the rest of the economy, then such temporary protection should increase incomes in the long run.

Thus, while no sensible economic theory offers the conclusion that autarky is preferable to an open economy, there are studies that argue for potentially positive outcomes from selective temporary protection of specific sectors under the "infant industry" and similar arguments (see Rodríguez and Rodrik 2000, pp. 267–272; O'Rourke 2000 for overviews). Such arguments highlight that specialization on the

production of "non-dynamic" (e.g., agricultural) commodities can, despite yielding static welfare maximization, lead to lack of development possibilities. Widening the domestic industrial base and aiding the self-discovery of nontraditional productive activities can lead to the evolution of new, more dynamic comparative advantages, which under direct world market pressure could not be effectively developed. If the resulting economic activities lead to higher economic growth and domestic knowledge development with concurrent spillovers – in the tradition of "new" endogenous growth theory – then temporary protection would be justified for the sake of long-term growth and development. However, as has already been highlighted above, foreign trade can also be a channel for knowledge transfer, and hence, trade barriers would act as barriers to the world technology pool and hence retard domestic productivity growth, so that successful "infant industry" protection would require both a wider growth-promoting macroeconomic environment and minimization of trade policy distortions.

Economic theory has thus been shaped by historical developments, and trade has been central to the development of economies over time and space and is thus a worthy focus of the efforts of cliometricians. In the following, we have surveyed papers from 2008 through 2017, plus older papers which were particularly relevant, although this is by no means a comprehensive study, and we rely on existing surveys where possible.

In relation to the cliometrics of international trade, we start by assessing the consequences of trade, which, according to standard theory, is directly related to understanding the sources of trade, since the standard textbook comparison of "autarky" and "free trade integration" predicts that adjustments in welfare, productive activity, factor remunerations, etc. will reflect these underlying sources. Hence, there is space for studies trying to assess the effect of trade, besides other "domestic" factors, on economic performance, as well as indirectly through the determinants of the latter (technological progress, technology transfer, institutions, and politics) and changes in the former (such as capital accumulation, natural population growth, and relative remunerations of factors of production), apart from the interplay between factor movements (foreign investments, migration) and trade. In the following, we provide a relatively concise survey focused more on methodology than on findings, since a recent chapter by Meissner (2014) in the *Handbook of Economic Growth* provides a comprehensive treatment of "Growth from Globalization," and Donaldson (2015) provides a similar survey on the "The Gains from Market Integration."

Turning to the big questions of the effects of international integration, two large questions stand out, to which economic historians have provided quantitative answers for the late nineteenth and early twentieth centuries: Does trade cause economic growth? And were trade and factor mobility substitutes or complements?

Regarding the first question, Irwin and Terviö (2002) use an identification strategy developed by Frankel and Romer (1999) to evaluate the impact of trade openness on growth net of the trade-enhancing effect of economic growth. This method consists of using standard gravity variables (distance, population, area, border, landlocked – see below) in a first stage to create "exogenous" trade shares

(aggregating bilateral trade per country) to be regressed onto income levels. Irwin and Terviö find that the coefficient for the trade share in their second-stage regressions for 1913, 1928, and 1938 is always positive but significant in only a few regressions, which might in part be due to small samples of 23–41 observations.

As for the second question, Collins et al. (1999) find that between 1870 and 1940, it is difficult to assess whether trade, capital flows, and international migration were substitutes or complements; although they quite clearly reject that trade and labor mobility were substitutes, for capital flows the findings are more ambiguous between complementarity and substitutability. They also highlight that both trade and migration policy might have influenced the actual historical outcomes. Both papers thus hint at history being richer and more complicated than standard theory might predict.

However, the Heckscher-Ohlin framework of relative factor prices and factor price convergence as a consequence of commodity market integration (see below) to explain the nineteenth-century globalization was behind the hugely successful research program leading to O'Rourke and Williamson's (1999) seminal monograph on *Globalization and History*. The underlying papers (O'Rourke and Williamson 1994, 1995, 1997; O'Rourke et al. 1997; O'Rourke 1997) have shown that commodity market integration went along with factor price equalization, especially regarding the ratio of wages to land rents, which increased in labor-abundant, land-scarce Europe but decreased in the land-abundant, labor-scarce New World, thanks to international migration, trade, and investments. Despite some criticism, for example, of the underlying data and interpretation of the Swedish case (Bohlin and Larsson 2007; Prado 2010), this account has become the standard reference in research and teaching of the nineteenth-century globalization.

Another central line of research focuses on the evolution of the early modern Atlantic economy, in which trade was not necessarily positive for welfare and development: Nunn (2008; see also Nunn and Puga 2012) finds that the slave trade had a clearly negative effect on the economic performance of the African regions that were most affected, not so much due to classical "direct" allocation effects but through the indirect impact via two not necessarily exclusive channels – boosting ethnic fragmentation and debilitating state capacity formation. This, of course, hints at the interplay between trade and domestic institutions and politics, a central topic in recent empirical growth economics. Acemoglu et al. (2005) find that, in Western Europe, the "central corner" of the Atlantic triangle, related trade was not large enough to directly boost economic growth significantly via capital accumulation or static gains from trade, but it increased the weight of merchants in political processes and thereby helped to tilt the political equilibrium toward institutional arrangements that favored trade and eventually economic growth via North and Thomas' (1973, p. 1) "efficient economic organization" via property rights and related "inclusive institutions."

This literature adds new layers onto an older literature regarding the role of trade in the "Great Divergence" with the "Rise of Western Europe," on the one hand, and African, Asian, and Latin American "backwardnesses" on the other. The relatively small importance for this trade on the European side has been highlighted by

O'Brien (1982) and is mirrored in Acemoglu et al. (2005), the O'Rourke and Williamson (2002b) assessment of the sources of early modern trade growth, as well as most recent discussions of the sources of the British Industrial Revolution. The latter often discard an important initial role for trade (Harley 2004; Mokyr 2009; McCloskey 2010), despite updated accounts on the volume and the working of the triangular trade (Inikori 2002) as well as selected links with welfare and economic activity in selected British ports (Draper 2008 for London shipbuilding, Richardson 2005 for Bristol), and for inventions and productivity in certain industries (Zahedieh 2013 for the British copper industry).

In an attempt to quantify the possible welfare losses for Britain from significantly reduced access to international markets, Clark et al. (2014) show in the context of a static standard computable general equilibrium model that relatively small welfare losses of 3–4% would have occurred in 1760, while increasing dependency on foreign trade, especially by the rapidly growing textile industry, would have implied substantial static welfare losses of 25–30% in 1850 by reducing access to foreign endowments and markets substantially. Beyond highlighting the importance of trade for the deployment of the industrial revolution, Allen (2003, 2011) has highlighted that the centrality of Britain in early modern international trade bore an important direct responsibility for the development of energy-intensive, labor-saving innovations that became a central feature of the industrial revolution, by raising real wages and making labor relatively expensive in Britain.

It is not only cliometricians of the British Industrial Revolution who have worked on the causal link between trade and economic performance and the role of international supply and demand versus domestic forces. A variety of studies with different approaches have emerged for mostly "peripheral" players in the emerging international economy of the nineteenth and early twentieth centuries. For Italy, Pistoresi and Rinaldi (2012) use cointegration analysis to assess (Granger-)causal relationships between imports, exports, and GDP. Bajo Rubio (2012) and Guerrero de Lizardi (2006) have conducted similar analyses, explicitly testing for the existence of balance of payment constraints to economic growth in Spain and Mexico, respectively, that is, structural limitations to conduct necessary imports for balanced economic growth. Other studies using cointegration analysis of the effect of trade on domestic economic activity include Greasley and Oxley (2009) on the pastoral boom in New Zealand after the invention of refrigerated long-distance transport and Boshoff and Fourie (2010) on the importance of both provisioning for ship traffic around the Cape of Good Hope and travelers stopping there during their journey to the East Indies, an early form of tourism, for agricultural activity in the Cape Colony. Somewhat connected to Allen's argument, Huff and Angeles (2011) show that globalization had a causal impact on urbanization in Southeast Asia prior to World War I, without leading to industrialization, simply by increasing demand from industrializing markets in the center of the world economy, fomenting commercial production and infrastructure investments, and accompanying overhead services in administrative and commercial centers.

Other authors have used different versions of input-output analysis to assess the relative importance of foreign versus domestic demand and supply forces in

structural models: Bohlin (2007) looks at Sweden before World War I and Kauppila (2009) at Finland during the Great Depression, and Taylor et al. (2011) show, using Leontief's original 1947 input-output table, that (mostly US financed) exports to Europe in the immediate postwar years (1946–1948) helped to avoid increasing US unemployment during the reconversion from a war-oriented to a civilian economy (Leontieff 1953). Ljungberg and Schön's (2013) comparative assessment of the drivers of industrialization in the Nordic countries shares a similar analytic framework but uses shift-share analysis.

Returning to internationally comparative studies and channels between trade and economic growth, Liu and Meissner (2015) derive a new, theoretically consistent measure of market potential and assess whether differences in domestic and foreign markets contribute to explain productivity differentials between the USA and the other countries on the eve of World War I. They find that productivity/GDP per capita is significantly related to market access but that its substantive significance vis-à-vis other factors is relatively minor. Madsen (2007) has shown that bilateral trade was a decisive channel for technology transfer and hence total factor productivity (TFP) growth and convergence for current OECD countries over the 135 years from 1870 to 2004, thereby extending findings by Coe and Helpman (1995) beyond recent periods. López-Córdova and Meissner (2008) examine the link between trade and democracy, and Huberman and Meissner (2010) show that bilateral trade was a diffusion channel especially for the adoption of basic labor protection legislation, such as factory inspection and minimum work ages for children. Vizcarra (2009) demonstrates how the Peruvian guano boom helped the country to return to international capital markets despite domestic political instability and a history of defaults. This finding seems to suggest that at least some forms of trade, controlled by foreign customers and investors, can be substitutes for "real" political and institutional reforms, a recurrent theme in the literature on modern commodity booms and the "resource curse" in developing countries. The importance of institutions[1] is also emphasized by Pascali (2017), who finds that the 1870–1913 globalization benefited only the small number of countries which were characterized by more "inclusive" institutions and thus was one of the main drivers of the economic divergence between rich and poor countries in that period. He identifies this effect by measuring the changes in trade distances brought about by the introduction of the steamship, which he also finds had a large effect on the patterns of trade worldwide.

In this context one final strand of literature, related to specialization resulting from international trade, merits attention: the debate on the role of the specialization in primary commodities for the growth perspectives of developing countries. This topic, promoted in economic history by Jeffrey Williamson and coauthors, for example, in his 2011 book on *Trade and Poverty* (Williamson 2011), has three strands: first, the original Prebisch-Singer finding of falling secular terms of trade for primary commodities that structurally harm the purchasing power of primary

[1]See also the survey by Nunn and Trefler (2014).

producers. Here, one recent comprehensive article by Harvey et al. (2010) underlines that over the last four centuries, for 11 out of 25 commodities studied, relative price trends were significantly negative, while for none was a significantly positive trend found, underlining that Prebisch-Singer forces are at work (Prebisch 1950; Singer 1950). A second strand focuses on deindustrialization and losses of dynamic development possibilities resulting from such specialization via Dutch disease forces or because of forces modelled, e.g., in Matsuyama (1992) and the infant industry literature. Hadass and Williamson (2003) and Williamson (2008) offer a comprehensive assessment of the effect of terms of trade on economic performance before World War I. Third, recent literature has highlighted that more than long-run trends in relative prices, the higher volatility of prices for primary products versus manufactures has harmed economic performance, investment, etc. in developing countries (Blattman et al. 2007; Williamson 2008; Jacks et al. 2011b). Country studies, conducted by Williamson and coauthors (e.g., Dobado González et al. 2008 for Mexico, Clingingsmith and Williamson 2008 for India; Pamuk and Williamson 2011 for the Ottoman Empire) and others (Federico and Vasta 2012 for Italy, Beatty 2000 for Mexico), serve to complement these comparative-econometric findings with historical case studies of channels, mechanisms, and their importance relative to domestic forces. For China, Mitchener and Yan (2014) find that as that country opened up to trade in the early twentieth century, its exports became more unskilled-intensive and its imports became more skill-intensive and identify this effect using the exogenous shock of the First World War, which led to greater demand for Chinese unskilled industries in particular and an improvement in its terms of trade. This might explain the decline in Chinese wage inequality over this period, in contrast to findings for developing countries during the recent globalization.

The impact of trade policy (to which we return in the last section) on the economy has been investigated in different frameworks. The first, already mentioned above and discussed in more detail below, is the gravity equation and the question whether tariffs and other trade policy components affect (reduce or divert) imports or exports. In a similar vein, researchers have asked if trade policy affects relative prices and factor incomes and, as exemplified in O'Rourke's (1997) study of the grain invasion, have found that this is normally the case. These findings imply that trade restriction via trade policy normally works, although trade policy might not translate 1:1 into the desired effects due to varying elasticities of demand and substitution between international and import-competing goods, both on the side of domestic suppliers and the preferences of domestic consumers.

In economic history, several studies since the seminal and controversial contribution of Bairoch (1972) have run growth regressions to estimate the impact of "average tariffs" on growth. The main finding is that of a "tariff-growth paradox" following the widely cited article by O'Rourke (2000) and subsequent papers by Vamvakidis (2002), Clemens and Williamson (2004), and Jacks (2006b). The robustness of these findings has been challenged by results with different methodologies and samples, including Foreman-Peck (1995), Irwin (2002), Athukorala and Chand (2007), Madsen (2009), Tena-Junguito (2010a), Schularick and Solomou

(2011), and Lampe and Sharp (2013).[2] Recent research has moved toward a clearer identification of the underlying channels of an existing or nonexisting tariff-growth paradox: Lehmann and O'Rourke (2011) find that before 1914, tariffs on manufactured goods were growth enhancing, while tariffs on agricultural commodities were probably harmful, and revenue tariffs on luxury goods and "exotic" products had no effect on growth. Tena-Junguito (2010a) finds that the skill bias of tariffs, one of the measures developed to assess not the average level but the structure of tariffs, is significantly related to growth before 1914. Lampe and Sharp (2013) have highlighted that the other side of a potential reverse causality circle is also of interest, since in many countries, tariff liberalization was preceded (and "Granger caused") by higher-income levels, presumably due to their effect on increased fiscal capacity to generate non-customs revenues (see, e.g., Aidt and Jensen 2009).

A novel investigation using prehistoric data by Maurer et al. (2018) suggests that locations on the Mediterranean which were better connected during the time of the Phoenicians, who were among the first to systematically cross the open seas, are associated with the location of Iron Age archaeological sites, suggesting that these areas were more developed due to trade. Otherwise, on a country level, Athukorala and Chand (2007) have studied the tariff-growth relationship for Australia over more than 100 years. Broadberry and Crafts (2010) have surveyed the interplay between trade openness, labor productivity, and structural change in Britain since 1870. Ploeckl (2013) shows that Baden's adhesion to the German Zollverein in 1836 had "traditional" effects on economic performance via increased market access but also led to the investment of Swiss entrepreneurs in Baden due to the higher external tariff Swiss exports faced toward the new customs area. Kauppila (2008) has studied the impact of tariffs on industrial activity and prices in interwar Finland. Tirado et al. (2013) combine new economic geography and an assessment of tariffs in their study of the effect of a gradual closing of the Spanish economy between 1914 and 1930 on the evolution of the regional wage structure. In the case of Spain, the post-Civil War (1936–1939) dictatorship under Generalísimo Franco is an especially interesting field of study, since it tried to run the country on an autarky basis. The macroeconomic consequences of this and the stepwise reforms during the 1950s have been ingeniously investigated by Prados de la Escosura et al. (2012); Martínez Ruiz (2008) has studied the impact of autarky policy on industrial efficiency (in 1958) via the domestic resource cost (DRC) indicator; and Deu and Llonch (2013) focus on the technological backwardness of the Spanish textile industry as a consequence of closed channels for embodied technology transfer. A related topic is import-substituting industrialization (ISI) in Latin America, whose strategies and results have been systematically investigated in Taylor (1998). Debowicz and Segal (2014)

[2]Lampe and Sharp look at a large number of individual country level ECMs for data on average tariffs and growth, identifying a multitude of different relationships for different countries and different periods. A similar framework has been adopted recently by Federico et al. (2017) for the trade-growth nexus, with similarly diverse results.

shed new light on the role of ISI for structural change and industrialization in a dynamic computable general equilibrium model for Argentina.

Finally, a few studies have used cliometric methods to study the effect of specific tariffs on the emergence of individual industries. The classical studies in this case are Head's (1994) study of the protection of US steel rails and Irwin's (2000) assessment of the US tinplate industry, which contrary to other iron and steel products faced a rather low tariff due to a misplaced comma in the 1864 tariff law. More recently, Inwood and Keay (2013) have studied the role of trade policy in modernizing and expanding the Canadian iron and steel industry in a comprehensive design including a novel identification strategy, and Juhász (2018) finds that regions in the French Empire which became protects from trade with the British during the Napoleonic Wars witnessed advances in the cotton spinning industry. Finally, Henriksen et al. (2012) demonstrate the relevance of the cheese tariff for the profitability of the Danish dairy industry before its eventual takeoff after 1880.

Having established the importance of trade in an historical context, we proceed by dividing this chapter into three further sections, which might be considered to follow a reverse causal structure. Thus, in the next section, we consider the extent of trade over time and space. How do we measure it? What different trade regimes can we identify in history? This, of course, can differ over both time and in the cross section and can be considered both in terms of trade volumes and in terms of market integration, which is measured by looking at prices in different markets. It also connects to the literature on the historical extent of "globalization." The section "What Determines Trade?" goes back one stage further and asks what is behind these different regimes, for example, institutions, technology, and trade policy. The latter deserves a particular mention given its importance for the pattern and extent of trade, as well as its central role, particularly in history, for the economic debate. Especially in the nineteenth century, politicians believed that by regulating trade, they were managing their whole economies. We thus devote the section "And What About Trade Policy?" to the issue of how to measure trade policy and its determinants.

Measuring the Extent of Trade and Market Integration

Before we can examine the effects of trade as discussed above, we need to be able to measure it. Thus, in this section we discuss the measurement of trade and market integration.[3] Clearly, the most direct way to measure the extent of trade is to look at the historical records of trade flows, which were often compiled by the customs authorities. Alternatively, or as a complement to this, cliometricians often measure the extent of market integration, which relies on price information.

[3]We ignore the sizeable literature on domestic market integration here, even though it obviously has a bearing on international trade, and the literature has contributed much to the methodological debate.

In very general terms, cliometricians have argued that the extent of market integration should be measured in terms of adherence to the (transaction cost adjusted) law of one price,[4] i.e., that integrated markets should enjoy an arbitrage-induced equilibrium, whereby prices cannot vary by more than the transaction costs of trading between them. Since market integration should be accompanied by more trade because of lower transaction costs, it should also lead to the effects outlined in the previous section.

The related work on globalization – a major part of the market integration literature – was inspired particularly by the new globalization of the late twentieth century, and the interest of cliometricians soon focused on the late nineteenth century, which they termed the "First Era of Globalization" (much of the early literature is summarized by O'Rourke and Williamson 1999). Exactly how to define globalization was, and is, a moot point. Clearly it should at least involve intercontinental trade, but the work by O'Rourke and Williamson cited in the introduction emphasized in particular that increasing *volumes* of trade were not a sufficient criterion for implying the presence of globalization – after all, intercontinental trade had expanded in previous eras, particularly perhaps with the European "discovery" of the Americas. Nor should it be defined by low-volume, high-price products such as the famous spices from the East, which have been traded for centuries. Instead it should be about the market integration of important, but basic, commodities, such as grains. Thus, in this literature, market integration was taken as an indicator of the increasing interdependence of markets and thus also their "globalization," and globalization is thus simply market integration on a global scale.[5]

To measure the extent of market integration, we simply need prices from different markets. The extent of trade and market integration is clearly linked, although markets might appear integrated even without trade, and there can be large volumes of trade with little market integration, as we discuss below. An important aspect of this is that trade regimes do not simply vary across time, for example, in the sense that the interwar years were more protectionist and with lower levels of trade and less market integration than the late nineteenth century. They also vary across space, so that, for example, Britain and Denmark were more free trading and consequently more internationally integrated in the late nineteenth century than France, the USA, and Sweden, for example. The market integration literature is heavily biased toward an understanding of the time dimension in the sense that many studies look at country pairs, or averages of several countries, and ask whether market integration is increasing or decreasing over time.

Turning first to the measurement of trade, much of the historical metrics have concentrated on tasks prior to the analysis of trade flows and their consequences,

[4]See the useful discussion on this in Persson (2004).

[5]This definition is not uncontroversial. De Vries (2010) distinguishes between soft globalization, which encompasses many things, and might well be applied to the changed trading world after 1500, and hard globalization, or "globalization as outcome," for example, market integration.

that is, the construction of databases and the examination of reliability and usefulness of key sources on cross-border trade. Starting with the most complex task, measuring the growth and geographical composition of world trade in the period prior to international statistical bodies like the UN, IMF, and World Bank and their classification (such as the Standard International Trade Classification) has been undertaken by a series of scholars, with the most recent estimates coming from Klasing and Milionis (2014), Federico and Tena-Junguito (2016), and Fouquin and Hugot (2016).

Klasing and Milionis (2014) calculate a world degree of openness (the ratio of imports and/or exports to GDP) for 1870–1949, which can then be chained with series from other sources such as the Penn World Tables. They contribute little to the understanding of the evolution of trade volumes, since they are aggregating available data from the Correlates of War database built by political scientists (Barbieri et al. 2009; Barbieri and Keshk 2012). Nevertheless, they provide a valuable service as they aim to derive non-PPP-adjusted estimates of national GDPs comparable to the non-PPP US-dollar-denominated trade flows they use; that is, they aim to undo Maddison's (2001) PPP adjustment based on a shortcut method for deriving the relationship of the difference of national to US price levels from a structural equation inspired by Prados de la Escosura (2000).

On the other hand, Federico and Tena-Junguito (2016) actually revise the whole literature on international trade flows from the beginning and succeed in the construction of comparable series from at least 1850 to 1938 based on a broad base of the cliometric literature and a more comprehensive use of historical statistical material. They also estimate world levels of (export) openness, using national export price indices to deflate trade series to make them comparable to Maddison's GDP series. Their work also gives a more detailed overview of previous estimates and yields annual growth rates of world trade and trade for the major regions from 1815 to 1938. In addition, they provide a large variety of price series and estimates of average transaction costs derived from CIF-FOB differences, which they show to be fairly constant over time (at about 7% of commodity values), apparently due to an increase in the average distance commodities traveled as a consequence of falling transport costs for given distances.

Such efforts are built upon two interrelated traditions: one of aggregating national statistics (Bairoch 1973, 1974, 1976; Maddison 1962; Lewis 1981) and the other, more relevant in the present context, of understanding the shortcomings and peculiarities of trade statistics as sources that economists often tend to brush over, while historians may in contrast have exaggerated (Platt 1971; Don 1968). Investigations of national cases like the Netherlands (Lindblad and van Zanden 1989), Belgium (Horlings 2002), Spain (Tena-Junguito 1995), Italy (Tena-Junguito 1989; Federico et al. 2012), China (Keller et al. 2011), and Argentina (Tena-Junguito and Willebald 2013) in the nineteenth and early twentieth centuries have unearthed a variety of peculiarities, most notably (Lampe 2008) underreporting due to smuggling or lack of legal requirement to declare, for example, duty-free imports or exports; differences in the definition, especially in differentiating retained ("special") imports and exports of domestic production from transit and reexport; unreliable practices of gathering

values or converting collected data on quantities into values; and different practices in recording countries of origin and destination, often proxied by last land border or port of consignment, as well as problems with port city entrepôts such as Hamburg for Germany or Hong Kong for China. For a comparative account on the international comparability of origins and destinations, the pioneering study is by Morgenstern (1963) for the first half of the twentieth century, reexamined later by Federico and Tena (1991) as well as Carreras-Marin (2012), Folchi and Rubio (2012), and Carreras-Marín and Badia-Miró (2008) for subsets of countries and commodities over the same period. Lampe (2008) offers a similar investigation for six European countries and the USA in the 1850s–1870s.

For the period prior to the nineteenth century, the problems are even greater since data on port entries, shipment manifests, customs revenues, etc. were in many cases not aggregated at a national level. As a result, they are often difficult to interpret and integrate into a meaningful picture. This leads to generally more qualitative than cliometric accounts, though national experiences and the relative endurance of their researchers provide differences in the state of knowledge.[6] Recently, sophisticated descriptions of international trade flows and shifting comparative advantages for individual countries have received renewed input through studies on Italy (Vasta 2010; Federico and Wolf 2013) and China (Keller et al. 2011), a study that also assesses changes to the intensive and extensive margin (number of available products and product varieties) over time. For late-Victorian Britain, Varian (2016) reaches the perhaps surprising conclusion that that country was at a marked comparative *dis*advantage in a number of manufacturing industries.

Finally, cliometricians are also now discovering the post-1945 period, where international statistics are easier to collect, and comparative accounts for countries and sectors can be more readily constructed. Examples for this include Serrano and Pinilla (2011) and Hora (2012).

Turning now to market integration, relatively little has been written on the general accuracy and usability of available price series. Although the issue is sometimes discussed in individual works, more often than not cliometricians work with "whatever they can get." A couple of useful studies by Brunt and Cannon (2013, 2014) have adopted a more critical stance, however. In the first, they offer a careful evaluation of the so-called *Gazette* prices of grain in England, which have been used in a vast number of studies. They find them to be of generally high quality, but they identify a number of limitations as a general indicator of the levels of prices due to fluctuations in quality, changes in the consumption share of domestic grains, and changes in the definition of the units of observation. In their second study, Brunt and Cannon build on this in order to examine the biases introduced to market integration studies when not taking the weaknesses of the statistics into account.

[6]See, for example, the comparative account of sources and knowledge on Spanish and British colonial trade as a subset of total foreign trade in Cuenca-Esteban (2008), work on Spanish cotton imports in the eighteenth century by Thomson (2008), and the export series for the British American colonies reconstructed by Mancall et al. (2008, 2013).

In particular, this problem arises from using infrequent data to measure the half-lives of price shocks, as we will touch on in the following discussion.

The literature on market integration and how to measure it is vast, and it is difficult to improve on the excellent survey provided by Federico (2012a). The following draws heavily on this. His survey includes everything written on market integration, including working papers, before 31 December 2009, and the reader is referred to this for a more complete survey of the literature prior to this date. Thus, we now summarize this literature and its conclusions but update it with the contributions of the last 5 years.

Within the market integration literature, a multitude of methodologies has been used to provide an econometric estimate of the extent of market integration. Likewise, conclusions differ about the extent of market integration, and a perennial question concerns that of "when globalization began." We start with the methodological debate. One of the main points Federico makes regarding this is that in order to understand market integration, there must be a clear theoretical framework. In particular, it should be understood that it consists of two, separable aspects[7]: first, that the equilibrium level of prices should be identical (the law of one price) and, second, that prices should rapidly return to this equilibrium after a shock (what he terms "efficiency").

Testing the first condition leads to the obvious problem that it is rarely if ever met in practice due to imperfect markets and the presence of transportation and other transaction costs. O'Rourke and Williamson (2004) suggest that the best approach is to look at trends and see whether or not prices are converging over time. However, although this works well for two markets, it becomes rather more complicated as the number of markets increases, and for this reason, most cliometricians have concentrated on price convergence between two markets. Thus, authors such as Persson (2004), Metzler (1974), and O'Rourke and Williamson (1994) have looked at simple graphs or have estimated simple regressions of price gaps or relative prices on trends.

Federico's preferred method, since it allows for the aggregation of price information from a number of markets simultaneously, is to calculate coefficients of variation and to regress these on a trend: a negative and significant coefficient implies integration (σ-convergence).The contribution of groups of markets to changes in dispersion can be calculated using simple variance analysis (Federico 2011, Sharp and Weisdorf 2013). Federico (2012a) notes, however, that inferences on the extent of market integration based solely on prices are risky, except with the addition of other information, particularly on the existence of trade. This is because a decline in the price gap *might* reflect a decline in transaction costs between the two locations, but it might also (or instead) reflect an increase in efficiency or availability of information, or it might reveal indirect arbitrage via other markets between which transaction costs have fallen.

Tests of "efficiency," i.e., the strength of arbitrage forces, on the other hand, follow a number of approaches, each of which also has particular weaknesses. First,

[7]Following Cournot (1838)

cointegration implies that the price differential will return to equilibrium after a shock due to arbitrage. Using the Vector Error Correction Mechanism (VECM), it is possible to both test for the presence of a cointegrating relationship and to estimate the half-life of a shock (see, e.g., Ejrnæs et al. 2008). As Taylor (2001) explains, however, this can lead to an overestimation of the size of the correction as long as transaction costs are positive. Thus, alternative approaches such as the threshold autoregressive (TAR) model have been suggested, which implies that prices only converge up to the "commodity points," i.e., the difference in prices beyond which arbitrage becomes profitable after the payment of transaction costs.[8]

The second approach, co-movement, implies that prices move together due to arbitrage. In its simplest form, this corresponds to the calculation of the coefficient of correlation between two prices or an OLS regression between them. To avoid bias if these prices share a common trend, the data can be de-trended, for example, by first differencing.[9] More recently, a Bayesian approach has also been applied (Uebele 2011). Third, variance tests can reveal that arbitrage has reduced the effects of local shocks, thus decreasing the volatility of prices,[10] although as Federico notes, such declines in variation could also be the result of changes to the weather or technology, for example.

Besides their weaknesses as discussed above, Federico is pessimistic about all these measures of efficiency, since they provide no indication of how to determine the relative strength of market integration (e.g., how close should the correlation between prices be before we claim "strong" integration?). Moreover, successful inference requires that it is possible to distinguish trading and non-trading locations, so we must be certain that common shocks unrelated to arbitrage are not biasing integration measures upward and that models that assume constant parameters (often over very long periods) are well specified. Moreover, it is not clear how the results for several country pairs, e.g., the correlation coefficients of their prices, can be aggregated into a more general and coherent picture. Other difficulties Federico notes are with the available data, which are often too infrequent to measure the speed of adjustment satisfactorily and only available for certain, possibly nonrepresentative, commodities (often grains), a point taken up again more recently by Brunt and Cannon (2014), who also measure the extent of the bias using data for England.

In the following we abstract from the more technical debate about how to test for market integration, and what exactly it means, and summarize some of the most important results from the literature. Federico (2012a) notes that most papers testing for market integration cover relatively short time periods and that there is a preponderance of work on the long nineteenth century, i.e., from the Napoleonic Wars to World War I. He explains that the results can be summed up quite simply. First, before the early modern period, there were waves of integration and disintegration

[8]See, for example, Obstfeld and Taylor (1997), Jacks (2005, 2006a, b).

[9]See, for example, Chartres (1995), Ljungberg (1996), Peña and Sánchez-Albornoz (1984), and Bessler (1990).

[10]See, for example, Shiue and Keller (2007), Persson (1999), and Bateman (2011).

both within Europe and between continents. Second, integration increased in the first half of the nineteenth century, but the process was slowed by increasing protectionism toward the end of the century, culminating in the well-known market disintegration of the interwar years. As Federico (2012a) also noted, the literature on the interwar market integration is perhaps surprisingly thin.[11]

Unfortunately, this generalization masks some debates. For example, although O'Rourke and Williamson (2002a) argue that there was no transatlantic integration in the early modern period, Rönnbäck (2009) sees waves of integration and disintegration, with great variation depending on which routes and commodities are being studied. Jacks (2005) was the first to suggest that markets started to integrate before the mid-nineteenth century. This is supported for the classic example of the trade between North America and Britain by Sharp and Weisdorf (2013), who document evidence for the importance of imports of wheat from the USA to Britain already in the middle of the eighteenth century, but with market integration being continuously disrupted, in particular by the French and Napoleonic Wars.[12] Similarly, but looking more generally at Europe and the Americas, Dobado-González et al. (2012), using a new methodology[13] to test for grain market integration between Europe and the Americas over the eighteenth and nineteenth centuries, find gradual integration with some setbacks. Going back further, more recent work by O'Rourke and Williamson (2009) demonstrates that the European Voyages of Discovery of the fifteenth and sixteenth centuries led to the integration of both European spice markets with those of Asia (despite the attempt to monopolize spice markets) and those within Europe. They would not, of course, classify this as evidence of globalization. De Zwart (2016) finds evidence of commodity price convergence between Europe and Asia for goods traded by the Dutch East India Company (VOC), although much of this was due to the ability of the VOC to control commodity markets. Federico and Tena-Junguito (2017) describe trends in trade and openness for the period 1800–2010, finding that the first globalization started around 1820 and was over by 1870 after which trade continued to grow (apart from during the Great Depression), but openness (and gains from trade) fluctuated greatly. It was only after 1970 that openness again resumed a steady upward trend. Finally, Fouquin and Hugot (2016) use a new database of historical trade 1827–2014 and a theory-based measure for assessing bilateral trade costs (an inverse measure of changes in trade), which are then aggregated as indices along various trade routes. They find evidence that globalization had already begun around 1840 (although it was mostly associated with a greater regionalization of trade patterns) and thus question the role of late nineteenth-century improvements in transportation technology and liberal trade policies.

[11]See the recent paper by Hynes et al. (2012).

[12]The effect of wars is also taken up by Jacks (2011), who looks at England during the French Wars to examine the effect of war on market integration and finds that it was mostly through the disruption of international trade linkages and the arrival of news regarding wartime events. This finding is supported by Brunt and Cannon (2014).

[13]Their methodology makes use of the residual dispersion of univariate models of relative prices between markets.

A similar debate exists for market integration within Europe, with Özcumur and Pamuk (2007) arguing against integration before the nineteenth century and Persson (1999) arguing for grain market integration across Europe already in the eighteenth century. More recent work by Bateman (2011) suggests that markets were as integrated in the early sixteenth as in the late eighteenth century, but with a severe contraction in between, while Chilosi et al. (2013) use a large database on grain prices for 100 European cities to demonstrate that market integration was gradual and stepwise rather than sudden for the period 1620 until World War I.[14]

What Determines Trade?

Trade theory, as outlined briefly above, provides the framework within which economists and cliometricians can understand the reasons for the patterns of trade which they observe. Direct tests of trade theory are, however, rare and often inconclusive, not just in a historical perspective but also for more recent periods. Estevadeordal and Taylor (2002) provide a series of tests of the Heckscher-Ohlin-Vanek theory of trade, that is, whether predicted and observed factor contents of trade for 18 countries, disaggregated by industry, correlated in 1913. For the standard factors of production, capital, and labor, correlations between predicted and observed factor contents are low, while for (especially nonrenewable) natural resources, their findings show that factor abundance and observed trade patterns seem to fit quite well.

A similarly motivated literature examines whether the factor endowment theory in its price version holds, that is, whether "autarky prices" of goods whose production uses a relatively abundant factor are relatively cheap. Normally, autarky prices cannot be observed, so the literature focuses on whether market integration, that is, a reduction in barriers to trade, leads to commodity and factor price convergence following Heckscher-Ohlin arguments. The main exponent of this literature is O'Rourke and Williamson's (1999) Globalization and History and its background papers. However, Bernhofen and Brown (2004, 2005, 2011) have used the actual opening of the isolated Japanese economy after 1853/1857 and its abundant available data for a direct evaluation of the autarky prices of its revealed exports after opening, finding that Heckscher-Ohlin type predictions cannot be rejected or are confirmed by this natural experiment.

Beyond this more or less strictly Heckscher-Ohlin-oriented literature, researchers trying to explain the growth of trade have used empirically less restrictive designs, mostly based on the gravity model, both to explain the growth of world trade in specific periods and when inferring determinants of trade from the immense variation to be obtained from comparing bilateral trade flows in cross section

[14]Analyses of markets outside Europe have generally been neglected, but see the recent study by Panza (2013). With a particular focus on the cotton industry, she shows that the Near East integrated into the global economy at the end of the nineteenth century.

or panel designs. The gravity model departs from a simple but theoretically micro-founded idea borrowed from Newtonian physics: the size of trade flows between two countries is (log) proportional to the size of their respective economies and the economic (geographical, institutional, cultural) distance that separates them. However, theoretical motivations and econometric applications have shown that the simple, "naïve" gravity equation, following Head and Mayer (2014, Eq. 4, p. 138)

$$X_{ni} = GY_i^a Y_n^b \phi_{ni} \tag{1}$$

where the *Y*s are importer and exporter GDPs, ϕ is distance, and *G* is a gravitational (cross-sectional) constant, has important flaws. Based on arguments prominently brought forward first by Anderson and van Wincoop (2003), empirical trade economists now recommend including proxies for the so-called multilateral resistance, that is, country-specific characteristics related to the idea of a "home bias" that make them more or less reluctant to trade internationally. Since these are normally assumed to be time varying, the typical approach is then to include country-year fixed effects, which, however, eliminates any other variable from the regression that is determined annually on the country level – such as GDP, GDP per capita, etc.

Thus, Estevadeordal et al. (2003) have used the gravity equation to assess the drivers behind the "Rise and Fall of World Trade" by first estimating gravity models including transport costs, tariffs, and the currency arrangement of the gold standard and then using the estimate to calibrate counterfactual situations for 1870, 1900, 1929, and 1938, in which these variables take their 1913 values. They find that world trade in 1870 would have been five times larger and world openness (trade/GDP) doubles the actual value. The higher counterfactual versus actual openness would be explained mostly by the spread of the gold standard and lower transport costs, as well as some income convergence, especially before 1900, while tariff changes played no role. The almost 60% higher counterfactual trade and 141% higher counterfactual openness in 1939 estimated by Estevadeordal et al. would have been achieved by avoiding increasing transport costs in the interwar period, maintaining the gold standard at its 1913 level, and avoiding the increases in tariffs that followed, especially after 1929. Some of these results have been reexamined in subsequent studies focusing on individual trade determinants, such as Jacks and Pendakur (2010), surveyed below.

O'Rourke and Williamson (2002b) provide a similar assessment of the drivers of a 1.1% annual growth rate in Europe's intercontinental trade between 1500 and 1800 but have to rely on much scarcer data, combining information on quantities and price gaps. They conclude that between half and two thirds of the post-Columbus trade boom is not explained by decreasing transport costs – which they find to be unstable and negligible due to "monopoly, international conflict, piracy, and government restrictions" (p. 426) – but by increases in European surplus income (i.e., land rent growth) spent on "exotic" commodities. This gave rise to a number of papers discussing "When did globalization begin?" which we survey in the context of the price-based market integration literature below.

For the period from about 1850 to 1940, as well as subperiods motivated by the research question of each study, researchers have used data on trade volumes in the context of the gravity model to investigate the significance and importance of different determinants of trade flows. The following offers a short survey of this literature. Although all gravity models include some proxy for country size (GDP) or productivity/purchasing power (GDP per capita), apart from Estevadeordal et al., the focus of the gravity-based literature is not directly on income growth or convergence as the main determinants of bilateral trade performance.

Distance, by contrast, has attracted considerable attention, especially since the now classical account of the late nineteenth-century globalization. O'Rourke and Williamson (1999) give (exogenous) innovations in transport technology, such as railways and steamships, as the main drivers of market integration during this period. The easiest way of incorporating distance, as done by Estevadeordal et al., is to calculate "effective distance" by multiplying geographic distance with a transport cost factor, traditionally taken from Isserlis' (1938) maritime freight rate index and improved by Mohammed and Williamson (2004). This, however, assumes homogeneity of trade cost developments across routes or actual mode of transportation.[15] Jacks and Pendakur (2010) use more refined data on transport costs by different routes and plausible instrumental variables to argue that it was not transport cost reductions which caused trade to increase but that increased bilateral trade led to increased demand and lower costs for transport services between 1870 and 1913. They then recalculate the sources of trade growth over this period, attributing 76% of it to income growth, 18% to income convergence, and relatively small shares to the gold standard (6%) and declining exchange rate volatility (2%), while the mild increases in average tariffs over the period would have contributed negatively (−1.4%).

However, in subsequent research, Jacks and coauthors (Jacks et al. 2008, 2010, 2011a) have derived a gravity-based measure of trade costs, which theoretically include all costs of conducting international trade as compared to national trade, that is, all determinants of bilateral trade increases not corresponding to income growth. They show that these costs vary significantly between country pairs and for the average of trading partners of individual countries, as well as over time; they are also significantly higher than existing ad valorem freight rate estimates for corresponding connections. For the period 1870–1913, they declined on average by 33%, increased (with considerable fluctuations) by 13% between 1921 and 1939, and decreased by 16% between 1950 and 2000 (Jacks et al. 2011a, pp. 190–192).

When estimating the determinants of these trade costs, distance, tariffs, the gold standard, the British empire, and joint railway density turn out to be significant determinants in the 1870–1913 period (Jacks et al. 2010, p. 135) as well as wider measures of fixed exchange rate regimes, common language, empire membership,

[15] A similar approach has been pursued more recently for the case of Argentina during the Belle Epoque by Pinilla and Rayes (2017), who find an important role for transportation costs for Argentinian export success but a less important role for tariffs.

and shared borders for all three periods (Jacks et al. 2011a, p. 194). Of the 486% growth in world trade between 1870 and 1913, 290% can be explained by the fall in trade costs and the rest mostly by increased output. For the period 1921–1939, they find a 0% increase in world trade, to which an increase in trade costs that would have led to a trade decline by 87% contributed negatively, while an almost equal contribution of income growth nullifies this (Jacks et al. 2011a, p. 195; cf. Jacks et al. 2008, p. 534). The Jacks-Meissner-Novy trade cost measure cannot be used as a measure of economic distance in gravity equations, since it is calculated based on the gravity equation itself. Assessing the importance of its components for systematic changes in trade would therefore imply first calculating the trade cost measure and its quantitative importance for trade and then estimating the determinants of trade costs and proceed from there to indirectly identify their effect on trade. So far, the literature in this direction has not extended beyond the initial contributions described here.

Researchers have, however, estimated the effects of all sorts of trade cost-related determinants of bilateral trade flows in the gravity framework. Related to transport and transaction costs, this includes physical transport infrastructure (railway mileage/density, e.g., in Lew and Cater 2006; Mitchener and Weidenmier 2008) and communication infrastructure to facilitate information flows and shipping coordination (telegraphs as proxied by the bilateral sum of telegrams sent in Lew and Cater 2006). To date nobody has included costs of information transmission or actual volumes of international traffic or information flows, although both, in the sense of Jacks and Pendakur (2010), might be endogenous to trade flows.

The role of exchange rate regimes, especially the gold standard, has also been central to the debate, given its prominence in accounts of both pre-World War I globalization and post-World War I instability and the Great Depression. For the first period, López-Córdova and Meissner (2003) find that the gold standard had considerable trade-enhancing effects: countries on the gold standard traded "up to 30% more with each other than with countries not on gold," so that, had the gold standard not spread widely, world trade in 1913 would have been approximately 20% below its actual level. In a similar fashion, Flandreau (2000), in what seems to have been the first cliometric gravity paper, and Flandreau and Morel (2005) assess the impact of the Scandinavian and Latin Monetary Unions and the Austro-Hungarian currency union on trade flows, finding insignificant effects for the Latin Monetary Union but a significantly positive contribution of the apparently more tightly coordinated currency unions in Austria-Hungary and Scandinavia on trade flows. Timini (2018) zooms in on the Latin Monetary Union and confirms the overall lack of significant trade effects but shows that in the earliest period (1865–1874), the Union had effects but that these were concentrated on the trade between France as its "hub" and the rest of the member countries as "spokes" but were negligible between the latter.

For the interwar period, the formation of trade and currency blocs has been analyzed with special care. Eichengreen and Irwin (1995) found that members of the Commonwealth [Ottawa signatories] and the Reichsmark bloc already traded more with each other in 1928, that is, before they formed "blocs" as a consequence

of the Great Depression. Ritschl and Wolf (2011) have reassessed the issue more formally, modelling endogeneity based on optimum currency area arguments. They essentially confirm that naïvely estimated trade creation among members of the different blocs disappears when accounting for the countries' self-selection into these blocs. Political scientists Gowa and Hicks (2013) have recently revisited the issue with a larger dataset. They confirm that none of the blocs increased trade between their members as a whole and underline political conflict and cooperation between the great powers (and "anchors" of the 1930s blocs) as an important component for understanding interwar trade patterns.

Recently, Eichengreen and Irwin (2010) have shown that, at least in the 1930s, flexible monetary policy and trade restrictions were substitutes, with trade restrictions being used when monetary policy, e.g., under the "straitjacket" of the gold standard, is limited when addressing domestic concerns. This leads us to the next classical determinant of foreign trade: trade policy. Studies have investigated two strands, tariffs (normally proxied by the average ad valorem tariff discussed below) and the effects of trade agreements, proxied by dummy variables. For the former, studies are limited, although Lampe (2008), Flandreau and Maurel (2005), and Estevadeordal et al. (2003) find indications of a significantly negative relationship before World War I. For the same period, Jacks (2006b) shows that both levels and changes of tariffs are positively correlated to a positive balance of payments (scaled to GDP), while Madsen (2001) finds a significantly negative impact of tariffs on trade in the interwar period. Regarding trade agreements, both the benign bilateralism of the mid- to late nineteenth century and the pernicious bilateralism of the interwar period have been evaluated using gravity models.

For the nineteenth-century most-favored nation clause trade agreements, both Accominotti and Flandreau (2008, period 1850–1880) and Ly most-favored nation clause (2003, period 1870–1913) find insignificant coefficients, with the former concluding that seeing the Cobden-Chevalier treaty of 1860 as a cornerstone of the nineteenth-century globalization would therefore be unjustified. Lampe (2009) has reexamined the evidence at the commodity level, arguing that nineteenth-century bilateralism did not actually intend to increase world trade but to exchange preference for specific commodities, for which he does find commodity-specific trade-enhancing effects for the first wave of the European Cobden-Chevalier network (1860–1875).

For the interwar period, apart from the literature cited above, de Bromhead et al. (2019) evaluate the effects of the British increase in tariff rates and quantitative restrictions from 1932 and its shift to imperial preference in the Import Duties Act of 1931 and the Ottawa Agreements of 1932. They show with a detailed commodity-wise database that increased protection explains about one quarter of the fall in British imports but over 70% of the shift in imports toward the production from the British Empire in the 1930s. In this case, protection did matter for total trade, but even more for its geographic composition. Similarly, Gowa and Hicks (2013) find, in a study with aggregate trade flows, that while the Imperial Preference System does not seem to have increased or redirected trade among members significantly, the trade of the UK within the system seems to have been redirected toward the

preference group. In contrast, Jacks' (2014) study of the effects of the 1932 Ottawa Agreements on Canadian trade patterns at the commodity level uses a difference-in-difference approach on trade flows at a quarterly frequency and shows that the Imperial Economic Conference had substantial anticipation effects on Canadian trade with the other signatories but very unclear direct effects once it was in place, leading him to conclude that "the conference was a failure from the Canadian perspective."

Another potentially transaction cost-reducing military-politico-economic institution related to the interwar trade blocs and imperial preference systems discussed above is colonialism, which, due to common economic and legal frameworks, bureaucratic practices, and preferential market access and potentially due to emigration, settlement, and homogeneous culture, might be trade enhancing. Mitchener and Weidenmier (2008) have examined the trade-enhancing consequences of colonial relationships using a large bilateral trade flow dataset for the 1870–1913 period (more than 20,000 observations) and find that empire membership had significantly positive effects on trade, with trade more than doubling between empire members as opposed to nonmembers. These were apparently largest for the relatively small empires of the USA and Spain but also substantial for the British, French, and German colonial empires. In a second step, they reestimate their models with a set of transaction cost (common language, years in empire, imperial currency union) and trade policy-related (empire customs unions and preferential market access proxies) variables and show that all of them are significant determinants of trade, confirming the trade cost decreasing function of empires. Head et al. (2010) have shown that these tend to persist even after independence but decrease over time, probably because of depreciating "trading capital."

Another form of changing political ties is the redrawing of national borders. The Versailles settlement after World War I provides a quasi-natural experiment, especially for parts of prewar Germany, the dissolution of the Habsburg Empire; the independence of Czechoslovakia, Hungary, and Poland; and the formation of Yugoslavia. Border effects are normally estimated from price data, but in a series of papers, Schulze, Wolf, and coauthors (Trenkler and Wolf 2005; Wolf 2005, 2009; Heinemeyer 2007; Schulze and Wolf 2009, 2012; Schulze et al. 2008, 2011) have estimated the effects of old and new borders on new and old political entities using trade statistics on railway shipments between regions and across old and new borders. Two central findings are that borders both tend to be endogenous and their effects persistent over time and, here, ethno-linguistic composition, that is, cultural ties, seems to play an important role for explaining trade flows (Schulze and Wolf 2009; see also Lameli et al. 2015).

Conflicts and military alliances have also been shown to be important determinants of trade flows. Gowa and Hicks (2013) highlight the importance of certain military alliances in the interwar period, while Rahman (2010) assesses the effects of being allied to central naval powers between 1710 and 1938. Glick and Taylor (2010) deal with the relationship between trade and wars and show that wars have a significantly negative impact on trade up to 8 years after they were fought and influence not just trade between opposed parties but also their trade with third

countries. They use their results to quantify the trade loss as a share of world GDP resulting from World War I and World War II at 10% and 17.6% of the respective prewar GDPs, with a corresponding trade-related GDP loss of 4.4% and 4.2%, respectively.

Related to this, some studies have also shown that democratic countries trade more with each other (Gowa and Hicks 2013). The importance of national institutional factors for trade orientation has also been stressed in papers with methodologies different from the gravity equation: Sánchez et al. (2010) have shown that lower levels of land conflicts and more secure land property rights helped raise investment in export-oriented coffee trees and production of coffee in the nineteenth- and early twentieth-century Colombia. Rei (2011) examines the determinants of institutional choices that determined the performance of early modern merchant empires in the long run.

What does the market integration literature contribute to this literature? Clearly, many of the factors identified as being determinants of trade, such as trade policy and wars, will also impact on market integration. Following Harley (1980), O'Rourke and Williamson (1999) are particularly associated with the idea that it was falling transatlantic transport costs which led to the globalization of the late nineteenth century, although Persson (2004) and Federico and Persson (2007) argue that it was largely domestic American transport costs that fell, particularly with the extension of the rail network, rather than transatlantic shipping costs. Their basis for so doing is the calculation of "freight factors," i.e., the cost of shipping a unit of a good divided by the price of the good. This can be considered as an ad valorem measure of shipping costs, equivalent to ad valorem measures of tariffs (see below), and a more accurate indicator of the impact of shipping costs on market integration than standard indicators of real freight rates.

Beyond transport costs, the market integration literature has largely focused on demonstrating the fact that markets integrated and disintegrated, rather than testing and estimating the factors behind this, although reasons are usually suggested. For example, O'Rourke (2006) demonstrates that mercantilist conflicts restricted commodity market integration in the eighteenth century, and Sharp and Weisdorf (2013) identify trade policy, war, and politics as being behind the fluctuating experience of market integration between America and Britain in the eighteenth and nineteenth centuries before the revolutionary changes in transport technology, which to a large part has inspired the nineteenth-century globalization literature. For Europe, specifically the Baltic Sea Region, Andersson and Ljungberg (2015) have demonstrated that the role of distance for market integration disappeared for wheat and rye (although not oats and barley), as it became integrated into the Atlantic Economy. At the other end of the First Era of Globalization, Hynes et al. (2012) show that the disintegration after 1929 was caused by trade barriers, the collapse of the gold standard, and the difficulty of obtaining credit.

A particularly notable contribution to this debate is Jacks (2006a), who directly focuses on the question of what drove commodity market integration in the nineteenth century. Using an impressively large panel of grain prices, he finds econometric evidence for the importance of transport technology, geography, monetary

regimes, commercial networks/policy, and conflict over both the cross-sectional and temporal dimensions. In more recent work, Ejrnæs and Persson (2010) have demonstrated the improvements in market efficiency between Chicago and Liverpool after the establishment of the transatlantic telegraph due to faster arbitrage (efficiency) and quantify the gains in terms of reduced deadweight losses. Steinwender (2018) takes this approach further, finding that the smaller information frictions after the telegraph led to greater, but more volatile, trade flows due to more efficient responses to demand shocks. Finally, using data from the transatlantic slave trade, Rönnbäck (2012) suggests that some of the market integration in the early modern period was due to the increased transit speed of ships, a finding somewhat backed up by Solar and Hens (2015), who find that the duration of voyages to Asia by English East India Company ships fell by between one fourth and one third between the 1770s and the 1820s, largely due to the adoption of copper sheathing, which both increased speed by around 11% and obviated the need for a stop at the Cape. Similarly, Kelly and Ó Gráda (2018) use a large database of daily log entries to estimate daily sailing speed from 1750 to 1850, finding that British, in contrast to Spanish and Dutch ships, saw steady progress.

And What About Trade Policy?

As mentioned in the first section, a key feature of trade in economic history and modern economics is the existence of policy barriers to trade. In principle, trade policy is any policy that affects the volume and value of imports coming into or exports leaving a country. This can be by levying tariff duties and other commodity-specific taxes, which, if not corresponding to exactly equivalent domestic taxes, will introduce changes in the relative prices between imported and domestically produced goods and probably also between the relative prices of different sorts of goods, depending on the rates of these duties and the elasticities of demand, supply, and substitution. Ideally, in order to study trade policy, we would wish to create an aggregate measure of all the various forms of duties, as well as accompanying legislation on related trade costs, such as monopolies, port duties, river and strait/sound tolls, prohibitions, regulations, etc. This is, however, theoretically difficult and practically impossible with the existing historical data.

Most studies thus proxy trade restrictiveness by the so-called average ad valorem equivalent tariff rate (AVE), which, as the name suggests, should proxy for the average ad valorem duty corresponding to the wide range of weight- or volume-specific rates and other duties importers or exporters would have to pay at the toll house or the customs office. In practice, this is normally estimated as the ratio of customs receipts to total imports, whenever possible separating import from export duty receipts. Among economic historians, this measure has received wide criticism on several accounts. First, it does not account for nontariff barriers, that is, prohibitions or restrictions like quotas or red-tape requirements that discourage trade. Second, it effectively weights rates for individual commodities by their share of imports, which would be affected by the structure of tariff rates if this is not perfectly

balanced out to be non-distortionary (Estevadeordal 1997, pp. 91–93). Third, it does not distinguish between protective tariffs, which effectively distort the domestic-to-world market price relationship, and the so-called fiscal tariffs, levied on demand-inelastic goods, and often those which are not produced domestically as an easy way to collect an indirect tax on the consumption of "luxury goods." This final point is particularly important in the nineteenth century, when large parts of government revenue in many countries are raised from such import duties (Tena-Junguito 2006a, 2010a), although the solution is not obvious, since the "fiscal commodities" taxed in this way should have had some domestically produced substitute and hence fiscal duties would distort prices in favor of the producers of those substitutes.

In practice, the wide use of AVEs is generally justified for a couple of reasons (see, e.g., Eichengreen and Irwin 2010, pp. 881–882; Lampe and Sharp 2013). First, given the data constraints, it is extremely difficult to imagine how superior measures might be calculated. Second, AVEs have been shown to correlate significantly with theoretically more consistent measures, both within one country (the USA over the nineteenth to mid-twentieth centuries (Irwin 2010)) and among a wide cross section of countries in the present (Kee et al. 2008). For researchers interested in using AVEs, the standard databases are those underlying Clemens and Williamson (2004), Schularick and Solomou (2011), and Lampe and Sharp (2013).

Alternative measures do exist, however. These are constructed to be more theoretically consistent and have been calculated for certain countries and periods. They include the so-called effective protection rates (Balassa 1965), trade restrictiveness indices (Anderson and Neary 2005), the nominal rate of assistance (Anderson et al. 2008), and Leamer's (1988) trade intensity ratio.

Effective protection rates combine information on tariffs for individual goods with input-output tables to assess the structure of protection between final products, primary materials, and intermediate inputs and weigh these rates accordingly in an overall index. Federico and Tena (1998, 1999) and Tena-Junguito (2006b, 2010b) have calculated effective protection rates for Italy and Spain in selected years between the 1870s and the 1930s based on individual tariff rates for 400–500 commodities and different input-output tables. Bohlin (2005, 2009) has undertaken similar work for Sweden.

The trade restrictiveness index (TRI) by Anderson and Neary (2005) in its simplified Feenstra (1995) and Kee et al. (2009) version is motivated by a computable general equilibrium framework and combines data on tariffs of individual commodities and import demand elasticities, thereby establishing a uniform ad valorem tariff rate calculation equivalent to the same welfare level as the existing structure of varying tariff rates; it can be converted straightforwardly into GDP-share equivalent static deadweight losses (DWLs) from protection. Irwin (2010) and Beaulieu and Cherniwchan (2014) have calculated TRIs and estimated DWLs for the USA and Canada over long periods since the mid-nineteenth century. Irwin (2005, 2007) developed a similar measure based on price data to assess the DWL of the Jeffersonian trade embargo of 1807–1809 (about 5% of US 1807 GDP) and the intersectoral transfers resulting from high tariffs in the USA in the late nineteenth century, for example, the classical transfer from consumers to producers via higher prices for import-competing goods.

Similar considerations are behind the "nominal rate of assistance," developed mainly to assess the degree of agricultural protection as "the percentage share by which government policies have raised (or lowered) gross returns of producers above what these returns would have been without the government's intervention" (Swinnen 2009, p. 1501) by comparing domestic to world market prices for individual goods, adding, if necessary, domestic subsidies to the calculations. Swinnen (2009) has calculated these for a variety of agriculture and animal husbandry products in Belgium, Finland, France, Germany, the Netherlands, and the UK from about 1870 to 1970.

Finally, Estevadeordal (1997) presents results on the "trade intensity ratio" of 18 countries in 1913. This measure estimates a Heckscher-Ohlin-based structural equation for trade flows based on endowments and compares the sum of predicted bilateral trade flows to the actual trade per country, interpreting the residual as a measure of protection (or openness) for the market of each country.

Recent research has also focused on assessing relative rates for different commodity groups, not overall average measures of protection, as in Tena-Junguito (2010a) and Tena-Junguito et al. (2012), who compare manufacturing tariffs and their potential skill bias for a large sample of countries in the nineteenth century, and O'Rourke and Lehmann (2011), who distinguish between agricultural, industrial, and revenue tariffs.

A different but related literature looks at tariffs for individual goods, sometimes only in one country. The major examples here are the British Corn Laws and their sliding scales (Williamson 1990a; Sharp 2010), discussed in a comparative perspective by Federico (2012b), or the US tariff on cottons (Irwin and Temin 2001) and a possible optimum export tariff on American raw cotton exports (Irwin 2003), a topic also worked on for interwar Egypt (Yousef 2000). That constructing comprehensive and comparable time series for individual tariff rates in the long run is a time-consuming and often complicated task is illustrated by Lloyd (2008), who estimates Australian tariffs on road motor vehicles, blankets, and beer from 1901–1902 to 2004–2005.

Other nontariff barriers to trade like prohibitions, quotas, licenses and capital constraints, import and production monopolies, marketing boards, etc. are normally only included in regression designs via proxies. At least for the period between the dismantling of mercantilist policies in the early nineteenth century and the introduction of all sorts of protective measures in the 1930s, nontariff barriers are generally said to have been small, at least outside a small group of commodities like live animals and meat, where public health concerns sometimes led to trade restrictions. For prohibitions, ad hoc adjustment assumptions have sometimes been made, such as twice the rate when imports started being permitted (Tena-Junguito et al. 2012) or 1.5 times the highest rate in other countries (Lampe 2011). Regarding nontariff barriers in the 1930s, Eichengreen and Irwin (2010) provide a summary of the scarce data available on quotas and exchange controls as a part of the trade and payments system. Finally, Ye (2010) investigates the political economy of US trade policy regarding the countries of the Pacific Rim from 1922 to 1962. Other measures of trade policy, like membership of trade blocs or trade agreements and most-favored nation status, have normally been proxied by dummy variables.

Despite the difficulties in defining the extent of trade policy as a simple numerical estimate, we might want to answer what explains it. The consensus seems to be that it emerges mainly as a result of political interest groups reacting to the changes brought by trade on national-, local-, and industry-specific "initial conditions." Thus, explaining trade policy involves disentangling the relative importance of these factors. This is normally done through contemplating just one sector or a relevant sample of the industries which are most affected in order to assess the specific impact on them and their reactions alongside the possibilities to affect policy making at the national level. In this sense, the studies by the political scientist Rogowski (1989) and the cliometrician O'Rourke (1997) on the European reaction to the late nineteenth-century grain invasion are outstanding examples of comprehensive trade policy studies, including initial factor endowments, changes in relative prices and factor incomes due to the inflow of cheap grain, formation of coalitions in policy formation, and trade policy outcomes. As Lehmann and Volckart (2011, p. 29) have summarized it, "Kevin O'Rourke (. . .) argued that where agriculture was concerned, the political choices of governments were related on the one hand to how the grain invasion affected land rents, and on the other to the weight of agricultural interests in domestic politics."

Thus, the key variables to describe agricultural trade policy are (following Swinnen 2009) the weight of agriculture in the economy, the relative income of agriculture, and political institutions and organizations, both as regards the level of democracy and the organization of agricultural interest groups. O'Rourke and Rogowski discuss and evaluate all of them in their comparative framework; Federico (2012b) provides a summary of the relevant forces behind an earlier central episode in agricultural trade policy, the repeal of the British Corn Laws, and parallel and subsequent liberalization of agricultural market access in Continental Europe, thereby summarizing a larger literature with important cliometric contributions (Kindleberger 1975; Bairoch 1989; Schonhardt-Bailey 2006; Montañés Primicia 2006; van Dijck and Truyts 2011). Recently, Lehmann (2010) and Lehmann and Volckart (2011) have studied voting behavior in key elections in Germany in the 1870s and Sweden in the 1880s and found that "agriculture," including small farmers, peasants, and rural workers, at least in Imperial Germany, voted "en bloc" for protection, hinting at low perceived possibilities for intersectoral mobility in the economy (a "specific factor model") by large parts of the rural population, as opposed to the opportunities of workers which might be derived from free trade and structural change. For Sweden, the results are less clear, apparently at least in part due to a much more restrictive franchise.

When assessing trade policy of more than one sector, the issue gets complicated by the fact that now not just the level of protection (e.g., on agriculture) has to be taken into account but also its level in comparison to protection or lack thereof for other sectors, i.e., the structure of trade policy. Thus, the political arena is much more complex. Pahre (2008) has written a whole book on the issue, offering a comprehensive theory of tariff setting, leading to six hypotheses on prices, interest group influence and compensation, country size and transport costs, two corollaries on tariff and price volatility, and several findings regarding the endogeneity and

exogeneity of fiscal revenue constraints and their dependence on customs duties and the interplay between democracy and tariff levels. The second step of his theory, regarding bilateral trade policy negotiations, is discussed below.

Blattman et al. (2002), Williamson (2006), and Clemens and Williamson (2012) provide systematic assessment of correlations between a wide set of variables and the "average tariffs," as measured by AVEs. They find population size (related to relatively low dependence on foreign trade), railroad penetration, urbanization, tariffs of other countries, and tariff autonomy (i.e., political independence versus formal or informal foreign control of trade policy) to be significantly and substantially correlated with tariff levels.

O'Rourke and Taylor (2007) investigate the link between tariffs and democracy and show that the relationship is contingent on the relative factor endowments of the national economy in question. In the case of the nineteenth-century globalization, the land-labor ratio is the most fitting operationalization. Irwin (2008) has highlighted that the use of tariff revenue for infrastructure provision was decisive for the American West to enter into a coalition with the North for high tariffs in the 1820s and 1830s and to swing toward more liberal trade policy later. Eichengreen and Irwin (1995, 2010) have shown that protective tariffs and otherwise restrictive policy can also emerge if no other opportunities for dealing with structural balance of payments deficits are available, in their case the unwillingness to or impossibility of devaluation under the interwar gold standard in the 1930s. Another recurrent aspect, especially in political science, is the importance of "hegemony" (McKeown 1983; Nye 1991; Coutain 2009) or the spread of "ideology" (Kindleberger 1975; Federico 2012b). The latter is especially difficult to measure. Finally, Chan (2008) has elaborated and indirectly tested an institutional economic model to explain the trade policy choices of the Chinese Song and Ming dynasties in the light of a trade-off between economic efficiency (and trade tax revenues) and political authority, a question motivated by the famous Needham puzzle of why modern economic growth did not start in China (Lin 1995).

Bilateral or multilateral negotiations to change trade policy have seldom been the subject of cliometric research, and if they have, the focus has been on their impact on trade flows as discussed above. In his book on the Agreeable Customs of 1815–1914, Pahre (2008) formulates nine hypotheses, three corollaries, two remarks, and one conjecture on the likelihood that individual countries cooperate in bilateral trade treaties and finds that, among other things, larger countries and countries with lower tariffs are more likely to cooperate and that "real" exogenous revenue constraints resulting from low fiscal capacity make cooperation less likely, while endogenous (i.e., politically chosen) revenue constraints increase the scope for cooperation. Lampe (2011) offers an assessment of the political and economic determinants of the Cobden-Chevalier network of bilateral MFN treaties in the 1860s and 1870s in the light of both Pahre's theory and recent contributions by economists Baier and Bergstrand (2004) and Baldwin (1995) as well as the political scientist Lazer (1999), and Lampe and Sharp (2011) use this framework for a cost-benefit analysis of bilateralism, the latter for Denmark, which, despite figuring as a free trader in classical accounts, concluded no substantial trade treaties during

this period. In the context of the effects of trade bloc formation in the 1930s, Ritschl and Wolf (2011) and others discuss its origins in the context of evaluating the endogeneity of these blocs and the resulting econometric challenges.

Conclusion

In this chapter we have argued for the importance of trade in economic history, in particular through its impact on growth. Today, domestic sources of growth play a much more important role, but trade might still be important – by establishing constraints, increasing competition, affecting coalitions and institutions, etc.

After discussing how to measure trade and its related concept of market integration, we then went one step back and discussed what factors were behind different examples of trade increases and declines and of market integration and disintegration. Finally, we honed in on trade policy as one of the most important determinants of trade, as well as perhaps the most policy relevant.

The literature is vast, but important questions remain. Moreover, much work is still being done on collecting trade databases and improving our measures of trade costs.[16] The cliometricians of the future will certainly have plenty of opportunities to make important contributions, not only for economic history but for economics in general.

References

Accominotti O, Flandreau M (2008) Bilateral treaties and the most-favored-nation clause: the myth of trade liberalization in the nineteenth century. World Polit 60(2):147–188

Acemoglu D, Johnson S, Robinson JA (2005) The rise of Europe: Atlantic trade, institutional change and economic growth. Am Econ Rev 95(3):546–579

Aidt T, Jensen PS (2009) Tax structure, size of government, and the extension of the voting franchise in western Europe, 1860–1938. Int Tax Public Finan 16(3):362–394

Allen RC (2003) Poverty and progress in early modern Europe. Econ Hist Rev 56:403–443

Allen RC (2011) Why the industrial revolution was British: commerce, induced invention, and the scientific revolution. Econ Hist Rev 64(2):357–384

Anderson JE, Neary JP (2005) Measuring the restrictiveness of international trade policy. MIT Press, Cambridge, MA

Anderson JE, van Wincoop E (2003) Gravity with gravitas: a solution to the border puzzle. Am Econ Rev 93(1):170–192

Anderson K, Kurzweil M, Martin W, Sandri D, Valenzuela E (2008) Measuring distortions to agricultural incentives, revisited. World Trade Rev 7:4

Andersson FNG, Ljungberg J (2015) Grain market integration in the Baltic Sea region in the nineteenth century. J Econ Hist 75:749–790

Athukorala PC, Chand S (2007) Tariff-growth nexus in the Australian economy, 1870–2002: Is there a paradox? Australian National University, Arndt-Corden Department of Economics working papers 2007–2008

[16]For the latter, see, for example, the recent working paper by Chilosi and Federico (2016).

Baier SL, Bergstrand JH (2004) Economic determinants of free trade agreements. J Int Econ 64(1):29–63
Bairoch P (1972) Free trade and European economic development in the 19th century. Eur Econ Rev 3:211–245
Bairoch P (1973) European foreign trade in the XIX century: the development of the value and volume of exports (preliminary results). J Eur Econ Hist 2:5–36
Bairoch P (1974) Geographical structure and trade balance of European foreign trade from 1800 to 1970. J Eur Econ Hist 3:557–608
Bairoch P (1976) Commerce extérieur et développement économique de l'Europe au XIXe siècle. Mouton, Paris- La Haye
Bairoch P (1989) European trade policy, 1815–1914. In: Peter M, Pollard S (eds) The industrial economies: the development of economic and social policies, The Cambridge economic history of Europe, vol VIII. Cambridge University Press, Cambridge, pp 1–60
Bajo Rubio O (2012) The balance-of-payments constraint on economic growth in a long-term perspective: Spain, 1850–2000. Explor Econ Hist 49(1):105–117
Balassa B (1965) Tariff protection in industrial countries: an evaluation. J Polit Econ 73:573–594
Baldwin RE (1995) A domino theory of regionalism. In: Baldwin RE, Haaparanta P, Kiander J (eds) Expanding membership of the European Union. Cambridge University Press, Cambridge, pp 25–48
Barbieri K, Keshk O (2012) Correlates of war project trade data set codebook, version 3.0. http://correlatesofwar.org
Barbieri K, Keshk O, Pollins B (2009) Trading data: evaluating our assumptions and coding rules. Confl Manag Peace Sci 26(5):471–495
Bateman VN (2011) The evolution of markets in early modern Europe, 1350–1800: a study of wheat prices. Econ Hist Rev 64(2):447–471
Beatty EN (2000) The impact of foreign trade on the Mexican economy: terms of trade and the rise of industry, 1880–1923. J Latin Am Stud 32(2):399–433
Beaulieu E, Cherniwchan J (2014) Tariff structure, trade expansion, and Canadian protectionism, 1870–1910. Can J Econ 47(1):144–172
Bernhofen DM, Brown JC (2004) A direct test of the theory of comparative advantage: the case of Japan. J Polit Econ 112:48–67
Bernhofen DM, Brown JC (2005) An empirical assessment of the comparative advantage gains from trade: evidence from Japan. Am Econ Rev 95:208–225
Bernhofen DM, Brown JC (2011) Testing the general validity of the Heckscher-Ohlin theorem: the natural experiment of Japan. CESifo working paper 3586
Bessler DA (1990) A note on Chinese rice prices: interior markets, 1928–1931. Explor Econ Hist 27:287–298
Blattman C, Clemens MA, Williamson JG (2002) Who protected and why? Tariffs around the world around 1870–1913. Paper presented the conference on the political economy of globalization. Trinity College Dublin, August 2002. http://scholar.harvard.edu/jwilliamson/publications/who-protected-and-why-tariffs-world-around-1870-1938
Blattman C, Hwang J, Williamson JG (2007) The impact of the terms of trade on economic development in the periphery, 1870–1939. J Dev Econ 82:156–179
Bohlin J (2005) Tariff protection in Sweden, 1885–1914. Scand Econ Hist Rev 53(2):7–29
Bohlin J (2007) Structural change in the Swedish economy in the late nineteenth and early twentieth century – the role of import substitution and export demand. Gohlin J papers in economic history 8
Bohlin J (2009) The income distributional consequences of agrarian tariffs in Sweden on the eve of World War I. Eur Rev Econ Hist 14:1–45
Bohlin J, Larsson S (2007) The Swedish wage-rental ratio and its determinants, 1877–1926. Aust Econ Hist Rev 47(1):49–72
Boshoff WH, Fourie J (2010) The significance of the Cape trade route to economic activity in the Cape Colony: a medium-term business cycle analysis. Eur Rev Econ Hist 14:469–503
Broadberry S, Crafts N (2010) Openness, protectionism and Britain's productivity performance over the long-run. Centre for Competitive Advantage in the Global Economy working paper 36

Brunt L, Cannon E (2013) The truth, the whole truth, and nothing but the truth: the English Corn Returns as a data source in economic history, 1770–1914. Eur Rev Econ Hist 17(3):318–339
Brunt L, Cannon E (2014) Measuring integration in the English wheat market, 1770–1820: new methods, new answers. Explor Econ Hist 52:111–130
Carreras-Marín A (2012) The international textile trade in 1913: the role of intra-European flows. Rev Hist Ind 49(2):55–76
Carreras-Marín A, Badia-Miró M (2008) La fiabilidad de la asignación geográfica en las estadísticas de comercio exterior: América Latina y el Caribe (1908–1930). Rev Hist Econ 26(3):355–374
Chan KS (2008) Foreign trade, commercial policies and the political economy of the song and Ming dynasties of China. Aust Econ Hist Rev 48(1):68–90
Chartres JA (1995) Market integration and agricultural output in seventeenth-, eighteenth- and early nineteenth-century England. Agric Hist Rev 43:117–138
Chilosi D, Federico G (2016) The effects of market integration: trade and welfare during the first globalization, 1815–1913. LSE Department of Economic History working paper no. 238
Chilosi D, Murphy TE, Studer R, Tuncer AC (2013) Europe's many integrations: geography and grain markets, 1620–1913. Explor Econ Hist 50:46–68
Clark G, O'Rourke KH, Taylor AM (2014) The growing dependence of Britain on trade during the industrial revolution. Scand Econ Hist Rev 62(2):109–136
Clemens MA, Williamson JG (2004) Why did the tariff-growth correlation reverse after 1950? J Econ Growth 9:5–46
Clemens MA, Williamson JG (2012) Why were Latin American tariffs so much higher than Asia's before 1950? Rev Hist Econ 30(1):11–44
Clingingsmith D, Williamson JG (2008) De-industrialization in 18th and 19th century India: Mughal decline, climate shocks and British industrial ascent. Explor Econ Hist 45(3):209–234
Coe DT, Helpman E (1995) International R&D spillovers. Eur Econ Rev 39(5):859–887
Collins WJ, O'Rourke KH, Williamson JG (1999) Were trade and factor mobility substitutes in history? In: Faini R, de Melo J, Zimmermann K (eds) Migration: the controversies and the evidence. Cambridge University Press, Cambridge, pp 227–260
Cournot A (1838) Recherches Sur les principles mathematiques de la theorie des richessesses. L. Hachette, Paris
Coutain B (2009) The unconditional most-favored-nation clause and the maintenance of the liberal trade regime in the postwar 1870s. Int Organ 63(1):139–175
Cuenca-Esteban J (2008) Statistics of Spain's colonial trade, 1747–1820: new estimates and comparisons with Great Britain. Rev Hist Econ 26(3):323–354
de Bromhead A, Fernihough A, Lampe M, O'Rourke KH (2019) When Britain turned inward: the impact of interwar British protection. Am Econ Rev 109(2):325–352
de Vries J (2010) The limits of globalization in the early modern world. Econ Hist Rev 63(3):710–733
de Zwart P (2016) Globalization in the Early Modern Era: New Evidence from the Dutch-Asiatic Trade, c. 1600–1800. J Econ Hist 76:520–558
Debowicz D, Segal P (2014) Structural change in Argentina, 1935–1960: the role of import substitution and factor endowments. J Econ Hist 74(1):230–258
Deu E, Llonch M (2013) AutarquLl y atraso tecnolltecn en la industria textil espaal e, 1939–1959. Invest Hist Econ 9:11–21
Dobado González R, Gómez Galvarriato A, Williamson JG (2008) Mexican exceptionalism: globalization and de-industrialization, 1750–1877. J Econ Hist 68(3):758–811
Dobado-González R, García-Hiernaux A, Guerrero DE (2012) The integration of grain markets in the eighteenth century: early rise of globalization in the west. J Econ Hist 72(3):671–707
Don Y (1968) Comparability of international trade statistics: Great Britain and Austria-Hungary before World War I. Econ Hist Rev 21:78–92
Donaldson D (2015) The gains from market integration. Annu Rev Econ 7:619–647
Draper N (2008) The city of London and slavery: evidence from the first dock companies, 1795–1800. Econ Hist Rev 61(2):432–466

Eichengreen B, Irwin DA (1995) Trade blocs, currency blocs, and the reorientation of world trade in the 1930s. J Int Econ 38:1–24
Eichengreen B, Irwin DA (2010) The slide to protectionism in the great depression: who succumbed and why? J Econ Hist 70(4):871–897
Ejrnæs M, Persson KG (2010) The gains from improved market efficiency: trade before and after the transatlantic telegraph. Eur Rev Econ Hist 14:361–381
Ejrnæs M, Persson KG, Rich S (2008) Feeding the British: convergence and market efficiency in the nineteenth-century grain trade. Econ Hist Rev 61(S1):140–171
Estevadeordal A (1997) Measuring protection in the early twentieth century. Eur Rev Econ Hist 1:89–125
Estevadeordal A, Taylor AM (2002) A century of missing trade? Am Econ Rev 92(1):383–393
Estevadeordal A, Frantz B, Taylor AM (2003) The rise and fall of world trade, 1870–1939. Q J Econ 118(2):359–407
Federico G (2011) When did European markets integrate? Eur Rev Econ Hist 15:93–126
Federico G (2012a) How much do we know about market integration in Europe? Econ Hist Rev 65(2):470–497
Federico G (2012b) The corn laws in continental perspective. Eur Rev Econ Hist 16:166–187
Federico G, Persson KG (2007) Market integration and convergence in the world wheat market, 1800–2000. In: Hatton TJ, O'Rourke KH, Taylor AM (eds) The new comparative economic history: essays in honor of Jeffrey G. Williamson. MIT Press, Cambridge, MA, pp 87–113
Federico G, Tena A (1991) On the accuracy of foreign trade statistics (1909–1935): Morgenstern revisited. Explor Econ Hist 28:259–273
Federico G, Tena A (1998) Was Italy a protectionist country? Eur Rev Econ Hist 2:73–97
Federico G, Tena A (1999) Did trade policy foster Italian industrialization? Evidence from effective protection rates 1870–1913. Res Econ Hist 19:111–138
Federico G, Tena-Junguito A (2016) World trade 1800–1938: a new data-set. EHES working paper no. 93
Federico G, Tena-Junguito A (2017) A tale of two globalizations: gains from trade and openness 1800–2010. Rev World Econ 153:601–626
Federico G, Vasta M (2012) Was industrialization an escape from the commodity lottery? Evidence from Italy, 1861–1939. Explor Econ Hist 47:228–243
Federico G, Wolf N (2013) A long-run perspective on comparative advantage. In: Toniolo G (ed) The Oxford handbook of the Italian economy since unification. Oxford University Press, Oxford, pp 327–350
Federico G, Natoli S, Tattara G, Vasta M (2012) Il commercio estero italiano 1861–1939. Laterza, Bari
Federico G, Sharp P, Tena A (2017) Openness and growth in a historical perspective: a VECM approach. EHES working paper no. 118
Feenstra RC (1995) Estimating the effects of trade policy. In: Grossman GM, Rogoff K (eds) Handbook of international economics, vol 3. Elsevier, Amsterdam, pp 1553–1595
Flandreau M (2000) The economics and politics of monetary unions: a reassessment of the Latin Monetary Union, 1865–1871. Financ Hist Rev 7:25–43
Flandreau M, Morel M (2005) Monetary union, trade integration, and business cycles in 19th century Europe. Open Econ Rev 16:135–152
Folchi M, Del Mar Rubio M (2012) On the accuracy of Latin American trade statistics: a non-parametric test for 1925. In: Yalch C, Carreras A (eds) The economies of Latin America: new cliometric data, perspectives. Pickering & Chatto, London, pp 67–89
Foreman-Peck J (1995) A model of later nineteenth-century European economic development. Rev Hist Econ 13:441–471
Fouquin M, Hugot J (2016) Back to the future: international trade costs and the two globalizations. CEPII working paper no. 2016–13
Frankel J, Romer D (1999) Does trade cause growth? Am Econ Rev 89:379–399
Glick R, Taylor AM (2010) Collateral damage: trade disruption and the economic impact of war. Rev Econ Stat 92(1):102–127

Gowa J, Hicks R (2013) Politics, institutions and trade: lessons of the interwar era. Int Organ 67(3):439–467
Greasley D, Oxley L (2009) The pastoral boom, the rural land market, and long swings in New Zealand economic growth, 1873–1979. Econ Hist Rev 62(2):324–349
Guerrero de Lizardi C (2006) Thirwall's law with an emphasis on the ratio of export/income elasticities in Latin American economies during the twentieth centuries. Estudios Econ 26:23–44
Hadass YS, Williamson JG (2003) Terms-of-trade shocks and economic performance, 1870–1940: prebisch and singer revisited. Econ Dev Cult Change 51(3):629–656
Harley CK (1980) Transportation, the world wheat trade, and the Kuznets Cycle, 1850–1913. Explor Econ Hist 17:218–250
Harley CK (2004) Trade: discovery, mercantilism and technology. In: Roderick F, Paul J (eds) Industrialisation, 1700–1860, The Cambridge economic history of modern Britain, vol I. Cambridge University Press, Cambridge, pp 175–203
Harvey DI, Kellard NM, Madsen JB, Wohar ME (2010) The Prebisch-Singer hypothesis: four centuries of evidence. Rev Econ Stat 92(2):367–377
Head K (1994) Infant industry protection in the steel rail industry. J Int Econ 37:141–165
Head K, Mayer T (2014) Gravity equations: workhorse, toolkit, and cookbook. In: Gopinath G, Helpman E, Rogoff K (eds) Handbook of international economics, vol 4. Elsevier, Amsterdam, pp 131–195
Head K, Mayer T, Ries J (2010) The erosion of colonial linkages after independence. J Int Econ 81:1–14
Heckscher E (1919) The effects of foreign trade on the distribution of income. Ekonomisk Tidskrift 21:497–512
Heinemeyer HC (2007) The treatment effect of borders on trade. The great war and the disintegration of Central Europe. Cliometrica 1:177–210
Henriksen I, Lampe M, Sharp P (2012) The strange birth of liberal Denmark: Danish trade protection and the growth of the dairy industry since the mid-nineteenth century. Econ Hist Rev 65(2):770–788
Hora R (2012) La evolución del sector agroexportador argentino en el largo plazo, 1880–2010. Hist Agraria 58:145–181
Horlings E (2002) The international trade of a small and open economy. Revised estimates of the imports and exports of Belgium, 1835–1990. NEHA-Jaarboek 65:110–142
Huberman M, Meissner CM (2010) Riding the wave of trade: the rise of labor regulation in the golden age of globalization. J Econ Hist 70(3):657–685
Huff G, Angeles L (2011) Globalization, industrialization and urbanization in Pre-World-War II Southeast Asia. Explor Econ Hist 48:20–36
Hungerland W (2017) The gains from import variety in two globalisations: evidence from Germany. EHES working paper no. 120
Hynes W, Jacks DS, O'Rourke KH (2012) Commodity market disintegration in the interwar period. Eur Rev Econ Hist 16:119–143
Inikori JE (2002) Africans and the industrial revolution in England: a study in international trade and economic development. Cambridge University Press, Cambridge
Inwood K, Keay I (2013) Trade policy and industrial development: iron and steel in a small open economy, 1870–1913. Can J Econ 46(4):1265–1294
Irwin DA (2000) Did late nineteenth century U.S. tariffs promote infant industries? Evidence from the tinplate industry. J Econ Hist 60:335–360
Irwin DA (2002) Interpreting the tariff-growth correlation of the late nineteenth century. Am Econ Rev (P&P) 91(2):165–169
Irwin DA (2003) The optimal tax on antebellum cotton exports. J Int Econ 60:275–291
Irwin DA (2005) The welfare cost of autarky: evidence from the Jeffersonian trade embargo, 1807–09. Rev Int Econ 13(4):631–645
Irwin DA (2007) Tariff incidence in America's gilded age. J Econ Hist 67(3):582–607

Irwin DA (2008) Antebellum tariff politics: regional coalitions and shifting economic interests. J Law Econ 51(4):715–741
Irwin DA (2010) Trade restrictiveness and deadweight losses from US tariffs. Am Econ J Econ Policy 2:111–133
Irwin DA, Temin P (2001) The antebellum tariff on cotton textiles revisited. J Econ Hist 61:777–798
Irwin DA, Terviö P (2002) Does trade raise income? Evidence from the twentieth century. J Int Econ 58:1–18
Isserlis L (1938) Tramp shipping cargoes and freights. J Royal Stat Soc 101(1):53–146
Jacks DS (2005) Intra- and international commodity market integration in the Atlantic economy, 1800–1913. Explor Econ Hist 42:381–413
Jacks DS (2006a) What drove 19th century commodity market integration? Explor Econ Hist 43:383–412
Jacks DS (2006b) New results on the tariff-growth paradox. Eur Rev Econ Hist 10(2):205–230
Jacks DS (2011) Foreign wars, domestic markets: England, 1793–1815. Eur Rev Econ Hist 15:277–311
Jacks DS (2014) Defying gravity: the 1932 imperial economic conference and the reorientation of Canadian trade. Explor Econ Hist 53:19–39
Jacks DS, Pendakur K (2010) Global trade and the maritime transport revolution. Rev Econ Stat 92(4):745–755
Jacks DS, Meissner CM, Novy D (2008) Trade costs, 1870–2000. Am Econ Rev (P&P) 98(2):529–534
Jacks DS, Meissner CM, Novy D (2010) Trade costs in the first wave of globalization. Explor Econ Hist 47(2):127–141
Jacks DS, Meissner CM, Novy D (2011a) Trade booms, trade busts, and trade costs. J Int Econ 83(2):185–201
Jacks DS, O'Rourke KH, Williamson JG (2011b) Commodity price volatility and world market integration since 1700. Rev Econ Stat 93(3):800–813
Juhász R (2018) Temporary protection and technology adoption: evidence from the Napoleonic Blockade. Am Econ Rev 108(11):3339–3376
Kauppila J (2008) Impact of tariffs on industries and prices in Finland during the interwar period. Scand Econ Hist Rev 56(3):176–191
Kauppila J (2009) Quantifying the relative importance of export industries in a small open economy during the great depression of the 1930s: an input-output approach. Cliometrica 3:245–273
Kee HL, Nicita A, Olarreaga M (2008) Import demand elasticities and trade distortions. Rev Econ Stat 90(4):666–682
Kee HL, Nicita A, Olarreaga M (2009) Estimating trade restrictiveness indices. Econ J 119:172–199
Keller W, Li B, Shiue CH (2011) China's foreign trade: perspectives from the past 150 years. World Econ 34(6):853–892
Kelly M, Ó Gráda C (2018) Speed under sail during the early industrial revolution. Econ Hist Rev https://doi.org/10.1111/ehr.12696
Kindleberger CP (1975) The rise of free trade in Western Europe, 1820–1875. J Econ Hist 45(1):20–55
Klasing M, Milionis P (2014) Quantifying the evolution of world trade, 1870–1949. J Int Econ 92(1):185–197
Krugman PR (1979) Increasing returns, monopolistic competition, and international trade. J Int Econ 9(4):469–479
Lameli A, Nitsch V, Südekum J, Wolf N (2015) Same but different: dialects and trade. German Econ Rev 16(3):255–389
Lampe M (2008) Bilateral trade flows in Europe, 1857–1875: a new dataset. Res Econ Hist 26:81–155
Lampe M (2009) Effects of bilateralism and the MFN clause on international trade: evidence for the Cobden-Chevalier network, 1860–1875. J Econ Hist 69(4):1012–1040

Lampe M (2011) Explaining nineteenth-century bilateralism: economic and political determinants of the Cobden-Chevalier network. Econ Hist Rev 64(2):644–668
Lampe M, Sharp P (2011) Something rational in the state of Denmark? The case of an outsider in the Cobden-Chevalier network, 1860–1875. Scand Econ Hist Rev 59(2):128–148
Lampe M, Sharp P (2013) Tariffs and income: a time series analysis for 24 countries. Cliometrica 7:207–235
Lazer D (1999) The free trade epidemic of the 1860s and other outbreaks of economic discrimination. World Polit 51(4):447–483
Leamer EE (1988) Measures of openness. In: Baldwin RE (ed) Trade policy issues and empirical analysis. Chicago University Press, Chicago, pp 147–204
Lehmann SH (2010) The German elections in the 1870s: why Germany turned from liberalism to protectionism. J Econ Hist 70(1):146–178
Lehmann SH, O'Rourke KH (2011) The structure of protection and growth in the late nineteenth century. Rev Econ Stat 93(2):606–616
Lehmann S, Volckart O (2011) The political economy of agricultural protection: Sweden 1887. Eur Rev Econ Hist 15(1):29–59
Leontieff WW (1953) Domestic production and foreign trade: the American capital position re-examined. Proc Am Philos Soc 97(4):332–349
Lew B, Cater B (2006) The telegraph, co-ordination of tramp shipping, and growth in world trade, 1870–1910. Eur Rev Econ Hist 10(2):147–173
Lewis A (1981) The rate of growth of world trade. In: Grassman S, Lundberg E (eds) The world economic order. Past and prospects. Macmillan, London
Lin J (1995) The Needham puzzle: why the industrial revolution did not originate in China. Econ Dev Cult Change 43(2):269–292
Lindblad JT, van Zanden JL (1989) De buitenlandse handel van Nederland, 1872–1913. Econ Soc Hist Jaarboek 52:231–269
Liu D, Meissner CM (2015) Market potential and the rise of US productivity leadership. J Int Econ 96:72–87
Ljungberg J (1996) European market integration and the behaviour of prices, 1850–1914. Lund papers in economic history, 54
Ljungberg J, Schön L (2013) Domestic markets and international integration: paths to industrialization in the Nordic countries. Scand Econ Hist Rev 61(2):101–121
Lloyd P (2008) 100 years of tariff protection in Australia. Aust Econ Hist Rev 48(2):99–145
López-Córdova JE, Meissner CM (2003) Exchange-rate regimes and international trade: evidence from the classical gold standard era. Am Econ Rev 93(1):344–353
López-Córdova JE, Meissner CM (2008) The impact of international trade on democracy. A long-run perspective. World Polit 60:539–575
Maddison A (1962) Growth and fluctuation in the world economy 1870–1960. Banca Nazionale del Lavoro Q Rev 15(61):127–195
Maddison A (2001) The world economy I: a millennial perspective. OECD, Paris
Madsen JB (2001) Trade barriers and the collapse of world trade during the great depression. Southern Econ J 67(4):848–868
Madsen JB (2007) Technology spillover through trade and TFP convergence: 135 years of evidence from OECD countries. J Int Econ 72:464–480
Madsen JB (2009) Trade barriers, openness, and economic growth. Southern Econ J 76:397–418
Mancall PC, Rosenbloom JL, Weiss T (2008) Exports and the economy of the lower south region, 1720–1772. Res Econ Hist 25:1–68
Mancall PC, Rosenbloom JL, Weiss T (2013) Exports from the colonies and states of the middle Atlantic region 1720–1800. Res Econ Hist 29:257–305
Martínez Ruiz E (2008) Autarkic policy and efficiency in the Spanish industrial sector. An estimate of domestic resource costs in 1958. Rev Hist Econ 26(3):439–470
Matsuyama K (1992) Agricultural productivity, comparative advantage, and economic growth. J Econ Theory 58:317–334
Maurer S, Pischke J, Rauch F (2018) Of mice and merchants: trade and growth in the iron age. NBER Working Papers 24825

McCloskey DN (2010) Bourgeois dignity. Why economics can't explain the modern world. Chicago University Press, Chicago
McKeown TJ (1983) Hegemonic stability theory and 19th century tariff levels in Europe. Int Organ 37(1):73–91
Meissner CM (2014) Growth from globalization? A view from the very long run. In: Aghion P, Durlauf SN (eds) Handbook of economic growth, vol 2. Elsevier, Amsterdam, pp 1033–1069
Metzler J (1974) Railroad development and market integration: the case of Tsarist Russia. J Econ Hist XXXIV:529–549
Mitchener KJ, Weidenmier M (2008) Trade and empire. Econ J 118:1805–1834
Mitchener KJ, Yan S (2014) Globalization, trade, and wages: what does history tell us about China? Int Ec Rev 55:131–168
Mohammed SIS, Williamson JG (2004) Freight rates and productivity gains in British tramp shipping, 1869–1950. Explor Econ Hist 41(2):172–203
Mokyr J (2009) The enlightened economy: an economic history of Britain, 1700–1850. Yale University Press, New Haven/London
Montañés Primicia E (2006) Reformas arancelarias y comercio exterior de trigo en España: El fin de la prohibición de importar trigo (1849–1869). Invest Hist Econ 6:73–104
Morgenstern O (1963) On the accuracy of economic observations, 2nd edn. Princeton University Press, Princeton
North DC, Thomas RP (1973) The rise of the Western world. A new economic history. Cambridge University Press, Cambridge
Nunn N (2008) The long-term effects of Africa's slave trades. Q J Econ 123(1):139–176
Nunn N, Puga D (2012) Ruggedness: the blessing of bad geography in Africa. Rev Econ Stat 94(1):20–36
Nunn N, Trefler D (2014) Domestic institutions as a source of comparative advantage. In: Handbook of international economics, vol 4, ch. 5, pp 263–315. Elsevier, Amsterdam
Nye JVC (1991) Revisionist tariff history and the theory of hegemonic stability. Polit Soc 19(2):209–232
O'Brien P (1982) European economic development: the contribution of the periphery. Econ Hist Rev New Series 35(1):1–18
Obstfeld M, Taylor AM (1997) Nonlinear aspects of goods-market arbitrage and adjustment: Heckscher's commodity points revisited. J Jpn Int Econ 11:441–479
Ohlin B (1933) Interregional and international trade. Harvard University Press, Cambridge MA
O'Rourke KH (1997) The European grain invasion, 1870–1913. J Econ Hist 57(4):775–801
O'Rourke KH (2000) Tariffs and growth in the late 19th century. Econ J 110:456–483
O'Rourke KH (2006) The worldwide economic impact of the French Revolutionary and Napoleonic Wars, 1793–1815. J Global Hist 1:123–149
O'Rourke KH, Lehmann S (2011) The structure of protection and growth in the late nineteenth century. Rev Econ Stat 93(2):606–616
O'Rourke KH, Taylor AM (2007) Democracy and protectionism. In: Hatton TJ, O'Rourke KH, Taylor AM (eds) The new comparative economic history: essays in honor of Jeffrey G. Williamson. MIT Press, Cambridge, MA, pp 193–216
O'Rourke KH, Taylor AM, Williamson JG (1997) Factor price convergence in the late 19th century. Int Econ Rev 37(3):499–530
O'Rourke KH, Williamson JG (1994) Late 19th century Anglo-American factor price convergence: were Heckscher and Ohlin right? J Econ Hist 54:892–916
O'Rourke KH, Williamson JG (1995) Open economy forces and late 19th century Swedish catch-up: a quantitative accounting. Scand Econ Hist Rev 43:171–203
O'Rourke KH, Williamson JG (1997) Around the European periphery 1870–1913: globalization, schooling and growth. Eur Rev Econ Hist 1:153–190
O'Rourke KH, Williamson JG (1999) Globalization and history: the evolution of a nineteenth-century Atlantic economy. MIT Press, Cambridge, MA
O'Rourke KH, Williamson JG (2002a) When did globalisation begin? Eur Rev Econ Hist 6(1):23–50

O'Rourke KH, Williamson JG (2002b) After Columbus: explaining Europe's overseas trade boom, 1500–1900. J Econ Hist 62(2):417–456
O'Rourke KH, Williamson JG (2004) Once more: when did globalization begin? Eur Rev Econ Hist 8:109–117
O'Rourke KH, Williamson JG (2009) Did Vasco da Gama matter for European markets? Econ Hist Rev 62(3):655–684
Özmucur S, Pamuk Ş (2007) Did European commodity prices converge during 1500–1800? In: Hatton TJ, O'Rourke KH, Taylor AM (eds) The new comparative economic history: essays in honour of Jeffrey G. Williamson. MIT Press, Cambridge MA, pp 59–86
Pahre R (2008) Politics and trade cooperation in the nineteenth century. The "agreeable customs" of 1815–1914. Cambridge University Press, Cambridge
Pamuk Ş, Williamson JG (2011) Ottoman de-industrialization, 1800–1913: assessing the magnitude, impact, and response. Econ Hist Rev 64(S1):159–184
Panza L (2013) Globalization and the near east: a study of cotton market integration in Egypt and Western Anatolia. J Econ Hist 73(3):847–872
Pascali L (2017) The wind of change: maritime technology, trade and economic development. Am Ec Rev 107:2821–2854
Peña D, Sánchez-Albornoz N (1984) Wheat prices in Spain, 1857–1890: an application of the Box-Jenkins methodology. J Eur Econ Hist 13:353–373
Persson KG (1999) Grain markets in Europe, 1500–1900. Cambridge University Press, Cambridge
Persson KG (2004) Mind the gap! Transport costs and price convergence in the nineteenth century Atlantic economy. Eur Rev Econ Hist 8:125–147
Pinilla V, Rayes A (2017) Why did Argentina become a super-exporter of agricultural and food products during the Belle Époque (1880–1929)? EHES working paper no. 107
Pistoresi B, Rinaldi A (2012) Exports, imports, and growth. New evidence on Italy: 1863–2004. Explor Econ Hist 49:241–254
Platt DCM (1971) Problems in the interpretation of foreign trade statistics before 1914. J Latin Am Stud 3:119–130
Ploeckl F (2013) The internal impact of a customs union; Baden and the Zollverein. Explor Econ Hist 50:387–404
Prado S (2010) Fallacious convergence? Williamson's real wage comparisons under scrutiny. Cliometrica 4:171–205
Prados de la Escosura L (2000) International comparisons of real product, 1820–1990: an alternative data set. Explor Econ Hist 37(1):1–41
Prados de la Escosura L, Rosés JR, Sanz-Villarroya I (2012) Economic reforms and growth in Franco's Spain. Rev Hist Econ 30(1):45–89
Prebisch R (1950) The economic development of Latin America and its principle problems. United Nations Department of Economic Affairs, Lake Success
Rahman AS (2010) Fighting the forces of gravity – seapower and maritime trade between the 18th and the 20th centuries. Explor Econ Hist 47:28–48
Rei C (2011) The organization of eastern merchant empires. Explor Econ Hist 48:116–135
Ricardo D (1817) On the principles of political economy and taxation. John Murray, London
Richardson D (2005) Slavery and Bristol's "golden age". Slavery Abolition 26:35–54
Ritschl AO, Wolf N (2011) Endogeneity of currency areas and trade blocs: evidence from a natural experiment. Kyklos 64(2):291–312
Rodríguez F, Rodrik D (2000) Trade policy and economic growth: a sceptic's guide to the cross-national evidence. NBER Macroecon Ann 15:261–325
Rogowski R (1989) Commerce and coalitions. How trade affects domestic political alignments. Princeton University Press, Princeton
Rönnbäck K (2009) Integration of global commodity markets in the early modern era. Eur Rev Econ Hist 13(1):95–120
Rönnbäck K (2012) The speed of ships and shipping productivity in the age of sail. Eur Rev Econ Hist 16:469–489

Sánchez F, López-Uribe M, Fazio A (2010) Land conflicts, property rights, and the rise of the export economy in Colombia, 1850–1925. J Econ Hist 70(2):378–399
Schonhardt-Bailey C (2006) From the corn laws to free trade. Interests, ideas and institutions in historical perspective. MIT Press, Cambridge, MA
Schularick M, Solomou S (2011) Tariffs and economic growth in the first era of globalization. J Econ Growth 16(1):33–70
Schulze M-S, Wolf N (2009) On the origins of border effects: insights from the Habsburg empire. J Econ Geogr 9(1):117–136
Schulze M-S, Wolf N (2012) Economic nationalism and economic integration: the Austro-Hungarian empire in the late nineteenth century. Econ Hist Rev 62(2):652–673
Schulze MS, Heinemeyer HC, Wolf N (2008) Endogenous borders? Exploring a natural experiment on border effects. Center for Economic Policy Research working paper 6909
Schulze M-S, Heinemeyer HC, Wolf N (2011) On the economic consequences of the peace: trade and borders after Versailles. J Econ Hist 71(4):915–949
Serrano R, Pinilla V (2011) The evolution and changing geographical structure of world Agri-food trade, 1951–2000. Rev Hist Ind 46(2):97–125
Sharp P (2010) "1846 and all that": the rise and fall of British wheat protection in the nineteenth century. Agric Hist Rev 58(1):79–94
Sharp P, Weisdorf J (2013) Globalization revisited: market integration and the wheat trade between North America and Britain from the eighteenth century. Explor Econ Hist 50:88–98
Shiue CH, Keller W (2007) Markets in China and Europe on the eve of the industrial revolution. Am Econ Rev 97:1189–1216
Singer H (1950) The distributions of gains between investing and borrowing countries. Am Econ Rev Paper Proc 40:473–485
Smith A (1776) An inquiry into the nature and causes of the wealth of nations. W. Strahan and T. Cadell, London
Solar PM, Hens L (2015) Ship speeds during the Industrial Revolution: East India Company ships, 1770–1828. Eur Rev Econ Hist 20:66–78
Steinwender C (2018) The real effects of information frictions: when the states and the Kingdom became United. Am Econ Rev 108(3):657–696
Swinnen JFM (2009) The growth of agricultural protectionism in Europe in the 19th and 20th centuries. World Econ 32(11):1499–1537
Taylor AM (1998) Peopling the Pampa: on the impact of mass migration to the river plate, 1870–1914. Explor Econ Hist 34:100–132
Taylor AM (2001) Potential pitfalls for the purchasing-power-parity puzzle? Sampling and specification biases in mean-reversion tests of the law of one price. Econometrica 69:473–498
Taylor JE, Basu B, McLean S (2011) Net exports and the avoidance of high unemployment during reconversion, 1945–1947. J Econ Hist 71(2):444–454
Tena-Junguito A (1989) On the accuracy of foreign trade statistics: Italy 1890–1938. Rivista di storia economica 6(1):87–112
Tena-Junguito A (1995) Una reconstrucción del comercio exterior español, 1914–1935: La rectificación de las estadísticas oficiales. Rev Hist Econ 3(1):77–119
Tena-Junguito A (2006a) Assessing the protectionist intensity of tariffs in nineteenth-century European trade policy. In: Dormois JP, Lains P (eds) Classical trade protectionism, 1815–1914. Routledge, London, pp 99–120
Tena-Junguito A (2006b) Por qué fue España un país con alta protección industrial? Evidencias desde la protección efectiva 1870–1930. In: Dobado R, Gómez Galvarriato A, Márquez G (eds) España y México. ¿Historias económicas paralelas? Fondo de Cultura Económica, México
Tena-Junguito A (2010a) Bairoch revisited: tariff structure and growth in the late nineteenth century. Eur Rev Econ Hist 14:111–143
Tena-Junguito A (2010b) Tariff history lessons from the European periphery. Protection intensity and the infant industry argument in Spain and Italy 1870–1930. Hist Soc Res 35(1):340–362

Tena-Junguito A, Willebald H (2013) On the accuracy of export growth in Argentina, 1870–1913. Econ Hist Dev Region 28(1):28–68
Tena-Junguito A, Lampe M, Tena-J Fernandez F (2012) How much trade liberalization was there in the world before and after Cobden-Chevalier? J Econ Hist 72(3):708–740
Thomson JKJ (2008) The Spanish trade in American cotton: Atlantic synergies in the age of enlightenment. Rev Hist Econ 26(2):277–314
Timini J (2018) Currency unions and heterogeneous trade effects: the case of the Latin Monetary Union. Eur Rev Econ Hist. https://doi.org/10.1093/ereh/hex027
Tirado DA, Pons J, Paluzie E, Martínez-Galarraga J (2013) Trade policy and wage gradients: evidence from a protectionist turn. Cliometrica 7:295–318
Trenkler C, Wolf N (2005) Economic integration across borders: the polish interwar economy 1921–1937. Eur Rev Econ Hist 9(2):199–231
Uebele M (2011) National and international market integration in the 19th century: evidence from comovement. Explor Econ Hist 48:226–242
Vamvakidis A (2002) How robust is the growth-openness connection? Historical evidence. J Econ Growth 7:57–80
Van Dijck M, Truyts T (2011) Ideas, interests, and politics in the case of the Belgian corn law repeal, 1834–1873. J Econ Hist 71(1):185–210
Varian BD (2016) The revealed comparative advantage of late-Victorian Britain. EHES working paper no. 97
Vasta M (2010) Italian export capacity in the long run perspective (1861–2009): a tortuous path to keep the position. J Modern Italian Stud 15(1):133–156
Vizcarra C (2009) Guano, credible commitments, and sovereign debt in nineteenth-century Peru. J Econ Hist 69(2):358–387
Williamson JG (1990a) The impact of the corn laws just prior to repeal. Explor Econ Hist 27(2):123–156
Williamson J (1990b) Latin American adjustment: how much has happened? Peterson Institute for International Economics, Washington, DC
Williamson JG (2006) Explaining world tariffs, 1870–1938: Stolper-Samuelson, strategic tariffs, and state revenues. In: Findlay R, Henriksson RGH, Lindgren H, Lundahl M (eds) Eli Heckscher, international trade, and economic history. MIT Press, Cambridge, MA, pp 199–228
Williamson JG (2008) Globalization and the great divergence: terms of trade booms, volatility and the poor periphery, 1782–1913. Eur Rev Econ Hist 12:355–391
Williamson JG (2011) Trade and poverty. When the third world fell behind. MIT Press, Cambridge, MA
Wolf N (2005) Path dependent border effects: the case of Poland's reunification (1918–1939). Explor Econ Hist 42(3):414–438
Wolf N (2009) Was Germany ever united? Evidence from intra- and international trade, 1885–1933. J Econ Hist 69(3):846–881
Ye L (2010) U.S. trade policy and the Pacific Rim, from Fordney-McCumber to the Trade Expansion Act of 1962: a political-economic analysis. Res Econ Hist 27:201–253
Yousef TM (2000) The political economy of interwar Egyptian cotton policy. Explor Econ Hist 37:301–325
Zahedieh N (2013) Colonies, copper, and the market for inventive activity in England and Wales, 1680–1730. Econ Hist Rev 66(3):805–825

Market Integration

Giovanni Federico

Contents

Abstract

This chapter outlines the main advances in the literature on market integration in the last 5 to 10 years. After a short review of the definition of market integration and of the main statistical tools to measure it, each of the three following section addresses one major issue. Section "The First Wave: Measurement" deals with the measurement of integration, updating and expanding the earlier review of Federico (2012) to integration in non-European countries and to transoceanic integration. Section "The Second Wave: The Causes of Integration" surveys the literature on causes of integration, arguing that changes depended mostly on political decisions about barriers to trade, with a substantial contribution of technical progress in the nineteenth and twentieth century. Finally, Section "The Third Wave: The Effects of Integration" deals with the effects of integration. Most papers quote the benefits from integration in the motivation for the choice of the issue, but there are very few and very partial attempts to measure them.

G. Federico (✉)
Department of Economy and Management, University of Pisa, Pisa, Italy

CEPR, London, UK
e-mail: Giovanni.Federico@unipi.it

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_68

Section "Conclusion: Taking Stock" concludes with a short summary of the main findings, stressing the relevance of recent work to integrate spatial and intertemporal arbitrage.

Keywords
Market integration · Market efficiency · Smithian economic growth · Gains from trade

Introduction

Economists at least since Adam Smith have argued that markets and trade are essential for the smooth functioning and development of economic systems, and thus it is not surprising that economic historians have shown a keen interest in their development. A comprehensive research program should address three main questions: (1) the measurement of the process (when and by how much did market develop?), (2) its causes (why did markets develop?), and (3) its effects (how did market development affect welfare and growth?). It is almost impossible to address these issues in historical perspective with data on trade. Reliable series of international trade are available only since the nineteenth century (Federico and Tena 2016), and data on domestic trade flows are extremely scarce before the late twentieth century, an exception being Germany (Wolf 2009). Thus, economic historians have turned to prices. Series of prices, especially of cereals, are available in Europe since the Middle Ages, and their collection had started in the 1930s by the International Committee for the history of prices. Asian sources, once tapped systematically, have proven to be as rich as European ones, at least from the eighteenth century onward. This chapter will survey the price-based literature on market integration, leaving to other chapters the parallel quantity-based literature on growth of trade.

Some pioneering scholars started to use prices to study market integration as early as the 1960s, but the issue leapt forward as a hot topic in economic history in the late 1990s. The first wave of research focused on levels and trends of integration within Europe and the Atlantic economy at large (i.e., question 1 above). A survey (Federico 2012) quotes 61 works published before December 2009. Since then, researchers have continued to pursue this line of research, but recently they have begun to deal with the causes and, to a much lesser extent, the effects of integration – i.e., questions 2 and 3. In spite of their efforts, we are still far from having a comprehensive view of integration in the long run. Some works deal jointly with measurement and causes, and very few with measurement and effects, but remarkably, only Bateman (2012) in her work on early modern Europe addresses all three questions with statistical tools.

This chapter will deal with the quantitative works and not on the literature on market institutions and/or the causes of their change, such as Epstein (2000) and Van Bavel (2016). These works provide important insights to interpret quantitative results, but their discussion would exceed the remit of this survey. Furthermore, the chapter makes three clear and somewhat painful choices in the interest of page

constraints. First, it focuses on the integration of markets for commodities, neglecting integration of factor markets. Wage convergence has been studied, although not so extensively, in the literature on migrations, while the integration of capital markets in early modern Europe has attracted a lot of interest in very recent years (Volckart and Wolf 2006; Chilosi and Volckart 2011; Li 2017; Chilosi et al. 2018). Second, the chapter relies as much as possible on Federico (2012) for the literature on measurement of integration in Europe, updating it when necessary. In contrast, it will survey in more detail the measurement of integration on other continents (most notably Asia) and across continents and, above all, the literature on the causes (question 2) and effects (question 3) of integration. When appropriate, we will consider also works from the parallel literature on gains from trade (Donaldson 2015). Last but not least, unlike Federico (2012), this chapter does not aim at being systematic. Rather, it will focus on papers that add novel methodological insights or bring major new results.

The next section deals with the theory of market integration. It reviews the basic concepts of integration and sketches out the main statistical methods, integrating Federico (2012), with a discussion of the recent methodological developments. Each of the three following section addresses one of the issues – the measurement of integration (Section "The First Wave: Measurement"), the analysis of causes (Section "The Second Wave: The Causes of Integration"), and the estimation of its effects (Section "The Third Wave: The Effects of Integration"). Section "Conclusion: Taking Stock" concludes with a short summary of the main findings and some general remarks on the future agenda for research.

The General Framework: What Is Market Integration and why it Is Relevant

The definition of market integration was first put forward almost 200 years ago by French mathematician Cournot. He characterizes an integrated market as "an entire territory of which the parts are so united by the relations of unrestricted commerce that prices take the *same* level throughout with *ease and rapidity*" [emphasis added] (Cournot 1971: 51–52). In other words, (i) the equilibrium level of prices must be equal (the law of one price) and (ii) prices must return easily and quickly to this level after any shock. Those two conditions are clearly separable, as prices can return slowly to the same level or quickly to a different level. Both conditions are necessary but not sufficient for a fully integrated market, and each of them must be tested with a different set of statistical tools. Unfortunately, as we will discuss in section "The First Wave: Measurement," this good practice is not always followed. Many authors focus on one condition only and label their own results a test of market integration. Thus, before delving into historical results, it is necessary to briefly review the key features of the available statistical tests in relation to the basic theory of market efficiency (see for more details and additional references Federico (2012)).

Testing the law of one price is apparently simple: what can be easier than comparing prices of the same good in two different locations at the same moment?

Alas, the results are likely to be seriously misleading. If the two locations trade, the law would almost always be violated, even for perfectly homogeneous goods. The price differential in any given time would be at least equal to transaction costs (or commodity points) and could exceed them in case of temporary disequilibria while reabsorbing location-specific shocks. On the other hand, if the two locations do not trade, the difference may move randomly between the commodity points or be determined by trading with a third market (Coleman 2007; Federico 2012). One might argue that the law should not be taken literally and define as integrated market if violations are small or, as suggested by some authors (e.g., Stigler and Sherwin (1985), Ejrnæs et al. (2008)), if the price gap is equal to transaction costs. Both solutions are not appealing. Any threshold for distinguishing small violations from substantial ones would be unavoidably arbitrary, while computing all transaction costs and the implicit risks is a challenging task to say the least, as shown by the excruciatingly careful estimate of gold points in the nineteenth-century transatlantic currency markets by Officer (1996). Even if this computation were possible, the method would label as (equally) well-integrated markets with hugely different price gaps (say 1 percent or 100 percent or even 1000 percent).

The obvious alternative to a static comparison is a dynamic approach, focusing on trends rather than on levels. A market is integrating (rather than integrated) if the price gap between two locations shrinks (price convergence) and vice versa. However, the interpretation of results can become cumbersome if the number of locations, and thus of possible pairs, increases above a (small) figure. Some authors have suggested selecting some representative pairs of locations or focusing on the convergence of prices of a number of peripheral markets toward a central one (Ravaillon 1987). The easiest solution, which has become a standard in the literature, is to measure σ-convergence by computing the coefficient of variation across all locations in each year. The results are likely to be imprecise if quality differs, but not necessarily systematically biased (Federico 2011). The coefficient of variation, as a dimensionless measure, is easily comparable in time and space, and the sequence of coefficients can be analyzed with time-series analysis looking for discontinuities and estimating long-run trend. The cumulated change as a percentage of the initial dispersion is a simple measure of the extent of integration (or lack of it), while detected discontinuities could be related to specific events, such as changes in trade policy or in transport infrastructure.

Cournot's second condition can be restated in modern terms as a test for market efficiency. As it is well known, a market is defined as efficient if prices take into account all available information and thus there are no opportunities for profitable arbitrage in space or time. The price gap between two trading locations must be equal to transaction costs and the difference between current and future (expected) price to the storage costs. The standard framework in the market integration literature assumes that agents know the local conditions and prices of the commodity in other locations and set the prices accordingly. If a shock causes price differentials to exceed from their equilibrium level, arbitrage by profit-maximizing traders would push the gap back to the equilibrium level. The faster this process, the more efficient is the market. The early works tested Cournot's second condition indirectly by

looking at co-movement of prices: the faster the adjustment to an exogenous shock, the closer that prices would move together. The workhorse of these early "integration" (or more precisely efficiency) tests was the pairwise coefficient of correlation between prices in different locations. Some authors have preferred to run an OLS regression, possibly adding distributed lags to capture the dynamic component of the adjustment (see, for a recent example, Panza (2013)). This latter approach introduces the additional hypothesis of a hierarchy between two markets, the central one where prices were set and the other, to which prices were transmitted via arbitrage à la Ravaillon (1987).

The co-movement tests share two shortcomings. First, by construction, they have to be estimated with a minimum number of observations and yield a single coefficient. Implicitly, they thus assume transaction costs to be fixed in the period, in contrast with the whole research agenda on market integration. No matter how carefully the author could choose the period of estimation, there is risk that changes in transaction costs return an invalid estimate. The shorter the period and thus the higher the frequency of the data are, the smaller this risk is. Unfortunately, in very many cases, authors are forced due to the availability of sources to use low-frequency data. Second, the results will be biased upward if the series feature a common trend or are subject to the same shocks (e.g., from weather). Scholars used different strategies to tackle this problem, such as computing correlations of first differences (e.g., Li (2000), Studer (2008)) or detrending price series (e.g., Dobado et al. (2012), Chilosi et al. (2013)). Furthermore, the interpretation of the results needs a somewhat arbitrary decision on the minimum coefficient for an integrated market. Is a coefficient of correlation of 0.6 enough? In a nutshell, how big is big?

Since the 1990s, this traditional indirect approach to testing efficiency via co-movements has been sidelined by advances in time-series econometrics. The co-integration test offers a simple answer to the general question whether a market was integrated. It is possible to compute the key parameter for Cournot's second condition (the speed of adjustments) by running a simple AR (1) model, such as

$$\Delta \left(P_{it} - P_{jt}\right) = \rho \left(P_{it-1} - P_{jt-1}\right) + \varepsilon_t \quad (1)$$

The speed of adjustment is conventionally defined as the half-life of a shock, or $\lambda = [\ln 0.5/\ln (1+ \rho)]$. This parameter, as the coefficient of correlation, can easily be compared across time and space, although there is not an accepted threshold for distinguishing a fast from a slow adjustment.

There are many variants to eq. 1. It is common practice to add dummies for specific shocks (e.g., wars), while Bernhofen et al. (2016) add changes in the average price across locations as a measure of common shocks. It is also possible to explore the pattern of diffusion of shocks and hierarchies of markets with a vector error correction model or VECM (Lampe and Sharp 2015). On the other hand, the AR framework has a quite serious flaw as a test of efficiency. It implicitly assumes that prices would converge even if the price gap is smaller than the transaction costs, when arbitrage would be unprofitable. This logical shortcoming can be addressed by

running so-called threshold auto-regression models (or TAR). They assume that prices converge only when their gaps exceed the commodity points and that the process stops when the gap is equal to transaction costs. Within commodity points, prices move randomly. On top of this, the TAR software estimates the most likely commodity points with a grid-search procedure, which can be used in further analysis.

The co-integration revolution has marked a great leap forward in the analysis of efficiency, but it is not trouble-free. First, as the co-movement measures, the co-integration tests need a minimum number of observations, which correspond to a period of time inversely related to the frequency of data (240 observations correspond to 20 years with monthly data, to four and half years with weekly data, and to less than 1 year with daily data). They return one coefficient which would be unbiased only if transaction costs do not change in the period of observation. Second, the estimated speed depends on the frequency of the data (Brunt and Cannon 2014). Results would be biased upward if the frequency of data is lower than the actual speed of adjustment – e.g., if data are monthly and adjustment takes 1 week (Taylor 2001). They would differ according to the nature of the data – i.e., whether they refer to a specific point in time, with its specific shock, or rather are obtaining averaging prices, thereby smoothing shocks. Last but not least, the TAR assumes that its estimate of commodity points coincides with the actual transaction costs. This assumption would not hold if price gaps are smaller than the commodity points because the two locations do not trade, but also if they systematically exceed the commodity points, because the market is not (weakly) efficient. This case is not so implausible. Traders might rationally decide to overlook opportunities for arbitrage if they perceive that risks (e.g., from fluctuations in exchange rates) outweigh the potential gains from arbitrage. In both cases, the estimates from the TAR would be biased, respectively, downward or upward, relative to actual transaction costs. Furthermore, one should note that the whole co-integration approach implicitly rules out the possibility that agents in a location have information on the conditions in other locations other than the prices (i.e., that the markets were "semi-strongly" efficient). This restriction is somewhat implausible but relaxing; it implies a quite different process of price setting. Traders are likely to react similarly to the same information on events which can affect prices (a war, a failure in harvests), and thus price gaps would not change. In a perfectly efficient market with perfect information for a homogeneous good, price differentials between markets would remain constant as long as transaction costs remain stable.

A parallel strain in the literature has measured integration with the variance of prices for a single market or, as suggested by Engel and Rogers (1996), the ratio of prices in two locations. This approach assumes that integration could reduce price volatility relative to a closed economy, where prices are inversely related to domestic production, and fluctuations are larger, the less elastic is consumption. This assumption is supported by the results by Clark (2015), for England in the Middle Ages, and Studer (2015) for Switzerland in the early modern period. They run a very simple model of price determination as function of local supply conditions (as measured by yields or weather fluctuations) and find that adding

nationwide prices, or prices in other markets, reduced the coefficient of the local supply variable. These results, however, may not be sufficient evidence for the variance-based measures of efficiency. They refer to very specific cases, and in general price volatility depends on many factors, most notably the elasticity of substitution with other goods, the covariance with their prices, and the extent of storage or intertemporal arbitrage (Foldvari and van Leuuwen 2011). Furthermore, any interpretation of decline in variance as evidence of increasing efficiency assumes that the size of underlying shocks is constant. This is by no means sure (see Brunt and Cannon (2014) for an example). For instance, irrigation or other modern technology may reduce the effect of weather on crop yield and thus the size of weather-related price fluctuations.

The recent literature has suggested a number of important, although not earth-shattering, improvements to these traditional measures of efficiency. Dobado et al. (2012, 2015) suggest the use of variance residuals from AR(1) models of price ratios, rather than the variance of the series, to highlight shocks. On the same line, Brunt and Cannon (2014) distinguish the transitory from the permanent component of price movements with a two-market VECM model and decompose the former into location-specific shocks, common shocks, and adjustment. Others have introduced the possibility of changes in parameters, either by using time-variant specifications of the baseline AR model (Craig and Holt 2017) or a rolling stock approach (e.g., Hynes et al. (2012), Bernhofen et al. (2016), Federico et al. (2018)). The most relevant addition to the statistical toolkit is the dynamic factor analysis (Uebele 2011; Anderson and Ljunberg 2015; Federico et al. 2018). It can be interpreted as a sophisticated version of the traditional co-movement measure. Instead of computing pairwise correlations between price series and averaging them somehow, the procedure extracts one single series, or common factor, which captures all common movements. The ratio of its variance to the overall variance of location-specific series measures the extent of co-movements and thus the efficiency of the market. Furthermore, the common factor can be used as a proxy for reference prices for the area, as an endogenous alternative to the a priori selection of a central market à la Ravaillon (1987),

The most relevant methodological improvement is arguably the integration of intertemporal arbitrage in the modeling of spatial arbitrage. By definition, intertemporal arbitrage is possible only if commodities can be stored for future consumption. The technology for long-term conservation of perishables, such as dairy products, is quite recent, but storage is as old as agriculture, as cereals and some other annual crops have to be stored for consumption throughout the year. In theory spatial and intertemporal arbitrage are substitutes: agents can decide to withhold products from the market if they expect future prices to exceed current ones by more than storage costs. In this case, storage would reduce market supply, causing the current prices to return to their intertemporal equilibrium level. If the market is efficient, expected prices would change only if new information is available. Steinwender (2018) introduces storage in her analysis on the effects of the telegraph on transatlantic cotton trade, as a fluctuation-smoothing mechanism. In her model, cotton was stored only in the destination market (the United Kingdom), and

sales from stocks met additional demand, dampening price increases until the arrival of additional supply from the United States. The latter depended on the combined time for transmission of information about excess demand from the United Kingdom to the United States and for actual shipping of cotton. The telegraph reduced the time of transmission and thus the need for storage. In another very recent paper, Craig and Holt (2017) attribute the decrease in the speed of adjustment of egg prices in the United States in the late nineteenth century to the introduction of refrigeration (cold storage). Without storage, traders had to sell immediately, irrespective of their expectations about future prices. Refrigeration made it possible for traders to withdraw the product from the market if they expected prices to rise above storage costs and thus reduced the local supply for arbitrage.

A different and more radical approach was first suggested by Shiue (2002) and then formalized by Coleman (2009). Unlike Steinwender, he assumes perfect information on prices and focuses on the difference between equilibrium conditions with and without storage. In his model, prices in origin markets (Chicago in his example) at any given time are equal to the expected prices in destination markets (New York) at the time of arrival, t + n. These latter are equal to prices in New York at time t plus storage costs. In equilibrium, Chicago prices are equal to New York spot prices less transport costs (the standard condition for spatial arbitrage) plus storage costs – i.e., the gap in prices is equal to transaction costs less storage costs. Therefore, the spatial-cum-intertemporal commodity points are narrower than spatial-only points. In this case, spatial arbitrage would be unprofitable as long as the demand in the destination market could be met by reducing inventories. The information that inventories are going to run down at the time of arrival causes prices in the destination market to rise, and thus the price gap to exceed transaction costs, triggering arbitrage. Thus, in theory, the model implies regular spikes in price gaps, which could be detected with high-frequency data. In practice, if the period n is short enough, the difference between commodity points with and without storage is likely to be small, and thus the results of the spatial plus intertemporal arbitrage model will not differ much from those of the standard spatial-only framework.

The First Wave: Measurement

Federico (2012) summed up results of almost 40 years of research on the integration of European markets by pointing out the difference between the early modern period and the "long" nineteenth century, from Waterloo to World War One. There was a wide consensus on a massive integration process along Cournot dimensions, price convergence, and growing market efficiency in the latter period and a substantial disagreement on the movements before the French Revolution. Recent work has confirmed the view of the nineteenth century as a period of integration [integration, nineteenth century] and has shed some light on previous trends – or lack thereof. First, Chilosi et al. (2013) have explored the geographical pattern of integration in

Europe from 1620 to 1913, by allocating their sample of about 100 cities in regions (from 7 to 11 according to periods) with a principal component analysis. The size of regions and the pattern of price convergence depended mostly on geography. The area around the North Sea, which benefitted from easy and cheap sea communications, was already quite well integrated at the beginning of the period, while landlocked regions were smaller and less integrated. In the long run, also the landlocked areas joined the European market, with the exception of Spain, which remained isolated well into the nineteenth century. Federico et al. (2018) have expanded the period back in time to the early fourteenth century and the size of the database to almost 600 locations, with up to 300 prices in the mid-eighteenth century. Unfortunately, most of these series are quite short, thus authors rely for the very-long-run analysis on a core sample of 15 cities, mostly from North-Western Europe, with some cities in Southern France and Northern Italy. In the early modern period [integration early modern period], price dispersion for the core sample fluctuated widely: it was fairly low in the fifteenth and sixteenth century, jumped in the early seventeenth century, declined again from about 1650 to the French Revolution (the period covered by Chilosi et al. (2013)), and soared during the Napoleonic wars. A dynamic factor analysis shows that efficiency varied much less than price dispersion in the early modern period and started to rise in the late eighteenth century. After Waterloo, prices started to converge and efficiency continued to grow, so that around 1870 the European market for wheat achieved an unprecedented level of integration along both dimensions. Federico et al. (2018) supplement this analysis by building larger samples for 80-year periods, selecting cities from their database with a grid procedure to be as geographically representative of the whole continent as possible. These additional data confirm the result from the 15 cities sample but shows that convergence outside the core areas was slower and less complete. Last but not least, Federico et al. (2018) show that, taking distance into account, the level of integration in England and the Netherlands was not exceptionally high in the sixteenth and early seventeenth century and soared afterward. These two papers provide a broadly consistent story on European integration in the long run. Unfortunately, the story is quite complex, with substantial differences between areas and periods and between trends in price dispersion and efficiency. These differences can explain how the earlier literature, using different statistical tools on small samples of differently located cities for different periods, had reached quite divergent conclusions.

The conventional wisdom about intercontinental integration has been deeply shaped by the research by K. O'Rourke and J. Williamson. They have argued, in different papers, that:

1. The opening of a new route to Asia at the beginning of the sixteenth century caused a small decline in real prices of spices in Europe and a modest increase in correlation among European markets (O'Rourke Williamson 2009).
2. In the seventeenth and eighteenth centuries, price wedges between Europe and Southeast Asia remained stubbornly large, well above the transportation costs, even factoring for the risks of the trip (O'Rourke and Williamson 2002).

3. The Atlantic economy integrated during the nineteenth century, with far-reaching consequences on facto incomes and distribution (O'Rourke and Williamson 1994).
4. The interwar years featured mixed trends, with some divergence in prices, but no major loss of efficiency (Hynes et al. 2012).

The recent works suggest a more sanguine view of the intercontinental integration in the early modern period. Dobado et al. (2012) and Sharp and Weisdorf (2013) show that the Atlantic wheat market was already quite efficient in the eighteenth century. Both Ronnback (2009) and de Zwart (2016) criticize the conclusions by O'Rourke and Williamson as based on a small number of series for a limited period. They find that levels of dispersion and trends in convergence differed substantially among routes and between products along the same route. In particular, de Zwart (2016) shows that price gaps between Europe and Southeast Asia started to decline for some goods (e.g., cloves) already in the seventeenth century, and that convergence extended to the majority of commodities, with exceptions such as mace and nutmeg, in the eighteenth century. In the early 1810s, the markups on Asian products in Europe were still quite large, from 70% to 220%, and substantially greater than the price gaps for American products (Chilosi and Federico 2015). The convergence of the first half of the century was correspondingly faster and deeper in the Indian Ocean than across the Atlantic.

Most of the early literature on market integration deals with Europe, but there were some pioneering studies of the integration of Asian markets in the 1970s and 1980s (Hurd 1975; Latham and Neal 1983; Brandt 1985; Bessler 1990). The number of such studies has substantially increased in the last year. Dobado et al. (2015) argue, with their preferred variance-based measure, that domestic rice markets in Japan were quite efficient already in the eighteenth century. In contrast, the admittedly much bigger Indian market showed neither price convergence nor increase in efficiency before the second half of the nineteenth century (Studer 2015). The integration of the Chinese domestic market has attracted much attention as part of a wider debate on the economic conditions of China in the eighteenth century. The so-called California school argued that China was as advanced as Western Europe in the eighteenth century according to almost all indicators, including the development of markets, and diverged only in the nineteenth century (Pomeranz 2000). This sanguine view did not tally well with the disintegration of the market for wheat and millet in the Northern province of Hebei, which included Beijing, from 1738 to 1911 (Li 2000). On the other hand, Pomeranz's optimism was buttressed by a very influential paper by Shiue and Keller (2007), who compared China and Europe in the eighteenth century according to three different static measures of efficiency (volatility of prices, average coefficient of correlation, and t-statistics of stationarity of residuals of a pairwise regression of prices). They argued that China as a whole was no less efficient than Western Europe, although China's most advanced area, the Lower Yangtze valley, was less integrated than England. China and Europe diverged in the nineteenth century when massive integration in Europe was not matched by a parallel process in the Heavenly Empire. Further work has given mixed results,

possibly because of the use of different measures of efficiency. The residual-based variance measure by Dobado et al. (2015) shows that the market for rice in three locations in "advanced" China was as efficient as the English market and that efficiency was growing until the 1840s. In contrast, Bernhofen et al. (2016), with an augmented AR model, confirmed the results by Li (2000) for the markets for wheat in North China (80 cities) and rice in South China (131 cities). The half-lives of price shocks were growing (i.e., the market was becoming less efficient) already toward the end of the eighteenth century, well before the mid-nineteenth-century crisis, and, in stark contrast with Shiue and Keller (2007), they were on average about six times longer than in the (admittedly smaller) European "national" markets for series of comparable frequency.

Integrating temporal and spatial arbitrage [intertemporal arbitrage], although theoretically sound, has proven to be empirically challenging due to the lack of data. The economic history literature has largely ignored Coleman's (2009) plea for storage-adjusted commodity points, although some authors have tried to assess the relevance of storage from price data. Shiue (2002) has argued that the resort to storage can be inferred from the impact of a weather variable (an index of low rain) on local prices. The coefficient is positive and significant but declining in time. Thus, she concludes that in the eighteenth century China's intertemporal smoothing via storage was imperfect but improving. This inference is not robust, however, as her specification does not control for the effect of spatial integration on prices by including prices in other locations. She only points out that the effect of weather was stronger in inland areas, which were less spatially integrated than the coastal ones. Clark (2015) argues that in Medieval England storage was widespread because prices were correlated across years (while output fluctuated randomly), and intertemporal arbitrage was efficient because the increase of prices over the crop year was aligned to storage costs.

The Second Wave: The Causes of Integration

It would be unfair to state that economic historians have been so obsessed by measuring integration as to neglect its causes altogether. Almost all of them discuss the issue, but the majority infer the causes of integration combining the results of measurement and anecdotal evidence on changes in trade costs, without any formal test. All authors explain the nineteenth-century price convergence with the joint effect of liberalization of trade and technical innovation in transportation: the steamship and railways. Some authors have tried to support their conclusions about the likely causes of integration with some simple statistical testing. Quite a few of them have grouped (pairs of) markets according to distance to show how transportation costs affected domestic integration. For instance, Studer (2015) finds that in India before 1830 the average coefficient of correlation among first differences of wheat prices was very high for close-by markets (less than 35 kilometers apart), but it dropped to almost zero beyond 70 km. After 1860, the correlation was 0.46 for pairs over 1500 km. – i.e., as high as for the 35–70 km bracket in the first

period but still well below the correspondent figures for Western Europe in the same years. This approach requires the setting of thresholds to classify pairs of markets, and is suitable only for analyzing domestic integration, which depends only on transport costs. Federico and Persson (2007) suggest a way to disentangle the domestic from the international components of change in price dispersion with a variance analysis [causes of integration, variance analysis approach]. The two components accounted for almost the same share of long-run convergence of wheat prices in Europe from the early 1750s to 1870, but all short-term changes were driven by the international component, which clearly reflects political shocks, such as the outbreak of the Napoleonic wars or the abolition of British Corn Laws (Federico 2011). Last but not least, Dobado and Marrero (2005) explore the effects of railways on domestic integration in the nineteenth-century Mexico by comparing the speed of adjustment before and after the construction.

All these computations, while clearly an improvement over inferences from visual inspection, still need a selection of relevant causes and/or thresholds. Thus, in recent times, scholars have moved to formal testing of the effects of changes in trade cost using a panel regression approach [causes of integration, panel regression approach], broadly inspired by gravity models in international trade. In their classical paper on the issue, Anderson and van Wincoop (2004, pp. 691ff) define trade costs as "all costs incurred in getting a good to a final user other than the marginal cost of producing the good itself: transportation costs (both freight costs and time costs), policy barriers (tariffs and non-tariff barriers), information costs, contract enforcement costs, costs associated with the use of different currencies, legal and regulatory costs, and local distribution costs (wholesale and retail)." This definitions implies a general model such as

$$\mathrm{MI} = \mathrm{f}(\mathrm{Tc},\mathrm{B},\mathrm{I},\mathbf{E},\mathbf{X}) \tag{2}$$

where MI can refer either to price convergence or efficiency; Tc measures transport costs, B barriers to trade, I information flows, and **X** all other trade costs; and **E** is a set of dummies to capture unusually large shocks (e.g., wars). In this setting, the expected sign of explicative variables varies according to the dependent variable: a positive sign corresponds to less integration if the dependent variable is a measure of dispersion (e.g., price gaps between two markets) and to more integration if the dependent variable is a measure of efficiency (e.g., the speed of adjustment or the coefficient of correlation). Furthermore, results may differ to the extent that the same causes could affect the two Cournot conditions (equilibrium price gaps and the speed of adjustment) differently.

By far the most common measure of convergence in formal testing is the ratio of prices between pairs of markets. Federico (2007) uses also the coefficient of variation as the dependent variable, but the main alternatives are the estimates of transaction costs as obtained from a TAR model. As noted earlier, each TAR estimate yields a single set of points, but one can get several sets by running the model for overlapping periods (Jacks 2006). In contrast, there is a wide range of pairwise measures of efficiency, such as the coefficient of correlation (Shiue 2002), the

variance of relative prices (Engel and Rogers 1996), and of course the speed of adjustment (Jacks 2006). Brunt and Cannon (2014), in the most detailed analysis so far, uses as dependent variable(s) the standard deviation of gaps between prices of pairs of neighboring markets (i.e., the Engel-Rogers measure) and four different measures of location-specific shocks (speed of adjustment, size, pairwise ratio, or correlation between shocks in two locations).

Measuring trade costs has proved to be challenging. Only a few authors have been able to find series of actual transport costs or duties, and none has actual data on information costs. Nevertheless, economic historians have been quite imaginative in substituting them with proxies, with varying success.

The standard measure for transport costs (Tc) in gravity trade models is birds' fly distance, and indeed it is quite common also in the literature on market integration. Buyst et al. (2006) measure the effect of transport costs in the Southern Low countries in the eighteenth century with distance, distance squared, and dummies for access to water (rivers or sea). Distance affected the commodity points significantly, and not linearly, but not the speed of adjustment. By their nature, time-invariant measures such as distance and location dummies can explain the level of integration, but not their changes in time. Indeed, most authors try to capture changes in infrastructure endowment with dummies for the existence of a rail or road connection between two cities or, more generally, for the access to the network. The results confirm the expectations about the positive effects of infrastructure on integration, but with a lot of variations that offer important insights. For instance, Buyst et al. (2006) find that the existence of a paved road at the beginning of the period of estimation (1765) made adjustment significantly quicker, in all likelihood by improving the circulation of information, but it did not affect the commodity points. This latter result might reflect the omission of roads built after 1765, itself a consequence of structure of estimation (one single coefficient, no panel), or simply the existence of tolls on paved roads. Brunt and Cannon (2014) find that in the late eighteenth-century England, better transportation, as proxied by the density of paved road and dummies for canals, reduced the volatility of prices and the size of shocks and increased their correlation, but it had mixed effects on the speed of adjustment**.** Paved (free) roads reduced the gaps in rye prices in Westphalia in the central decades of the nineteenth century, while the effect of waterways was mixed, and railroads had little impact, possibly because the network was not yet fully developed (Uebele and Gallardo-Albarran 2015). Railroads have a very special position in this literature as the main factor of integration of domestic markets. Keller and Shiue (2008) estimate that building railways accounted for about four fifths of price convergence in Germany in the nineteenth century. The case of Italy was somewhat different (Federico 2007): the building of railways substantially reduced the price gaps between landlocked markets in the North, but had little impact for the rest of the country, where rail transportation was outcompeted by coastal trade. For the whole continent, Jacks (2006) confirms that the mere existence of a railway connection between two locations reduced price gaps by 5%, while the effects on efficiency are mixed: the speed of adjustment is positively related to the existence of a connection but negatively to its length. The effect of railways on integration of Indian markets

has drawn special attention within the debate about the economic benefits of British rule. Hurd (1975) argued that railways were the main driver on integration by comparing trends on coefficients of variation for counties with and without railroads and Studer (2015) concurs, although without a formal test. This conventional view has been questioned by Andrabi and Kuelhwein (2010), who find that convergence in cereal prices started before railway building and estimate that railways accounted for only a fifth of the overall decline in price gaps. They infer that most of the convergence can be explained by other integrating forces, such as a common currency, language, and administration, but that one could put forward an alternative interpretation. They use retail prices of cereals; thus the (likely) stable markup on wholesale prices may have dampened the effect of a decline of trade costs on wholesale prices. Indeed, the interpretation seems inconsistent with the recent estimates by Donaldson (2018) on the costs of transporting salt from a common source to the whole of India by different means. He estimates that the elasticity of trade costs to distance was around 0.25 and that both road and, somewhat surprisingly, river were more expensive than rail transport, by eight and slightly less than four times, respectively.

The econometric analyses have largely confirmed the relevance of changes in barriers to trade to foster or hamper the integration of commodity markets. The abolition of Piedmontese [duties] duty, the only relevant one, in 1849 accounted for about a fifth of the price convergence among Italian states before the unification of the country in 1860 (Federico 2007). Keller and Shiue (2014) estimate that the entry of a state into the Zollverein reduced the wheat price gaps with other Zollverein cities by 28% on average, although this decline cannot be interpreted exclusively as the consequence of the abolition of trade barriers because there are no data on duties on wheat before the accession. Chilosi and Federico (2015) find that the Corn Laws had a significant effect on price gaps between the United States and England. Finally, for the whole of Europe, Jacks (2006) finds that outright prohibition or protection against imports increased dispersion in wheat prices and slowed down the adjustment after a shock.

The effects of another trade barrier, the monopoly that European trading companies [European trading companies, monopoly], such as the Dutch *Vereenigde Oost-Indische Compagnie* (VOC) and the British East India Company (EIC) enjoyed on commerce with their Asian territories, has attracted a lot of attention and some very interesting controversy in recent times. O'Rourke and Williamson (2002) argued that only these restrictions could explain the huge gaps in spice prices between Asia and Europe in the seventeenth and eighteenth centuries. Their hypothesis tallies well with the situation of the spice market in Europe in the early sixteenth century, when the competition between the Portuguese and Venetians increased the correlation in spice prices in Europe and caused real prices to decline (O'Rourke and Williamson 2009). Along the same line, de Zwart (2016) interprets the differences in price convergence of various products between South Asia and Europe as a consequence of different market conditions. Prices converged earlier and faster in competitive goods, including pepper, while gaps remained wide in products such as cloves and nutmeg, where the trading companies succeeded in enforcing their monopoly. The

impact on welfare ultimately depended on the relative importance of these goods on total consumption. Chilosi and Federico (2015) show how trading monopolies still mattered in the first half of the nineteenth century. The French competition had disappeared, while the EIC maintained its monopoly on trade with India until 1815, and the Dutch recreated the monopoly of the failed VOC with the Nederlandsche Handel-Maatschappij (NHM) in the 1820s. The dummies for these companies are positively and significantly related to price gaps between England and the Netherlands for a wide range of commodities. In contrast, most commodity-specific market interventions in the interwar years, such as the Agricultural Adjustment Act (1933) or the Dutch marketing board for sugar (VSP and later NIVAS), had little impact on price gaps.

Information costs play an essential role in integration, especially efficiency, but measuring them has proved exceedingly difficult. Brunt and Cannon (2014) measure the proportion of cities with at least one local newspaper as a proxy for diffusion of news in the late eighteenth- to early nineteenth-century England. The variable has no effect on the speed of adjustment, but it is negatively related to the volatility of relative prices and positively to the correlation of shocks. In other words, prices in different locations reacted in the same way to presumably common news. Chilosi and Federico (2015) capture the impact of the telegraph on transatlantic price gaps with a simple dummy, which is negative (the telegraph reduced gaps) and significant. Steinwender (2018) explores in detail the impact of the reduction in the time for transmission of information (from about 10 days before the telegraph to less than 1 after it) on daily prices of cotton in New York and Liverpool. As expected, she finds that the telegraph reduced the information-adjusted price differentials (i.e., the gap between the New York price and the latest known Liverpool price). She also finds that the faster information increased the average and volatility of daily shipments but finds little support for the hypothesis that the telegraph affected storage.

The set of controls $\mathbf{X}$ includes a potentially boundless list of factors that might affect integration, which in panel regressions are lumped together in the fixed effect under the implicit assumption that they did not change in the period of estimation. Several authors have tried to be more specific, adding additional variables (or, more frequently, proxies and dummies), which can be grouped into four main categories:

1. Monetary factors. Currency fluctuations increase risks for arbitrage, and thus one would expect integration to be greater, ceteris paribus, under fixed exchange rates. Several authors have tested this by including monetary variables, such as exchange rate volatility (Jacks 2006), or dummies for monetary unions (Keller and Shiue 2008; Jacks 2006) or the gold standard (Jacks 2006). As a rule, results tally with expectations.
2. Ethnicity. Common ethnicity is expected to foster integration by increasing trust among traders. Usually it is proxied by common language, and, indeed, linguistic fractionalism increased price dispersion in Austria-Hungary (Schulze and Wolf 2009), and common language reduced price gaps between pairs of European cities and increased the speed of adjustment (Jacks 2006).

3. Measures of institutional quality. It is highly likely that efficient and universally fair legal systems fostered integration, but only Bateman (2012) tries to test this hypothesis for early modern Europe. She proxies quality with indexes of fiscal centralization or the activity of (traditional) parliaments, but neither variable is significant. This result should not be considered as conclusive, as these variables are arguably poor proxies for the efficiency of the legal system.
4. The existence of a political border. Engel and Rogers (1996) pioneered this measure by adding a dummy for pairs of markets in the United States and Canada to highlight the impact of institutional and other unaccounted-for differences on the volatility of relative prices. Jacks (2009) has reproduced their simplified model for a large number of markets, adding only volatility of exchange rates and, as expected, finds a long-term decline in the border coefficient in the long nineteenth century. Also, Andrabi and Kuelwhein (2010) find a significant effect of borders between British India and the princely states on price gaps. By its residual nature, one would expect the border dummy to be more relevant in bare-bone specifications. In fact, it comes out insignificant and wrongly signed in the analysis of integration of the Italian market by Federico (2007), which features among explicative variable yearly series of transport costs and duties. It is thus somewhat puzzling that the border dummy is still significant in the article by Jacks (2006), which features a rich set of proxies for other trade costs. Furthermore, the variable is positive in both regressions – i.e., ceteris paribus a border increased price dispersion as expected but also, unexpectedly, reduced efficiency.

By their nature, trade costs have a potentially permanent effect on integration, while wars are by definition temporary, if not exceptional, events. The effect of wars on price gaps is unambiguously negative, while there is a margin of uncertainty on their effect on efficiency. Wars disrupted the orderly working of markets, but in some cases, especially of backward societies, they could increase circulation of information, with positive effect on efficiency. Jacks (2006) tests the effect of six different cases of domestic and foreign war, with no-war as the default. In all these combinations except "neutral" (defined as the combination of a market in a neutral country and another in a country at war), the war dummy is significant: wars increased dispersion and, more tellingly, slowed down the adjustment. Chilosi and Federico (2015) systematically test the effect of a number of political events on intercontinental price gaps. The results are mixed. Some events did disrupt trade (the anti-slavery campaign in the 1830s for sugar, World War One on routes for India and Indonesia), while others (the Java war of the 1820s, the Indian mutiny, US Civil war, and World War One across the Atlantic) had no significant effect. The last result does not necessarily imply that these events did not affect price gaps. In fact, the regression includes a yearly series of trade costs, so the dummy captures additional effects, such as the shortage of ships for Asian trade during World War One.

This short review highlights two points. First, the same variable yields different results in explaining convergence or efficiency. This is not really surprising and is fully consistent with the results of measurement (cf. e.g., Dobado et al. (2012),

Hynes et al. (2012), Federico et al. (2018)). In an efficient market, a shrinking in commodity points, as a consequence of a decline in transport costs or of the abolition of barriers to trade, causes convergence, but does not necessarily increase the speed of adjustment after a shock.[1] On the other hand, an improvement in information flows (a decrease in costs or an increase in speed of transmission) or in institutions may increase the speed of adjustment, but does not necessarily affect the equilibrium price gaps. They might affect the measured price gaps to the extent that price differentials are computed as averages of actual prices, inclusive of periods of disequilibrium. The average gap would shrink if the latter become shorter because arbitrage is more efficient and/or less risky, even if the underlying equilibrium gaps remain constant.[2]

Second, a list of significant variables is not a sufficient answer to question two without an idea of their relevance. It is necessary to know how big is big or to borrow the title of a paper by Jacks (2006) "what drove market integration?" The question could be properly addressed only by including in the regression measures or proxies for all main drivers of integration. As is clear from previous survey, Jacks (2006) work is by far the most comprehensive, as he tests the effect of no less than 19 variables on convergence and efficiency. In Table 5, he sums up his results by ranking separately the impact of continuous variables (as measured by the impact of standard deviations on the dependent variable) and dummies (as measured by their coefficients). This division makes the comparison between the two categories of variables difficult. The author nevertheless concludes that "trade costs seem to be more responsive to changes in the choice of monetary regimes than changes in the underlying technology of transport," while "speeds of price adjustment present a more balanced account as transport, monetary, and commercial variables all seem to play a part" (Jacks 2006: 405). Federico (2007) and Chilosi and Federico (2015) estimate the contribution of each explicative variable to total price convergence by multiplying the coefficients from a log-log specification by the actual changes of the underlying variables (omitting nonsignificant ones). The results highlight a major difference between the first half of the nineteenth century ("early globalization") and the period from about 1870 to World War One ("heyday of globalization"). The early convergence was mostly determined by changes in barriers to trade, such as the abolition of trading monopolies by Western companies (the British EIC in 1815–1816 for India and the Dutch NHM around 1850) and the liberalization of British imports of wheat in 1842. After 1870 prices continued to converge, thanks to the telegraph lines and to technical progress in sea transportation, but the political decisions mattered more because most total convergence predated 1870.

[1]Note that a reduction in commodity gap increases ceteris paribus the likelihood of co-movements as it shrinks the scope for independent movements. However, this appears as an increase in efficiency only because co-movement measures are imperfect measure of efficiency.

[2]As stated earlier, in principle, the TAR models should solve the problem by endogenously selecting the "normal" observations to estimate the equilibrium commodity points. However, the results can be interpreted as effective equilibrium transaction costs on the twin assumptions of efficient markets and constant transaction costs in the period of estimation.

It may be premature to draw any conclusion from this still ongoing work, but the results so far point to an explanation for the difference between the early modern period and the "long" nineteenth century. This latter was the golden age of integration because of the coincidence of trade liberalization and a spurt of technical progress. The latter lowered transportation costs, especially overland, and increased the speed of transmission and the amount of available information, which increased efficiency first by increasing the proportion of common shocks (as determined by common information) and later, after the telegraph, also by speeding the adjustment.[3] But throughout history, waves of integration/disintegration were determined mainly by political decisions, such as barriers to trade, and by political events, most notably wars.

The analysis of causes has made impressive strides in the last 10 years, but its results have to be taken with an (abundant) pinch of salt. First and foremost, in most cases, the lack of data makes it impossible to consider potentially relevant causes. A major example is the role of the network of state granaries in China, set up by the Qing dynasty in the second half of the seventeenth century. Several scholars attribute to it the high level of "integration" (or efficiency) of Chinese domestic markets in the eighteenth century and to its decadence the disintegration of the early nineteenth century. This inference is plausible, as the system was designed to reduce the effect of local harvests on local supply and thus on price fluctuations. The inference would also be testable with information about the location and activities of granaries. Unfortunately, these do not exist, or at least have not been exploited so far.

Finally, there are three specific econometric issues, about endogeneity, the choice of market pairs, and the measurement of trade costs.

First, endogeneity is a widespread concern in modern-day econometrics that also affects the literature on market integration. Quite a few authors have used instrumental variables to address two potential sources of reverse causation. Existing trade flows may have determined the choice of railways or roads to be built, and political decisions, such as access to the Zollverein (Keller and Shiue 2008, 2014). In addition, transportation costs could be endogenously determined by the amount of trade (Jacks and Pendakur 2010). In these cases, a simple OLS regression might overstate the contribution of new infrastructures or liberalization of trade to integration. These concerns might be excessive. It is surely plausible that total trade affected the railway planning but less likely that trade in a single commodity (most frequently cereals) was so important as to determine it. It is even less likely that any specific commodity/route flow was large enough to affect the level of freights in the competitive world market for shipping.

[3]Let's assume that a war started in – say – Germany. Newspaper would transmit the news roughly at the same time in Amsterdam and London, and agents were likely to react in the same way by raising prices. This movement would appear as a common shock in the Brunt and Cannon (2014) framework and would increase the share of explained variance in DFA (Federico et al. 2018). In this case, there would be no market-specific shock to adjust to. Newspapers did not cut the time of reaction to any such market-specific shock, which still depended on the physical transmission of pieces of paper in the days before the telegraph.

Second, in contrast with the concern for endogeneity, most authors do not seem to bother about the existence of trade flows between each pair of locations, even if, as discussed earlier, this is an essential condition for price gaps to be a meaningful measure of integration. There are exceptions. For instance, Brunt and Cannon (2014) explicitly state that they prefer to run their VECM models only with pairs of neighboring counties, which were more likely to trade than faraway ones. They can afford to be so restrictive because their database includes a very large number of nearby counties to work with. To be sure, data on trade flows between specific cities are often unavailable, but the anecdotal information is abundant, and thus authors should defend their choices of market pairs beyond the obvious ones (e.g., nobody would doubt that Amsterdam and Konigsberg traded in the seventeenth century).

Third, the frequent resort to proxies for trade costs [trade costs, proxies for], admittedly quite common in gravity models, may introduce biases in estimation. For instance, using a common dummy for protection would yield imprecise coefficients if the duty is similar across pairs of markets and constant in time and a biased estimate if duties differ substantially among countries or change in time. This was the norm in the nineteenth century: Italian duties on wheat were first imposed in 1887, at about 10% of the world price, rose to a peak around 50% in the mid-1890s, and declined back to 30–35% on the eve of WWI. The problem is even more serious for transportation costs. By definition, distance is time-invariant and thus is a bad proxy for measuring the contribution of declining costs to integration. Unfortunately, precisely measuring transportation costs, especially overland ones, is difficult. The task is comparatively easy for water transport: the distance is fixed (although the length of a meandering river can exceed the distance by bird fly), and data on seaborne freights is readily available, although not necessarily route- and commodity-specific. In contrast, the distance of land transport depended on the shape of the network (the opening of a new road or of a new rail line could shorten it), and the data on unit costs are either scarce, as for road transportation, or very abundant, but often difficult to collect and unwieldy to manage, as for railways. The evidence suggests that the costs of road transport did not change that much before the diffusion of the internal combustion engine in the twentieth century, but the utilization of roads depended on the competition from railways. The productivity of railways was growing and thus one would expect rates to decline fairly steadily if the market were competitive. In most countries, this was not the case, thus the gains accrued to railway owners, including in quite a few cases the state, rather than being transferred to customers. Thus, one should look at actual fares, which differed by product, route, size of shipment, and so on. This combination of changes in network and in unit fares causes total costs of rail transportation to vary considerably. For instance, consider the case of rail transportation between Genoa, the main Italian port, and Rome. The distance as the bird flies and by sea is about 400 km. The first railway link via Florence, Bologna, and Turin (905 Km) was established in 1866, and the length of the connection was slashed by 40% (to 553 km) after the opening of the new line along the coast in 1875. The combined effect of shorter rail distance and cuts to unit fares lowered the total cost of rail transportation in 1890 to 22% of its 1866 level (Federico 2007). A dummy for rail connection in 1866–1890 would

measure only the initial impact of rail, and miss the subsequent decrease, thus understating the overall effect of railways on convergence. Most authors simply ignore the issue, while others use simple solutions such as interacting dummies (or rail distances) with time trends (Jacks 2006, 2009), segmenting the period of observation to get different coefficients for dummies (Uebele and Gallardo-Albarran 2015), or using ruggedness as a time-invariant proxy for railway costs (Keller and Shiue 2016). It is unclear whether these solutions can address the issue.

One might expect that the combined effect of poorly measured dependent variable and massive use of not so accurate proxies would yield poor results. As is evident from this review, this is not the case: most authors get significant coefficients from their imperfect measures. Either they are quite lucky,or underlying effects are quite large.

The Third Wave: The Effects of Integration

Authors in the market integration literature routinely motivate their interest in the field by reminding that the division of labor was a major driver of ("Smithian") economic growth in the past. Consequently, one would expect a massive effort to estimate the gain from integration benefits. Indeed, social savings from railroads had just such a prominent role in the early years of the cliometric revolution, and standard trade theory offers a wide range of estimation techniques. Price gaps depend on barriers to trade, and adding the impact of transportation costs is straightforward. Yet, somewhat contradictorily, estimates of benefits from integration [benefits from integration, estimates] are few and somewhat on the fringe of the literature.

Ejarnes and Persson (2010) and Steinwender (2018) focus on efficiency gains from the layout of telegraph cables between Europe and the United States in July 1866 with a broadly similar framework. Ceteris paribus, the amount of trade was positively related to the speed of transmission of information, as uncertainty in prices at destination led to a reduction in shipments. The fall in the time of transmission reduced this uncertainty and thus increased American exports to the United Kingdom. The authors use different specifications for different goods (wheat and cotton) with different data (monthly versus daily), but the results are strikingly similar – an increase of exports by 3–8% relative to their pre-telegraph level. The figure may seem impressive, but they correspond to an increase of 0.10–0.20% of American GDP in the 1860s.[4]

Gains from price convergence can be estimated by adjusting the standard partial equilibrium analysis of the welfare effects from trade liberalization to the more realistic case of a partial cut rather than a complete abolition (Hufbauer et al. 2002). It implies adding to the well-known Harberger triangles another term,

[4]The figure is obtained assuming that value added accounted for 90% of the export price of cotton and wheat, that wheat and cotton accounted for half of American exports, and that the export/GDP ratio was about 0.05.

where gains from integration of a given product are proportional to the absolute difference between shares of that product on consumption and output of the country. Thus, the benefits from integration are greater the more relevant the product and the more specialized it is. Federico and Sharp (2013) estimate the losses from the regulation of American rail fares, which prevented a full transfer of productivity gains to consumers and an adjustment to collapsing prices for agricultural products during the Great Depression. They ranged from 0.6 to 3.1% of GDP in the 1930s, depending on levels of specialization and demand and supply elasticities. Chilosi and Federico (2016) deal with the benefits from transatlantic price convergence in the nineteenth century. Gains were similar for the United Kingdom (1–3% of GDP in 1913), India (1.2–1.7%), and Indonesia (0.6–2.2%), while they were much lower for the United States (only 0.10–0.3%).

One might argue that gains below 3% are small, and consequently market integration was less economically relevant than assumed. However, this conclusion would seem too hasty, for three reasons.

First, the estimate refer to few, albeit important, products only, and changes in trade costs although sizeable (a decline by a third in costs) are much smaller than other historical ones, such as the convergence in domestic prices after the introduction of railways. It is likely that more extensive coverage of individual products would augment the size of total benefits, although it would be rash to mechanically extend the product-specific estimates to all goods. The effect of unit changes in transportation costs on price gaps depended on many factors, most notably the bulkiness of products. As a result, gains were likely to be smaller for manufactures than for commodities such as wheat. On the other hand, barriers to trade in manufactures were often higher.

Second, all estimates are partial-equilibrium static models, thus neglecting both the static general equilibrium effects and the dynamic benefits from Smithian growth. Donaldson has put forward two different strategies to estimate general equilibrium effects of market integration [benefits from integration, general equilibrium estimates], both broadly inspired to new trade theory. Costinot and Donaldson (2016) use a linear programming approach to estimate the differences between maximum potential and actual yields and local and New York prices in each American county for benchmark years from 1887 to 1997. Then they measure the gains from market integration (technical progress) by comparing the actual output with a counterfactual one, which they compute as prices (yields) of the current year times the yields (prices) of the next benchmark. The results are quite impressive: market integration increased agricultural output in the United States by 62% from 1887 to 1920 and by 55% from 1954 to 1997, about as much as technical progress (respectively, 30% and 70%). The first figure implies that market integration accounted for about a tenth of total growth in American GDP before 1920.[5] Donaldson (2018) uses a three-step regression-based approach to estimate the effects

[5]Agriculture produced about 27% of GDP in 1890, and the additional gross output corresponded to about 18% of GDP (assuming a GDP/output ratio around 0.90).

of railways on agricultural output of Indian districts (quite a good proxy for GDP in the case at hand) from 1870 to 1939. He shows that his price-based estimate of trade costs determined trade flows; the existence of a railway connection increased district level agricultural income by 16%; and that about 85% of this increase is explained by the additional trade from railways.

Finally, the lack of estimates of the benefits of integration for economic growth [benefits from integration, dynamic estimates] is hardly surprising, given that economists have not yet found a way to get country-level estimates. The market integration literature faces the additional problem of a dearth of measures of local GDP. Yet, at least two authors have tried to use the standard growth regression approach for market integration, using different proxies for GDP. Keller and Shiue (2016) use total population as a proxy for the level of development of each German city in the nineteenth century and find it to be inversely correlated with the average size of price gaps with all other cities in the sample. Bateman (2012) has estimated the effect of price gaps and volatility on level and trends of real wages and urbanization rates in early modern Europe. In her case, the results are poor as the integration variables are not significant. Neither results are really conclusive, given the poor measure of GDP and the well-known shortcomings of the growth regression approach.

Conclusion: Taking Stock

Federico (2012) concluded his survey of the literature on European integration in a "less than upbeat" tone. He pointed out the almost exclusive focus on measurement (question 1 from the introduction), the high proportion of studies on cereals (as opposed to other commodities and above all to manufactures), the limited size of samples, the prevailing interest in state-of-the-art econometric technicalities, and, by contrast, the insufficient attention to economics of arbitrage (as shown by the confusion between Cournot's two conditions). He also argued that price convergence may be more important than an increase in efficiency. An efficient market increases welfare by making adjustments to external shocks faster and easier, but does not affect the allocation of factors and consumption by much (unless the starting point is a totally inefficient market, which prevents prices from adjusting to changes in trade costs). In contrast, convergence would affect relative prices in the long term, with potentially massive effects on agents choice and ultimately on the economy. Last but not least, he put forward an ideal (and perhaps too ambitious) research agenda for measuring integration.

Seven years later, these shortcomings are still quite evident, but there are reasons for hope. There is potentially a more comprehensive theoretical framework, integrating time, via storage, with spatial arbitrage. The most recent works on measurement of integration rely on substantially larger samples of cities for longer periods of time, and have extended the range of goods, most notably to spices, even if they might not be representative (Solar 2013; De Zwart 2016). The literature on causes of

integration is flourishing with a common (albeit still somewhat questionable) framework, and some authors have started to measure effects. We know much more than before, and the new research suggests a more complex and nuanced, although not so different, story.

Markets have become more integrated in the long run, but the process has been far from steady. In pre-industrial Europe, price dispersion and, to a lesser extent, efficiency featured huge fluctuations with large differences across areas, and it is very difficult to find any clear common trend. There is some evidence of price convergence in Asian trade and of increasing efficiency in the late eighteenth century both in Europe and in the Atlantic market, which could compensate for the disintegration of the Chinese domestic market. A massive process of integration almost everywhere, with the likely exception of mainland China, characterized the nineteenth century. Integration most likely peaked sometime between 1860 and World War One and remained close to that peak until the Great Depression. The disintegration of the market during the Depression appears less devastating from the point of view of prices than of quantities, possibly because commodities were less affected than manufactures by the protectionist backlash.

There is widespread consensus on the importance of technical progress in transportation and transmission of information for price convergence in the long nineteenth century and possibly for the increase in efficiency since the late eighteenth century. In contrast, technical change does not seem to have played a major role in determining integration in the early modern period, although this statement is based more on circumstantial evidence than on quantitative testing. Thus, most authors explain the fluctuations in the level of integration in general, and especially in dispersion, in the early modern era to changes in barriers to trade. Furthermore, liberalization of trade was in all likelihood as important as technical progress for the nineteenth-century spurt. Thus, political decisions may have been the driving force of integration in the very long run.

The research on gains from market integration is still beginning. There are very few estimates, which yield fairly small figures. In all likelihood they represent a lower bound, for a number of methodological and empirical reasons.

The measurement of the benefits of integration is arguably the most difficult, but also the most rewarding topic in the field. The partial equilibrium static models are useful as a first approximation, but their results downplay the true benefits of integration. On the other hand, adding spatial dimension for all integrating areas into classical CGE models seems very difficult. Models from new trade theory are promising, but it is too early to tell how helpful they will be. A fortiori this conclusion holds true for the dynamic effect of integration. The growth regression approach, even if perfect data were available, is plagued by the twin problems of collinearity and too many potential explicative variables. On the other hand, there seems to be no simple alternative approach.

The second main task seems to be the improvement of the databases. The progress in coverage by product and area is substantial but still insufficient, and the work on explicative variables is rather backward (a problem that also bedevils the

macroeconomic literature on trade). Poorly measured explicative variables may be significant and still give misleading results about the importance of different factors of integration.

More generally, in spite of all progress so far, the ambitious agenda by Federico (2012) is far from fulfilled. There is still a lot to do.

Cross-References

► Cliometric Approaches to International Trade

References

Anderson J, van Wincoop E (2004) Trade costs. J Econ Lit 42:691–751
Andersson F, Ljungberg J (2015) Grain market integration in the Baltic Sea region in the nineteenth century. J Econ Hist 75:749–790
Andrabi T, Kuehlwein M (2010) Railways and price convergence in British India. J Econ Hist 70:351–377
Bateman V (2012) Markets and growth in early modern Europe. Pickering and Chatto, London
Bernhofen D, Eberhard M, Li J, Morgan S (2016) Assessing market (dis)integration in early modern China and Europe. CEPR DP 11288
Bessler D (1990) A note on Chinese Rice prices: interior markets, 1928-1931. Explor Econ Hist 27:287–298
Brandt L (1985) Chinese agriculture and the international economy 1870-1930: a reassessment. Explor Econ Hist 22:163–198
Brunt L, Cannon E (2014) Measuring integration in the English wheat market, 1770-1820: new methods, new answers. Explor Econ Hist 52:111–130
Buyst E, Dercon S, von Campenhout B (2006) Road expansion and market integration in the Austrian Low Countries during the second half of the 18th century. Hist Mes 21:185–219
Chilosi D, Federico G (2015) Early globalizations: the integration of Asia in the world economy, 1800-1938. Explor Econ Hist 57:1–18
Chilosi D, Federico G (2016) The effects of market integration: trade and welfare during the first globalization, 1815–1913 LSE Economic History WP 238/2016
Chilosi D, Volckart O (2011) Money, states, and empire: financial integration and institutional change in Central Europe, 1400-1520. J Econ Hist 71:762–791
Chilosi D, Murphy T, Studer R, Tuncer C (2013) Europe's many integrations: geography and grain markets. Explor Econ Hist 50:46–68
Chilosi D, Schulze MS, Volckart O (2018) Benefits of empire? Capital market integration North and South of the Alps, 1350-1800. J Econ Hist 78:637–672
Clark G (2015) Markets before economic growth: the grain market of medieval England. Cliometrica 9:265–287
Coleman A (2007) The pitfalls of estimating transaction costs from price data. Evidence from trans-Atlantic gold-point arbitrage, 1886-1905. Explor Econ Hist 44:387–410
Coleman A (2009) Storage, slow transport and the law of one price: theory with evidence from 19th century US corn markets. Review Economics and Statistics 91:332–350
Costinot A, Donaldson D (2016) How large are the gains from economic integration? Theory and evidence from US agriculture, 1880–1997 NBER WP 22496
Cournot A (1971, English edition, New York) Recherches sur les principles mathematiques de la theorie des richessesses (Original date of publication 1838)

Craig L, Holt M (2017) The impact of mechanical refrigeration on market integration: the US egg market, 1890-1911. Explor Econ Hist 66:85–105
De Zwart P (2016) Globalization in the early modern era: new evidence from the Dutch-Asian trade, c 1600-1800. J Econ Hist 76:520–558
Dobado R, Marrero GA (2005) Corn market integration in Porfirian Mexico. J Econ Hist 65:103–128
Dobado R, Garcia-Hiernaux A, Guerrero D (2012) The integration of Western hemisphere grain Markets in the Eighteenth Century: early rise of globalization in the west. J Econ Hist 72:671–707
Dobado R, Garcia-Hiernaux A, Guerrero D (2015) West versus Far East: early globalization and the great divergence. Cliometrica 9:235–264
Donaldson D (2015) The gains from market integration. Annual review of economics 7:619–647
Donaldson D (2018) Railroads of the raj: estimating the impact of transportation infrastructure. Am Econ Rev 108:899–934
Engel C, Rogers JH (1996) How wide is the border? Am Econ Rev 86:1112–1125
Ejrnæs M, Persson KG, Rich S (2008) Feeding the British: convergence and market efficiency in the nineteenth century grain trade. Economic History Review 61 S1:140–171
Ejrnæs M, Persson KG (2010) The gains from improved market efficiency: trade before and after the transatlantic telegraph. Eur Rev Econ Hist 2010:361–381
Epstein SR (2000) Freedom and growth. The rise of states and markets in Europe. Routledge, London, pp 1300–1750
Federico G (2007) Market integration and market efficiency. The case of 19th century Italy. Explor Econ Hist 44:293–316
Federico G (2011) When did the European market integrate? Eur Rev Econ Hist 15:93–126
Federico G (2012) How much do we know about market integration in Europe? Economic history review 65:470–497
Federico G, Persson KG (2007) Market integration and convergence in the world wheat market, 1800-2000. In: Hatton T, O'Rourke K, Taylor AM (eds) The new comparative economic history: essays in honour of J. Williamson. MIT Press, Cambridge, MA, pp 87–114
Federico G, Sharp P (2013) The cost of railroad regulation: the disintegration of American agricultural Markets in the Interwar Period. Economic history review 66:1017–1038
Federico G, Tena-Junguito A (2016) World trade, 1800–1938: a new data-set. EHES Working paper no. 93
Federico G, Schulze M, Volckart O (2018) European Goods Market Integration in the Very Long Run: From the Black Death to the First World War LSE Economic History Working papers 277/2018
Foldvari P, van Leuuwen B (2011) What can price volatility tell us about market efficiency? Conditional heteroscedasticity in historical commodity prices series. Cliometrica 5:165–186
Hufbauer G, Wada E, Warren T (2002) The benefits of Price convergence: speculative computations. Institute for International Economics, Washington, DC
Hurd J (1975) Railways and the expansion of Markets in India, 1861-1921. Explor Econ Hist 12:263–288
Hynes W, Jacks D, O'Rourke K (2012) Commodity market disintegration in the interwar period. European Review Economic History 16:119–143
Jacks DS (2006) What drove 19th century commodity market integration. Explor Econ Hist 43:383–412
Jacks D (2009) On the death of distance and borders: evidence from the 19th century. Econ Lett 105:230–233
Jacks DS, Pendakur K (2010) Global trade and the maritime transport revolution. Review Economics and Statistics 92:745–755
Keller, Wolfgang and Carol Shiue (2008) Tariffs, trains and trade: the role of institutions versus technology in the expansion of markets NBER WP 13913., April 2008

Keller W, Shiue C (2014) Endogenous formation of free trade agreements: evidence from the Zollverein's impact on market integrations. J Econ Hist 74:1168–1204
Keller W, Shiue C (2016) Market integration as a mechanism of growth. CEPR DP 11627
Lampe M, Sharp P (2015) How the Danes discovered Britain: the international integration of the Danish dairy before 1880. Eur Rev Econ Hist 19:432–453
Latham AHJ, Neal L (1983) The international market in Rice and wheat, 1868-1914. Economic History Review 36:260–280
Li L (2000) Integration and disintegration in North China's grain markets, 1738-1911. J Econ Hist 60:665–699
Li L-F (2017) Arbitrage, communication, and market integration at the time of Datini. Eur Rev Econ Hist 21:414–433
Officer L (1996) Between the gold-sterling gold points. Cambridge University Press, Cambridge
O'Rourke K, Williamson JG (1994) Late nineteenth-century anglo-american factor price convergence: were Heckscher and Ohlin right? J Econ Hist 54:892–916
O'Rourke K, Williamson JG (2002) When did globalization begin? Eur Rev Econ Hist 6:23–50
O'Rourke K, Williamson JG (2009) Did Vasco de Gama matter for European markets? Economic History Review 62:655–684
Panza L (2013) Globalization and the near east: a study of cotton market integration in Egypt and Western Anatolia. J Econ Hist 73:847–872
Pomeranz K (2000) The great divergence. Princeton University Press, Princeton
Ravaillon M (1987) Markets and famines. Oxford University Press, Oxford
Ronnback K (2009) Integration of global commodity markets in the early modern era. Eur Rev Econ Hist 13:95–120
Schulze M-S, Wolf N (2009) On the origins of border effects: insights from the Habsburg empire. J Econ Geogr 9:117–136
Sharp P, Weisdorf J (2013) Globalization revisited: market integration and the wheat trade between North America and Britain from the 18th century. Explor Econ Hist 50:88–98
Shiue C (2002) Transport costs and the geography of arbitrage in 18th century China. Am Econ Rev 92(5):1406–1419
Shiue CH, Keller W (2007) Markets in China and Europe on the eve of the industrial revolution. Am Econ Rev 97:1189–1216
Solar P (2013) Opening to the east: shipping between Europe and Asia, 1770-1830. J Econ Hist 73:625–661
Steinwender C (2018) Real effects of information frictions: when the States and the Kingdom became United American Economic Review 108:657–696
Stigler G, Sherwin RA (1985) The extent of the market. J Law Econ 28:555–585
Studer R (2008) India and the great divergence. Assessing the efficiency of grain markets in 18th and 19th century India. J Econ Hist 68:393–437
Studer R (2015) The great divergence reconsidered. Europe, India and the rise to global economic power. Cambridge University Press, Cambridge
Taylor AM (2001) Potential pitfalls for the purchasing-power-parity puzzle? Sampling and specification biases in mean-reversion tests of the law of one price. Econometrica 69:473–498
Uebele M (2011) National and international market integration in the 19th century: evidence from co-movement. Explor Econ Hist 48:226–242
Uebele M, Gallardo-Albarran D (2015) Paving the way to modernity. Prussian roads and grain market integration in Westphalia, 1821-1855. Scand Econ Hist Rev 63:69–92
Van Bavel B (2016) The invisible hand? How market economies have emerged and declined since AD 500. Oxford University Press, Oxford
Volckart O, Wolf N (2006) Estimating financial integration in the middle ages: what can we learn from a TAR model? J Econ Hist 66:122–139
Wolf N (2009) Was Germany ever united? Evidence from intra- and international trade 1885-1933. J Econ Hist 69:846–881

Part IV

Institutions

African-American Slavery and the Cliometric Revolution

Richard Sutch

Contents

Abstract

This chapter explores the significant contributions to the history of African-American slavery made by the application of the tools of cliometrics. As used here "cliometrics" is defined as a method of scientific analysis marked by the explicit use of economic theory and quantitative methods. American slavery of the late antebellum period (1840–1860) was one of the earliest topics that cliometricians focused on and, arguably, the topic upon which they made the largest impact.

R. Sutch (✉)
Economics Department, University of California, Riverside, CA, USA

National Bureau of Economic Research, Cambridge, MA, USA
e-mail: richard.sutch@ucr.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_62

Keywords
Slavery · Slave trade · Slave breeding · Slave markets · Cotton plantations · African-Americans · Cliometrics · Antebellum economic growth

Introduction

No historical topic has been more central to the rise, influence, and refinement of the cliometric approach than African-American slavery. The first self-conscious cliometric contribution to American history was a 1957 conference paper, "The Economics of Slavery in the Antebellum South" by Alfred Conrad and John Meyer, which subsequently became an article in the *Journal of Political Economy* (1958). Many of the academic economists associated with the pioneering decades of the cliometric revolution, including myself, examined slavery equipped with the tools of economic theory and quantitative methodology: Douglass North (1961), Richard Sutch (1965), Peter Temin (1967), William Parker (1970b), Robert Gallman (1970), Gavin Wright (1970), and Robert Fogel and Stanley Engerman (1971b). We were not alone. *The Bibliography of Historical Economics to 1980* lists no fewer than 96 individual contributors to the examination of the history of slavery (McCloskey and Hersh 1990).

In the midst of this flood of publications, an ambitious and contentious book on the topic, provocatively titled *Time on the Cross*, touched off a scholarly whirlwind about the potential, the limitations, and the misuse of the cliometric method (Fogel and Engerman 1974). The controversy turned into an academic battle on two fronts, the historians against the cliometricians (Herbert Gutman 1975; Kenneth Stampp 1976) and a Civil War among the cliometricians (David et al. 1976). Although the clamor has since died down and the pace of new findings has slowed, the topic continues to fascinate. Cliometric studies periodically feed the appetite for quantitative evidence on the American horror story (recent examples include Wanamaker 2014; Bodenhorn 2015; Olmstead and Rhode 2015; Calomiris and Pritchett 2016; Pritchett and Hayes 2016).

What explains the original excitement and the sustained attention? First, and I think most important, the Cliometric Revolution really was a revolution. It was so viewed at the time (North 1963: 128). It appears so in the statistics of the scholarship (Whaples 1991). Consequentially, the revolution redefined the scope, method, and goals of economic history as a discipline. Economics departments staffed by professors with degrees in economics commandeered the field of economic history, which was subsequently abandoned by history departments and professors with degrees in history. Like with all revolutions, the young were in the frontlines. Assistant professors and graduate students published papers overturning the interpretations of distinguished full professors and scholars who had been held in high repute for several generations. That was empowering and tended to make true believers of those who joined the revolution. As true believers, they persisted for the long haul.

The first efforts that engaged cliometric methods addressed easy questions and there were many. One of the most enthusiastic advocates of econometric history, Robert Fogel, claimed that any random page from a history book contained "either an explicit or implicit quantitative statement that needed to be measured" (Fogel 1990: 28 (350)). The economic theory and statistical expertise required was often elementary. The only difficult work was gathering data from the archives. While tedious, this required no special talents, but the resulting studies achieved great success. They revised the old consensus interpretation of slavery and swept aside the sentimental (and racist) descriptions of the Old South and its "peculiar institution" seemly without protest. As Peter Temin has noted, that impact made the topic and even the tedious data collection "an enormous amount of fun" (Temin 1999: 45 (433)).

The topic of slavery held special relevance at the time because the Conrad and Meyer 1957 conference paper appeared during the Civil Rights Era. That year federal troops were sent to enforce school integration in Little Rock. Martin Luther King gave his "I have a Dream" speech at the 1963 March on Washington. The Civil Rights Act of 1964 prohibited discrimination based on race, color, religion, or national origin. In 1965, President Lyndon Johnson gave his "We Shall Overcome" speech to Congress and the Voting Rights Act was passed. In that context, a "new" view of slavery seemed to be required to accompany the movement for racial equality. Many of the first cliometricians were personally committed to the civil rights movement. For example, Gavin Wright's direct involvement led him to pursue graduate study in economics and to specialize in economic history and the study of slavery (Wright 2013).

Race remains a hot-button topic and any hard look at the history of slavery is controversial; discussions about black slavery are often fraught with emotion. Articles and books on the subject, however dispassionate on their surface and from whatever quarter, still attract close attention and agitated debate (for a recent example, see Murray et al. 2015). My task, however, is not to offer a behind-the-scenes, I-was-there memoir, but to put emphasis on the concrete contribution of cliometrics, quantification, and social science to our knowledge of the history of slavery.

Conrad and Meyer

The pioneering contribution of Conrad and Meyer addressed a simple, straightforward question. Was slavery profitable? Their positive answer directly confronted a view on the economics of slavery then current. Despite the important work of historians such as Lewis Gray (1933) and Kenneth Stampp (1956), it was still widely taken for granted that American slavery had become economically moribund by late antebellum period, the decades of the 1840s and 1850s. The interpretation then prominent in the social sciences had been promulgated by Ulrich Bonnell Phillips in "The Economic Cost of Slaveholding in the Cotton Belt," a journal article published

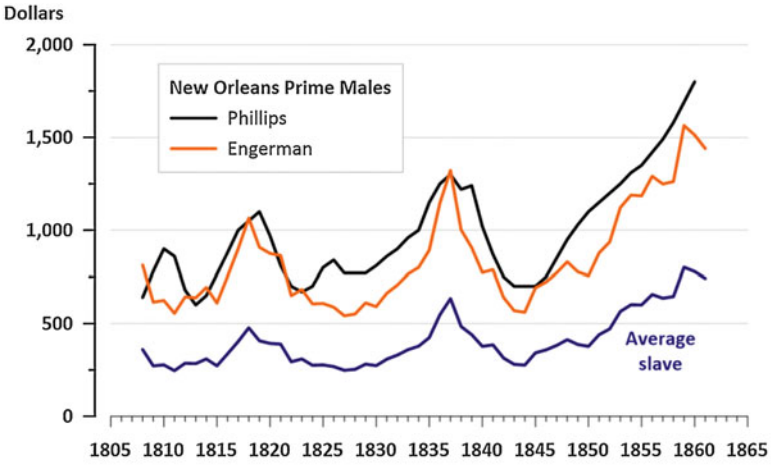
in 1905 in *Political Science Quarterly.* Phillips was born in Georgia in 1877. Although he was a life-long Southerner, he taught history at the University of Wisconsin from 1902 to 1908. Later, he joined the faculty at the University of Michigan and then at Yale University. Phillips was a racist, an apologist for slavery, and (in his time) a distinguished professional historian. Following publication, he emerged as the leader of a now-discredited academic campaign designed to excuse slavery as a positive good, portray the enslaved as largely contented, and celebrate the Confederate cause as noble. Phillips' belief in the Lost Cause was so strong it led him to change his given name from Ulysses to Ulrich.

Phillips' principal thesis was that slavery had become an economic burden to the last generation of antebellum planters. Based on his prejudice rather than evidence, Phillips believed that black people were inherently "unintelligent" and were incompetent laborers except under strict guidance and close supervision. Owning slaves, he suggested, had become a losing business when the price of slaves, driven by speculation, rose rapidly after 1845.[1] The failing system was nevertheless maintained and defended as a means to preserve racial peace. There was a cultural motivation behind the reluctance to end slavery as well. Slavery, in Philips' eyes, was a civilizing institution made necessary to kept black peoples' "savage instincts from breaking forth" (Phillips 1905: 274–275, 259). These ideas may seem absurd – shocking – today, but even as late as the mid-1950s, mainstream historians continued to portray black people as simple and carefree and slavery as a benign institution needed to civilize them and maintain social control.

Conrad and Meyer did not attempt to address the questions of the aptitude of black workers or the civilizing influence of the plantation. Instead, they focused on calculating the profitability of slavery. The only quantitative evidence presented by Phillips was the time series of slave prices plotted with a black line in Fig. 1.[2] Conrad and Meyer, adopting a standard econometric technique, demonstrated that the rising price of slaves was insufficient evidence to conclude that cotton growing was unprofitable. They specified two production functions specifying cotton output as a function of the plantation's inputs of capital, land, and out-of-pocket spending for food, clothing, medical care, taxes, overseers' wages, and cotton marketing expenses. One production function defined the output of cotton produced by male prime field hands. The second function defined the joint outputs of cotton and slave children produced by enslaved women. To put quantitative meat on these bones, Conrad and Meyer assembled data on slave prices (using Phillips' own data), the prices of cotton and land, and the out-of-pocket costs. Combined with estimates of yields per hand, "saleable children" per woman, and longevity (expectation of life at birth), they simulated the annual flow of revenue from an investment in slaves

[1]Phillips elaborated on his 1905 argument in Chapter 19 of *American Negro Slavery* (1918).

[2]Philips published a chart of prices (1918: 371) that he later revised (1929: 177). Figure 1 is based on visual inspection of the revised chart by Conrad and Meyer (1958: Table 17, p. 117). Time series on typical slave prices have since been further refined. Two alternatives to Phillips's series are also plotted in Fig. 1. For these data and a discussion of the sources and methods used to assemble them, see Ransom and Sutch (1988: appendix) and Engerman et al. (2006: Table Bb209–214).



Sources: *New Orleans Prime Males*: Phillips 1929: 177; Engerman, Sutch, and Wright 2006: series Bb210. *Average slave*: Ransom and Sutch 1988: table A-1 (column 4), pp. 150-151.

Fig. 1 Slave prices – prime male field hands sold in New Orleans and a U.S. average of all slaves: 1808–1861

separately for males and females (Conrad and Meyer 1958: Tables 9, 10, and 11). They calculated the rate of return from these hypothetical purchases and compared those numbers to the yields on various alternative capital investments in the American economy.[3]

An important feature of Conrad and Meyer's approach recognized that each of the variables and parameters needed for their calculation varied over some range, but remained interrelated. Land that generated higher yields of cotton per acre would rent for more than poor land. Cotton prices varied from year to year depending on the yields. Conrad and Meyer made calculations for 12 plausible cases involving different combinations of capital outlays, yields per hand, slave prices, and farm prices for cotton. Their estimates for the rate of return on an investment in males spanned a range: from 4.5% to 6.5% for the typical case, as high as 8% for somewhat better land, and over 10% for the very best land on Mississippi alluvium. For women their estimates ranged from 7% to 8%. Since alternative capital investments ranged from 6% to 8% or below, they concluded that slavery was profitable – as profitable as plausible alternative undertakings.[4]

[3]Of course, these calculations did not measure the actual profitability of a would-be planter who purchased slaves in the mid-1840s or after. With the Civil War and the abolition of slavery, slave owners lost the stream of income from owning slaves.

[4]Conrad and Meyer were not the first to recognize the profitability of slavery. Historians such as Lewis Gray (1933), Thomas Govan (1942), and Kenneth Stamp (1956) had reached the same conclusion. Their work on this and other features of the slave economy had undermined the Phillips thesis to such an extent that it toppled with the final blow delivered by Conrad and Meyer.

Conrad and Meyer demonstrated that more evidence than the rising price of slaves had to be considered to establish, as Phillips thought he had, that owning slaves was an unprofitable undertaking. An important point that Phillips had missed was that the price of a slave woman was justified only in part by the yields of cotton she could help produce but also by the value of the children born to her. This is true whether the children were sold when they reached young adulthood or were kept in the owner's plantation. In the latter case, the market value of the owner's holdings – his "portfolio" of enslaved black people – would rise as the enslaved population grew.

Conrad and Meyer concluded that slavery was profitable throughout the South, "the continuing demand for labor in the Cotton Belt ensuring returns to the breeding operation on the less productive land in the seaboard and border states" (Conrad and Meyer 1958: 121). Phillips had ignored the increase of the enslaved population because he was blinded by his insistence that slaves were not bred for sale (Phillips 1918).

Following Conrad and Meyer

The work on the profitability of slavery begun with Conrad and Meyer soon inspired critics and imitators interested in testing the reliability of their findings, thus underscoring cliometricians' self-image as practitioners of a true science. New data on the underlying variables and parameters were assembled, and alternative specifications for the profit simulations were proposed. This effort at reproducibility supported the conclusion that slavery was profitable. If anything Conrad and Meyer probably underestimated the rate of return from an investment in male slaves. Table 1 presents some of the alternatives reported in the literature. From today's vantage point, the best guess for the midrange of profits is between 6% and 8% as an average for the South as a whole.[5]

An oft repeated claim of the Lost Cause historians was that the Civil War was not about slavery but about states' rights, or the tariff, or something else. A key argument supporting this claim was the proposition, best articulated by Charles Ramsdell, that slavery was unsustainable and would have died out on its own in "a little while – perhaps a generation, probably less" (Ramsdell 1929: 171). The Civil War, lamented Ramsdell, was tragically unnecessary.[6]

What seems surprising in retrospect is that so much effort had been invested in establishing the profitability of slavery when all contributors to the cliometric research on the topic agreed from the beginning that profitability per se was irrelevant to the long-run viability of slavery. Viability was the key issue raised by Phillips and

[5]The most prominent cliometric contribution to argue otherwise was offered by Edward Saraydar who recalculated the Conrad and Meyer estimate for prime field hands and arrived at a rate of return between zero and 1.5% (Saraydar 1964: Table 1). For a number of reasons discussed by Sutch (1965), Conrad (in Conrad et al. 1967), and Fogel and Engerman (1971b), his results have been rejected (but see Saraydar 1965).

[6]For a discussion of the causes of the Civil War, see Ransom and Sutch (2001).

Table 1 Alternative estimates of the rate of return to slavery.

Average for the South, various years

	Date	Rate (%)	Source reference
Conrad and Meyer [1958]			
prime field hands[a]	1846-1850	4.5-6.5	Table 9: p. 107
prime field wench	1846-1850	7.1-8.1	P. 109
Evans [1962]	1846-1850	12.6-17.0	Table 21, p. 217
	1856-1860	9.5-10.3	
Sutch [1965]	1849	5.7-6.4	Table 7, p. 376
	1859	5.8-6.3	
Foust and Swan [1970]	1849	9.3	Table 6, p. 55
	1859	6.9	
Fogel and Engerman [1974]	1860	10	Volume 2: p. 78
Ransom and Sutch [1977]	1859	6.3-8.0	Pp. 212-214

[a]Fogel and Engerman present evidence that Conrad and Meyer underestimated the productivity of prime-age males (1971: 327–328)

Ramsdell. Economic theory suggests capital values (slave prices) would adjust to expected returns. Thus, profitability was "more or less guaranteed," as Conrad and Meyer noted, "if the proper market mechanisms existed" (Conrad and Meyer 1958: 110). If Conrad and Meyer's calculations were wrong and cotton production employing enslaved men and women had become unprofitable, it would not mean that slavery was doomed. If no alternative uses for the labor of enslaved workers could be found, the price of slaves, the price of land, or both would have fallen until owning slaves and growing cotton once again became profitable. Slavery would always be able to compete with free labor through adjustments in the price of slaves (Evans 1962; Sutch 1965; Foust and Swan 1970; Yasuba 1961; Fogel and Engerman 1971b).

Despite the asserted irrelevancy of the profitability research, the initial foray by cliometricians into the economic history of slavery precipitated the collapse of the entire Phillips edifice. It was, of course, ripe for replacement based as it was on the presumed racial inferiority of black people and a whitewashed romanticized portrayal of plantation life.[7] Meanwhile, scholarly attention shifted to several related issues relevant to the economics of slavery: self-sufficiency, slave breeding, and the economic development of the antebellum South.

[7]An excellent review of the profitability and viability issues raised by Conrad and Meyer is provided by Hugh Aitken's commentary in his book, *Did Slavery Pay?* (1971). Aitken reprints the major contributions including those of Phillips, Conrad and Meyer, Yasuba, Evans, Saraydar, and Sutch.

Self-Sufficiency of the Plantation

Douglass North formulated an elegant model of antebellum economic development that proved influential for understanding the economics of slavery (North 1961). Starting from a theory of regional specialization and export-led growth he had previously formulated, North considered the American economy as if composed of three regional economies. The East specialized in the export of manufactured products (textiles, boots and shoes, locks and clocks, books and combs) as well as financial and commercial services. The West specialized in the export of wheat flour and corn (some considerable fraction of which was transported in the form of corn whiskey and salt pork). The South specialized in the export of cotton, sugar, and other staples produced by an enslaved labor force. Each region's economic prosperity, North argued, was tied to the success of its export trade and hence to the demand from other regions for the products of their predominant industry. The staple South imported food stuffs from the agricultural West to provision her enslaved population and relied on the industrial East for cheap cloth and shoes to clothe and shod the slaves as well as for banking, credit, and factorage. The South's cotton was sold to the East where it became raw input for the booming textile industry. The East was also a major consumer of Western agricultural products. The West imported manufactures and services from the East. The three interdependent regions were linked by these interregional trade flows, each region pursued its own comparative advantage, and each was more productive than it might have been if walled off and forced to be self-sufficient. The entire national economy, according to North's export-led model, was propelled throughout the antebellum era by the expanding exports of cotton to Europe during the era when England set out to clothe the world with cotton rather than animal skins.

Douglass North's model focused attention on a weak evidential link in Conrad and Meyer's calculations. Cotton output per slave would depend in part on how self-sufficient in food the typical plantation was. Self-sufficiency would imply that some labor effort had to be diverted from the staple crops to food production and home manufacture. North accepted the traditional view that the South depended upon the West "for a *large part* of its food supply and on the East for *the bulk* of its manufactured goods and *very largely* for the conduct of its commerce and banking."[8] The highlighted phrases suggest minimal self-sufficiency, at least on the large plantations. But, whatever the extent of the South's dependence on Western food imports, there is an implication for the profitability calculations. An important magnitude required by the production function approach is the magnitude of the out-of-pocket cost for food and clothing. That cost would vary widely depending on the degree of self-sufficiency. Conrad and Meyer considered three alternatives. One case imagined that plantations were almost completely self-sufficient with the annual cost of purchases required for maintaining a prime field hand estimated between $2.50 and $3.46. But that low cost would have come at the expense of a lower

[8]The quotation, with my emphasis added, is actually from Louis Schmidt. North quotes this passage with approval (Schmidt 1939: 820; North 1961: 103).

production of cotton per hand as labor had to be diverted from cotton to food production. Another possibility was that the out-of-pocket costs were between $7 and $10 "where some ready-made clothing and meat, fish, and other food 'delicacies' were purchased." And, if all provisions were purchased, the cash cost might be as high as $25 to $40 (Conrad and Meyer 1958: Table 5). In that last case, the output of cotton per hand could be maximized. It is important to specify the typical degree of self-sufficiency correctly and then to match that with the appropriate yield of cotton per hand. Conrad and Meyer *assumed* a cost of provisions in the middle range of $7 to $10 and then claimed the typical output would range between 3.5 and 4 bales of cotton per slave per year. However, they offered no evidence to justify the relevance of that yield to the assumed cost of purchased food and clothing. North's work was troubling since it implied the costs for food and clothing imported from outside the South should be higher, perhaps much higher, than the $7–$10 claimed.

Albert Fishlow reached a different conclusion about the magnitude of the South's imports from the West. Assuming that virtually all imports of Western products to the South traveled down river to New Orleans, he calculated from port receipts that "imports were truly minute compared with the [South's] production of foodstuffs." "The South was neither a major market for Western produce nor in dire need of imported foodstuffs" (Fishlow 1964: 352, 357).

Fishlow was challenged by Robert Fogel who countered with the suggestion that "a sizable" volume of Western products was shipped through New York and Baltimore and then down the coast to the South Atlantic states. He had no direct evidence of this alternative route but made rough calculations of the 1860 pork and beef consumption in the Southern seaboard states that implied a substantial deficit when compared to domestic production (Fogel 1964). The exchange between Fishlow and Fogel reached a stalemate however when it bogged down in a disagreement about the relative slaughter weights of Western and Southern hogs and cattle. Furthermore, the relevance of interregional shipments to the question of the plantation's self-sufficiency was not assured given the possibility that large plantations might have been supplied by local, small-scale, non-slave-owning farmers who specialized in grain and meat production.

The debate could not be resolved without new research. This had to wait several years until William Parker and Robert Gallman drew a statistical sample of 5,229 Cotton South farms and plantations from the 1860 census of agriculture. The census enumerators' manuscript returns from these agricultural operations were linked to each farm owner's returns from the population census and also to the slave census returns linked to the owners. The result was a cross-section micro-data sample. William Parker accurately described it as "the best-selected, most comprehensive, and most carefully processed body of statistical evidence which historians have ever employed in the study of history" (Parker 1970a: 2). Immediately the Parker-Gallman sample was applied to address the question: Did cotton farms grow their own food? The answer was "yes" (Gallman 1970; Battalio and Kagel 1970). The sample allowed the question to be addressed not in the aggregate, but farm-by-farm and plantation-by-plantation. Figure 2 presents Gallman's estimates of grain output and grain requirements by farm size. Grain is expressed in corn-equivalent bushels.

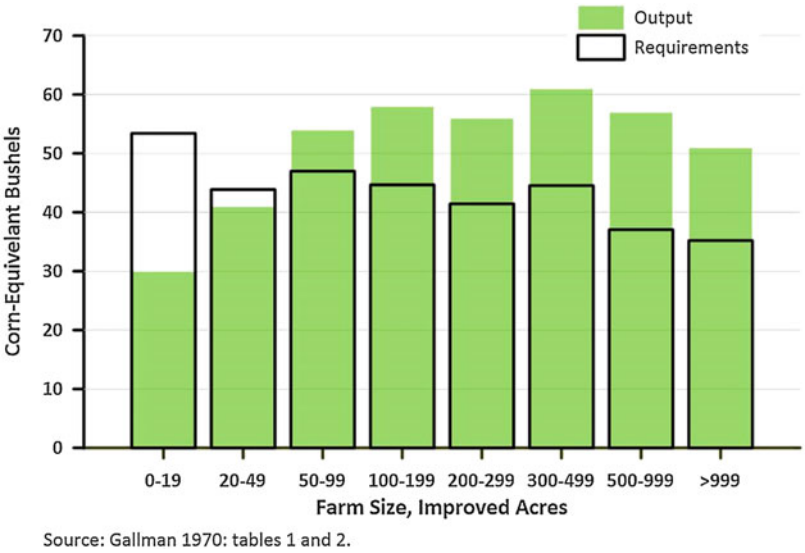


Fig. 2 Per Capia production of grain, Parker-Gallman sample, 1860

This measure aggregates corn, wheat, rye, barley, buckwheat, cow peas, and beans weighted by their nutritional contents. Per capita production (represented by the green bars) was significantly higher on the large farms than on small ones. There is no evidence in these findings that small farms were able to supply grains to large plantations.

Grain requirements were estimated by Gallman in the same units by aggregating estimated on-farm consumption by the enslaved and free populations, the work animals, cows, other cattle, sheep, and young animals – but not swine. Allowance was made for seed requirements, poultry, and on-farm waste. To avoid exaggerating self-sufficiency, Gallman made what he considered to be generous estimates of slave ratios – "more likely to be too high than too low" (Gallman 1970: 12–19). As illustrated in Fig. 2, farms with 50 acres or more of improved acreage had an estimated output in excess of consumption requirements. The residual was sizable. Gallman recognized that these residuals were likely converted to pork by feeding corn and other grains to swine. He assumed that it took 10 bushels of corn to produce 100 lbs of pork arguing that this would likely underestimate pork production. After considering whether his estimates of the pork produced were reasonable given the count of swine reported in the census and the extensive data he assembled on the slaughter weight of hogs, Gallman stated,

> we can have some assurance that the procedures by which claims on grain and output of pork were derived do not *underestimate* plantation demands nor *overstate* plantation production capacities. That is, the check suggests that my original object – to give the self-sufficiency hypothesis a strong test – has been met. (Gallman 1970: 16, emphasis in the original)

After estimating the quantities of meat consumed by both the enslaved and the free populations living on the farms, Gallman concluded:

> The surplus of all the farms in the cotton South (which the sample represents) would have been large enough to feed all of the slaves and one-sixth of the [free population] living in the South outside the sample universe. The farms of the cotton South, far from being dependent on external sources of basic foods, were in a position to supply food to outsiders on an impressive scale. (Gallman 1970: 19)

From an analytical point of view, one of the contributions of Conrad and Meyer was to view the slave owner as a businessman, a capitalist in a competitive business who retained slaves to employ them in a profit-making enterprise. The planter sought to engage the slaves at tasks that would earn a sufficient return to justify their price. That large plantations were self-sufficient in food production is not inconsistent with this view or the proposition that the Southern plantation had a strong comparative advantage in cotton production. Ralph Anderson and Robert Gallman pointed out that a slave was a form of fixed capital. The owner not only had access to the enslaved person's entire labor but was also responsible for his full maintenance (Anderson and Gallman 1977). The labor requirements of cotton production varied seasonally, heavy during planting season and again during picking season, but relatively light in between requiring only some animal husbandry and cultivating to remove weeds. The slave owner would not want his charges idle during the slack season since he was obliged to pay for their upkeep and since they might prove troublesome if left without work. The principal means by which the slave force was kept busy throughout the year was diversification into corn and pork production. This despite the fact that Southern yields per acre in corn were substantially below the yield achieved in the West (about 12.5 bushels per acre vs. 32.4 in both 1849 and 1859) (Parker and Klein 1966: Table 10). One explanation for the low corn yields per acre was that corn was planted and harvested at dates that did not interfere with those for cotton, but were not optimal for corn. During August when neither crop required much attention, the slaves were set to pulling fodder, stripping the corn stalks of their green leaves. The fodder provided cattle feed, but the practice wounded the growing corn plant and reduced grain yield between 10% and 18% (Ransom and Sutch 1977: 395–396, n62).

Other research projects provided evidence that subregions on the periphery of the South supplied grain and meats to the South's population not living on farms. Kentucky and Texas were both exporters of meat. The Upper South states of Kentucky, Tennessee, Virginia, and North Carolina produced a sizable grain surplus and "engaged in large-scale exportation of both wheat and corn" (Lindstrom 1970: 101). By implicitly redrawing regional borders and defining "The South" more narrowly and the West more broadly than had Douglass North, this work recovered the importance of interregional trade links that he had emphasized.[9]

[9]However, the role played by cotton exports to England in propelling US economic growth has been questioned. Irving Kravis argued that exports were supplementary factors: "handmaidens, not engines, of growth" (Kravis 1972: 405).

The Interstate Slave Trade and Slave Breeding

It is hard to question the widely accepted narrative evidence that the interregional slave trade operated by professional slave traders was substantial. The market functioned well and slave prices were flexibly determined. Slaves could be "reared on the poorest of land and then sold to those owning the best" (Conrad and Meyer 1958: 110). It is also clearly established that the slave traders preferentially selected prime-age slaves for removal to the new South.[10] In other words, the migration of complete plantation populations with their labor force intact including young and old, males and females, infants and children – together with their original owner and his family – was not the only way to transport slaves into the rapidly filling West. In case there was any doubt, Conrad and Meyer specifically addressed "the old historical question about whether the typical Southern gentleman planter could bring himself to indulge in the slave trade" (Meyer and Conrad 1957: 539). They argued that if a reluctance to sell slaves thwarted trade, the age and sex distribution of the enslaved population would be approximately the same in all states. Yet the census revealed that the "buying" states of the South had a larger proportion of slaves in the prime working ages than the "breeding" states (Conrad and Meyer 1958: 114–115 and Table 14).

Richard Sutch carried the argument a step further. He used census data and survival rate methods to calculate the volume of net exports (or imports) for each state by age and sex between 1850 and 1860. The overall export rates for Delaware, the District of Columbia, and Maryland ranged from 20% to 33% of the population. Six other states (in descending order: Kentucky, South Carolina, Virginia, Tennessee, North Carolina, and Georgia) were also exporters. During the final decade of slavery, over 10% of the enslaved population of these states was forced to take part in this remarkably large migration. Figure 3 presents Sutch's estimated exportation rate for each age and sex cohort during the last decade of slavery. This rate is defined as the number of enslaved persons living in the state in 1850 who were removed from the state over the decade as a percentage of the state's 1850 slave population that would have survived to 1860. The markedly higher rate of export for those 20–29 of age in 1860 (approximately aged 15–24 at the time of migration) is strong evidence that many slaves were sold singly into the domestic trade rather than making the move West in the company of their owner and their family members (Sutch 1975a).[11] The precise proportion of slave removals that were transferred by domestic slave traders was debated in the 1970s and 1980s without much progress. However, recent research by Johnathan Pritchett and Michael Tadman seems to have settled the

[10]The narrative literature (as distinct from the cliometric contributions) on the domestic slave trade is enormous. The classic is by Frederick Bancroft (1931). A modern treatment is by Ned Sublette and Constance Sublette (2016). Detailed documentary evidence on the trade that transported enslaved men and women from Maryland to New Orleans is provided by Ralph Clayton (2002). A blend of traditional and cliometric methods is employed in Michael Tadman's study which also provides detail on exports from South Carolina (1989).

[11]A similar pattern of exports by age was found by Tadman (1989: Fig. 2.4).

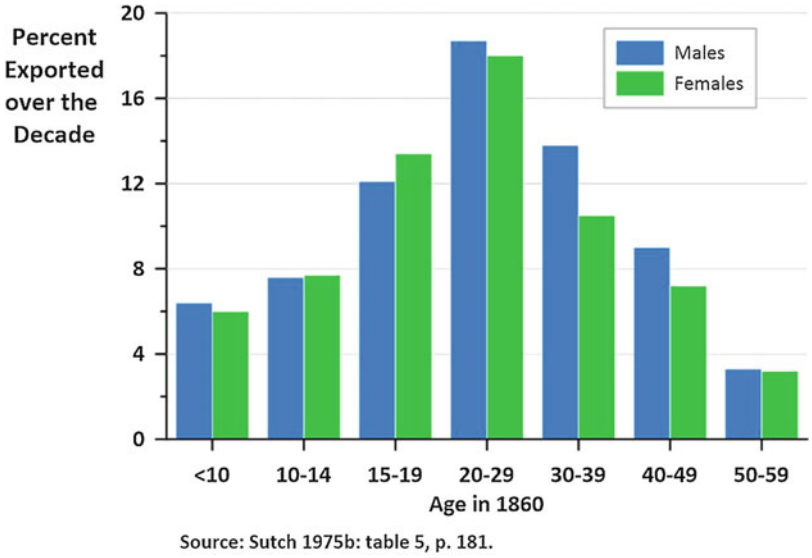


Fig. 3 Exportation rate, by age and sex of slave selling states, 1850–1860

issue with a more sophisticated analysis of the changing age and sex composition of the slave population. Pritchett calculated that approximately 50% of slaves transported across state lines had been sold to traders who resold them (Pritchett 2001). Michael Tadman put the percentage even higher at 60–70% (Tadman 1989: 29–31 and Appendix 3). For regions where exports were unusually large, the percentage sold to traders was probably even higher. Slaves shipped from Baltimore, for example, had been sold by their local owners to traders about 80% of the time (Clayton 2002: 641).

Despite statistics' matter-of-fact quality, these huge percentages considered together with the sheer enormity of the transplanted population focus attention on the misery and terror wrought by the slave trade. The young men and women sold South were separated from their families (parents, wives, partners), friends, and owner, forcibly transported to a foreign environment, sold to a new master not of their choice, and settled alone on an unfamiliar plantation. The numbers are heartbreaking. A teenager in 1850 born and raised in Maryland, for example, had about a 12–16% chance of being sold to a trader before 1860. Maryland was an extreme case. In Virginia, the chances were 11–14%, in North Carolina 9–10%, and those born in South Carolina faced an 8–11% chance.[12] For the exporting states taken together, the median age at sale was 16. Teenage males were 50% more likely than a woman to be sold during the decade. Stephen Crawford extracted reports of the sale of young slaves away from their families

[12]Calculated from the net exportation rates for each state (Sutch 1975a: Appendix Table 5) and the assumed probability of being sold to a trader of 60–70% (Tadman 1989: 31).

from the narratives related by ex-slaves in the 1920s and 1930s. Based on 42 reports, he concluded that "through age sixteen, the slave child faced roughly a 20% chance of being sold away from family" (Crawford 1992: 341–342, Table 11.6).

The teenager caught up in this trade would have been sold, perhaps at auction; jailed in a slave-holding pen in Baltimore, Charlestown, or another port city; then chained and marched into the hold of a ship; transported to New Orleans, a voyage of perhaps 30 days; chained again and marched to another "slave hotel"; sold at auction; and then moved to a new home.[13] Not only was the purchased slave uprooted, but those left behind were also devastated and terrified by their own obvious vulnerability. The children remaining behind could only look forward to facing the same brutal odds as they grew to adulthood. These forced travels West have been called, quite appropriately, the "Second Middle Passage."

No doubt remains about the general picture of a large coerced migration of enslaved black people from the "exporting states" to the "importing states" and the significant role played by professional traders in facilitating the movement. There remains, however, some concern about how to label things. Conrad and Meyer called importing regions the "buying states" and the exporting regions "breeding states." Although they were adopting established terminology, the term "breeding" seems unfortunate in retrospect. This emotionally charged term offended some, angered others, and led several less-than-careful readers to enormous misunderstandings about what practices constituted slave breeding. Conrad and Meyer's insight was that the enslaved were not only workers but also capital and, in the case of every child born to an enslaved mother, they were also product – "an intermediate good" in Conrad and Meyer's words. "Whether systematically bred or not, the natural increase of the slave force was an important, probably the most important, product of the more exhausted soil of the Old South" (Conrad and Meyer 1958: 96, 113–114).

Richard Sutch sought to bring some quantitative evidence to bear on the "breeding hypothesis." This is the idea that the existence of the interregional slave trade and the structure of slave prices established in that market gave slave owners an economic incentive to increase the number of children born on their plantations. The hypothesis entails the twin propositions that,

1. The increasing slave population was a by-product of the agricultural operation – not the main objective; and
2. The slave owner's incentive was to increase the number of children born for a given investment in working adults – not to maximize female fertility (children born per women of child-bearing age).

Female fertility in the antebellum South was astonishingly high. Sutch used the Parker-Gallman sample to estimate fertility on slave farms. The number of children, 0–14, per thousand woman-years of prime fertility experience was 295: 323 in the

[13]For a description of slave trading at the origin of the voyage, see Ralph Clayton's discussion of Baltimore (2002). For the receiving end, consult Walter Johnson's discussion of the New Orleans market (1999). Johnson also provides a narrative description of the forced journey West.

exporting states and 260 in the receiving states (Sutch 1975a: Table 10). In South Carolina where the calculated rate was highest, it reached 355. That would be close to the apparent biological maximum of 371 observed in the population with the highest gross reproduction rate ever recorded: the experience of the women of the Cocos-Keeling Islands born between 1873 and 1927.[14] For the exporting states, these numbers imply that on average, seven surviving children would be born during a woman's reproductive life. To reach levels of female fertility, this high would require early marriage, short lactation periods, and no economic pressure to restrain population growth.

This evidence by itself does not discriminate between two possibilities: that the high fertility on the plantations, like that on the Cocos Islands, was the natural result of ample resources, good health, and the early timing of first births or evidence that the planters were interfering in the sexual life of their slaves by encouraging (coercing?) early marriage and abbreviated breast feeding and discouraging abstinence. If the slave population's reproductive rate was "natural," as some slave owners insisted, how could individual planters increase the number of children born on their plantation when female fertility was already near the maximum? This could be done by skewing the sex ratio of adults toward women of child-bearing age. Since women as well as men worked in the fields, agricultural output could be maintained, while the number of children born *per adult worker* could be increased.

Before the Parker-Gallman microdata was available, the published census results seemed to undercut this possibility. The sex ratios of slaves, 15–39, in the exporting states were not different than that observed in the importing states, and both were almost precisely equal to 1.0. The sample of slave farms, however, told a different story. When farms without any women are excluded, the female-to-male ratio in the exporting states was 1.3 (1.2 in the importing states). Slaves employed in non-agricultural occupations were predominantly males, and farmers who owned only one slave had a revealed preference for males.

Table 2 illustrates the effectiveness of skewed sex ratios on farms with five or more adult women. Since children represent the "output" of a supposed breeding operation and the number of adults represents the "inputs," "productivity," measured as the ratio of children to adults, rises with the ratio of women to men. It is also remarkable that less than 40% of the larger farms in the exporting states had rough gender balance.[15] One-third had a ratio of women to men greater than 1.5; 17% had

[14]The number of years of "prime fertility experience" is an index of potential fertility. Each year of the woman's life during the preceding 15 years is weighted by the reproductive experience of the Cocos Island women at the same age. Sutch also presented calculations to suggest that the differences between exporting and importing states could not be explained by selective migration of women without children (Sutch 1975a: Tables 8–10).

[15]Stanley Engerman suggested in his comments on Sutch that moral influences led this minority of owners to noninterference and encouragement of family stability and that family stability might increase the female fertility rate (Engerman 1975). There is certainly evidence of that, as female fertility was actually higher on the farms with balanced sex ratios (Sutch 1975a: table 12). But the breeding hypothesis proposes that it is not female fertility that would matter, but the reproduction rate of the adult population on plantations, counting women and men together.

Table 2 Sex distribution and child-adult ratios, exporting states, 1860.

Parker-Gallman Sample
Farms with five or more women

R = Number of women per man, 15-44	Ratio of children, 0-14, to adults, 15-44	Farms with five or more women	
		Number	Percent
R > 2.0[a]	1.36	41	16.5
1.5 < R ≤ 2.0	1.27	42	16.9
1.1 < R ≤ 1.5	1.14	66	26.6
R ≤ 1.1	1.00	99	39.9

Source: Sutch 1975a, Table 12, p. 193
[a]Includes farms with no men.

two or more women per man. Sutch raised the possibility that polygamy or promiscuity might help explain the high birth rates on these farms despite the gender imbalance. On the other hand, enslaved husbands might have resided off the plantation on neighboring farms. Fogel and Engerman asserted that this was "quite common" (1992b: 462). However, Ann Patton Malone's thorough study of the slave family and household structure in Louisiana found little evidence of off-plantation marriages. She suggests that the frequency of "abroad marriages" has been exaggerated, "perhaps out of an overabsorption with … heralding the standard nuclear family" (Malone 1992: 227–228, 262–263). In any case, the need to sell left many individual farms with more adult women than men.

The large number of children born on plantations with substantially more women than men raises the possibility that some of the women were impregnated by men other than their husbands, with or without their consent. Given the racial and sexual domination inherent in Southern society, enslaved women may have felt they had little choice but to submit to intercourse, particularly when approached by a white man (Steckel 1980; Bodenhorn 2015). There is some limited evidence that rape by overseers was not uncommon (Malone 1992: 221–222). These acts would be crimes of passion, not necessarily motivated by the supposedly dispassionate calculations of a slave breeder. However, it is worth noticing that the enslaved woman's best insurance against sexual abuse by a slave breeder was to marry early and produce many children within that marriage.

Economic Growth and Manufacturing Development of the South

That slavery was profitable and viable did not mean that slavery was compatible with economic growth or development. This point was raised by Douglas Dowd in his comment on Conrad and Meyer, and later became the subject of a famous panel

discussion on "Slavery as an Obstacle to Economic Growth" held at the Economic History Association meetings in 1967 (Conrad et al. 1967).[16] As Dowd put it, "slavery and all that it entailed was fundamental in inhibiting industrial capitalism – and economic growth – in the South" (Dowd 1958: 441). Indeed, there was little doubt in the 1960s that the antebellum Southern economy was backward and lagged behind the North. Northern and Southern observers of the time, abolitionists and slaveholders, all noted the contrast and all agreed that slavery was to blame (for a review, see Harold Woodman 1963). The cliometric literature on the topic opened with a challenge to this view by Stanley Engerman (1967).

For the purposes of his reinterpretation, Engerman reworked regional personal income relatives estimated by Richard Easterlin to produce rough estimates of personal income per capita (counting enslaved persons in the population total). His results are presented in Table 3.[17] Engerman directed attention to the growth rates he calculated for the global South and North. Between 1840 and 1860, by these measures, the Southern economy grew at the average of 1.45%. This is faster than the growth rate of 1.30 for the North.[18] How could the observations of contemporary witnesses and the calculations of cliometrics be so at odds? The answer is the nature of the economic growth experienced by the South was quite unusual. It was not really noticeable locally; it was only manifest at the interregional level. Engerman calculated the rate of economic growth for the South viewed as a whole, while the contemporaries were commenting on what could be seen from the perspective of a local observer ("on the ground," so to speak). Note that the calculated rate of growth for the South (1.45%) is higher than the rate of growth in each of the South's three subregions measured separately (1.21, 1.28, and 0.82%).[19] This counterintuitive phenomenon arises because the westward migration was away from regions of poor soil and relatively low per capita income to regions with rich soils and relatively high income. Much of the apparent antebellum Southern growth came from

[16]Panel members: Alfred H. Conrad, Douglas Dowd, Stanley Engerman, Eli Ginzberg, Charles Kelso, John R. Meyer, Harry N. Scheiber, and Richard Sutch. For a colorful, emotional, and hyperbolic description of the discussion that ensued, see Robert Fogel and Stanley Engerman (1974, volume 2: 11–19).

[17]Table 3 is based on numbers reported by Easterlin (1961: table 1), Engerman (1971), and Fogel and Engerman (1971b: 335) and calculations by Ransom and Sutch (1977). See Ransom and Sutch for details. The states included in each region by Easterlin are the following: *New England*, Connecticut, Rhode Island, Massachusetts, Vermont, New Hampshire, and Maine; *Middle Atlantic*, New York, Pennsylvania, New Jersey, Maryland, and Delaware; *East North Central*, Ohio, Michigan, Indiana, Illinois, and Wisconsin; and *West North Central*, Iowa and Missouri. In 1860, Minnesota, Nebraska, and Kansas are also included: *South Atlantic*, Virginia, (including present-day West Virginia), North Carolina, South Carolina, Georgia, and Florida; *East South Central*, Kentucky, Tennessee, Alabama, and Mississippi; and *West South Central,* Arkansas and Louisiana.

[18]Engerman, departing from Easterlin, included Texas in the 1860 definition of the West South Central region. This adjustment boosted the South's growth rate to 1.67% (Fogel and Engerman 1971b: 335). Table 3 reports the figures for the South defined to exclude Texas. Texas was an independent republic in 1840. It became a state in 1845.

[19]Engerman's calculations probably exaggerate the progress made in the South's subregions between 1840 and 1860. They are primarily based on the census returns for the crops of 1839 and 1859, but 1859 was an unusually good year for cotton production.

Table 3 Personal income per capita by geographic divisions, 1840 and 1860.

1860 prices

Geograhic division	Personal income ($) 1840	Personal income ($) 1860	Annual rate of growth (%)	Population weights (US=1) 1840	Population weights (US=1) 1860
North	**109**	**141**	**1.30**	**0.62**	**0.65**
New England	126	186	1.97	0.13	0.10
Middle Atlantic	130	178	1.58	0.30	0.26
East North Central	64	90	1.72	0.17	0.22
West North Central	72	86	0.89	0.02	0.07
South	**72**	**96**	**1.45**	**0.38**	**0.33**
South Atlantic	66	84	1.21	0.20	0.14
East South Central	69	89	1.28	0.15	0.13
West South Central[a]	140	165	0.82	0.03	0.06

Source: Ransom and Sutch 1977: Table F.6, p. 267. Population weights from Easterlin 1961: Table 3, p. 535.
[a]Excludes Texas. See footnote 18

territorial expansion onto more productive lands. This extensive growth would not be relevant to anyone who did not move West.

Still, Engerman's point is well taken. The South's slave economy was not stagnant. The transformations however were largely driven by the interregional slave trade and planter migrations. Physical productivity in cotton production rose throughout the antebellum period (Conrad and Meyer 1958). This was probably true at the local level, even on the "exhausted" soils of the Atlantic coast states. The advance is attributed to improvements in transportation and marketing, the application of guano as a fertilizer, the adoption of new cotton varieties, and improvements in the machinery for ginning and baling cotton. The real price paid for cotton was also rising as advances in technology for manufacturing cotton thread and textiles enabled manufactures to make more effective use of the South's upland cotton varieties. But the growth in the South Atlantic and East South Central regions was modest in comparison to the growth in the economies of New England, the Middle Atlantic states, and the East North Central region.[20] While the planters could reap the benefits of higher prices and the innovations in marketing and production, the slaves did not. In the North, by contrast, the local economies were markedly dynamic, growth was indigenous, intensive, and the benefits were widespread. Northern

[20]The frontier regions, the West North Central and West South Central, were thinly populated in 1840 with only 2 and 3% of the population, respectively.

agriculture was transformed, local manufacturing appeared, young women left farms to work in factories, cities grew, and a flood of "Yankee gadgets" were filed with the Patent Office.[21]

A westward migration was common to the histories of both the North and the South. In the North, however, the migration was from the high-income regions in the East (New England and the Middle Atlantic states) to regions in the West (primarily the East North Central division) with lower per capita incomes. Northern migration was not produced by individuals giving up high-paying jobs for low-income ones. Rather it was primarily a migration out of low-productivity agriculture to the more fertile agriculture of the West. Agricultural income per worker was 25% higher in the Old North West than New England in 1840 and 45% higher in 1860 (Easterlin 1974: Table B.1). The figures in Table 3 support the notion that the local Southern slave economies were indeed lagging behind the industrial economies of New England and the Mid-Atlantic and behind the new agricultural economy of the East North Central states.[22]

But why should this be? The South's relative lack of manufacturing, the absence of a connected network of towns and villages, and the nonexistent provision of public education for white children and laws actually prohibiting the education of slaves are acknowledged features of the Southern landscape.[23] If growth is equated with manufacturing development, urbanization, and education, there was obvious work for cliometricians. Why did the South miss out?

The study of Southern manufacturing was initiated by Fred Bateman, James Foust, and Thomas Weiss. Based on a sample of manufacturing firms drawn from the manuscript census of 1850 and 1860, they report two findings which, taken together, led them to a surprising conclusion. For the South as a whole, there was a lack of participation by planters in manufacturing enterprises. As the wealthiest members of Southern society, slave owners might be expected to have stepped into the role of local entrepreneurs and venture capitalists. But only 6% of planters engaged in manufacturing, and yet their investments comprised around 25% of all manufacturing capital. "This suggests the South suffered a very conventional but perhaps economically rational fate – a limited transfer of resources from agriculture to manufacturing" (Bateman et al. 1974: 297). The low participation is particularly surprising in light of calculations establishing that the rate of return was two to three

[21]Technological creativity affected every facet of Northern life. On the transformation of Northern agriculture, see Jeremy Atack and Fred Bateman (1987). On the "rain of gadgets" and the impact on agriculture, see Peter McClelland (1997).

[22]A recent re-examination of regional growth in the antebellum period over a longer period confirms that the South lagged far behind the North. Between the benchmark years of 1800 and 1860, Peter Lindert and Jeffrey Williamson calculate a growth rate for the South Atlantic region of 0.9% per annum (that is below the rate they consider required for "modern economic growth"). By comparison, New England grew at the rate of 1.94%, and the Middle Atlantic states grew at 1.66% (Lindert and Williamson 2016: 101–107 and Figs. 5.1 and 5.2).

[23]Despite law and custom, a small minority of slaves did learn to read and write. A non-cliometric history by Heather Andrea Williams tells the story (2003).

times higher in Southern manufacturing than in Southern agriculture. Clearly, the region did not live up to its economic potential. The authors describe the situation as one of entrepreneurial failure, "investor irrationality." Manufacturing opportunities would have been more remunerative, yet planters were unwisely plowing their earnings back into cotton agriculture and the purchase of slaves. "While the South was not nearly so devoid of industry as conventionally believed, it no doubt could have done better. That it did not, largely reflects upon the behavior of Southern investors" (Bateman and Weiss 1976: Table 13, pp. 26 and 39).[24]

The scarcity of manufacturing helps explain the South's low urbanization. Factory workers needed to live near factories, and the factories needed to locate at a power source (a mill stream perhaps). In the antebellum North East and North West, urban places grew up around these nonagricultural sources of employment. And these urban enclaves became self-propelling. They fostered a business culture. When technology changed and opportunities emerged, homegrown entrepreneurs and lenders of capital were willing to assume the risks (Parker 1970b). Without urban centers, the South lacked self-conscious businessmen and lenders that would provide the dynamics to industrialize.

Claudia Goldin tackled the topic of slavery in the cities. Her concern, however, was not so much with the lack of urbanization as with the relative lack of slaves in the few cities that did exist. In 1850, 10% of the white population lived in the urban areas of the South, but only 4% of the slaves were city residents (Goldin 1976: Table 1). Goldin also established that urban slavery was on the decline during the 1850s. According to her statistical analysis of the supply and demand for urban slaves, this was almost entirely explained by the "rapid increases in the price of slaves in general, which led urban owners to cash in on their capital gains" (Goldin 1975: 449).

Leaving aside the South's woeful investments in basic education, which in the case of the enslaved black population was clearly the consequence of slavery, the initial attempts to explain low levels of investment in manufacturing did not directly implicate slavery as the cause. Instead, they attributed the failure of industrialization to the shortage of entrepreneurial imagination. Extending this discussion into questions of distinctive regional character traits, planter irrationality, risk aversion, business culture, and the like would take this essay too far from its assignment. Economic development and its index of success, economic growth, are complex concepts. Questions such as "Why didn't the South grow faster?" and "Why didn't she develop a booming and diverse industrial sector?" require a comparative approach. Engerman set out to compare the growth of the South with that of the North. Bateman and Weiss explicitly compared Southern manufacturing with Northern industrialization. Goldin compared slavery in the cities with slavery in the countryside. Gavin Wright made yet another comparison. He contrasted the economic progress of the Old South before the war to the progress of the New South following the Civil War. Wright saw a "basic change in the principles and directions of entrepreneurial energies":

[24]Bateman and Weiss collect and summarize their work in a book (1981).

> When slavery was abolished, investment strategies, entrepreneurial designs, and political schemes whose end purpose was to increase the productivity and value of *land* came to the fore.

The real cause of the lackluster prewar performance, Wright argued, was the lack of saving:

> The level of savings is a function of wealth as well as income. By analogy to the burden of the national debt, capitalizing labor satisfies the desire to accumulate wealth over the life cycle, and hence reduces the savings available for investment in physical capital. This model could well explain the evidence recently presented by economists Fred Bateman and Thomas Weiss that rates of return in antebellum southern manufacturing were generally high, yet failed to attract investment on a large scale. (Wright 1986: 19–20 Emphasis in the original)

Wright cited a paper on slave owners' saving by Ransom and Sutch that relied on Franco Modigliani's life-cycle model of saving and its implication for the burden of the national debt (Ransom and Sutch 1988; Modigliani 1961). Life-cycle models predict the existence of a desired wealth-income ratio that depends upon the worker's expectations about the age profile of income over the life span and the need to build a stock of wealth. The wealth is required both to establish a rainy-day fund and to provide security in old age when earning power is anticipated to be low. Because the saver is indifferent to whether the accumulated wealth is in the form of physical capital or government bonds, an increase in "national debt will tend to displace private tangible capital on a dollar per dollar basis" (Modigliani 1966: 200). A lower stock of physical capital in the economy reduces the flow of output. If you replace the phrase "national debt" with the words "slave capital" in the preceding quotation you have the essence of Ransom and Sutch's point. Slaves were a form of wealth that took the place of manufacturing capital.

The life-cycle model not only helps explain the lack of Southern manufacturing investment, it also offers a partial explanation for the puzzling upward jump in aggregate savings rates that appeared following the Civil War. With emancipation, a significant fraction of the saleable wealth in the South evaporated. The value of slaves in 1860 was nearly 60% of the total capital invested in agriculture and completely overshadowed the physical capital invested in Southern manufacturing (Ransom and Sutch 1977: Table 3.5; 1988, Table A.1). When the wealth represented by slaves disappeared, the wealth-income relationship was catapulted far out of equilibrium. In response the savings rate accelerated (Ransom and Sutch 1988). The quickened pace of capital formation was first brought to the attention of the profession in 1966 when Robert Gallman published his estimates of gross national product for the nineteenth century (Gallman 1966). Capital formation rates rose from approximately 13 to well over 17% of GNP. Subsequent revisions by Gallman of the figures on output and gross saving are presented in Table 4. Physical capital formation jumped from less than 15% to well over 20% according to the revised estimates, and from 16% to 22% using the slave-economy concept of gross national product that includes the value of the increase in slave wealth in aggregate capital formation.

Table 4 Share of gross capital formation in gross national product, United States.

Gallman's estimates and slave-economy GNP
Decade averages, 1839-1888

Decade	Gallman's estimates excluding the increase in slave wealth		Slave-economy concept including the increase in slave wealth
	1966	Revised 2006	
1839-1849	11.5	10.8	14.0
1844-1854	12.9	13.0	14.9
1849-1859	13.3	14.7	16.2
1869-1878	17.4	22.1	22.1
1874-1883	17.3	20.8	20.8
1879-1888	18.9	23.5	23.5

Sources: Gallman's 1966 estimates computed from the underlying sources by Ransom and Sutch (1988: Table 4: 149) using methods described by Gallman (1966: 10–14). Revised Gallman-estimates are calculated as averages of annual data presented in Carter et al. (2006: Tables Ca192–207 (for 1869–1888) and Ca219–232 (for 1839–1858)). The pre-Civil War shares based on the slave-economy concept are calculated from the same source (Table Ca233–240). See Paul Rhode and Richard Sutch for a discussion of the slave-economy gross national product (2006: 16)

Simply put, the slave South did not establish the educational, social, and financial institutions upon which to promote innovation, entrepreneurship, and economic development. The South's impulse to invest and expand the stock of wealth was satisfied by the increase in the slave population. Slavery denied the South even the potential for modern economic growth.[25] Taking the long-term perspective, the South was in peril. Immigrants self-selected the North with its rapidly growing urban centers and her expanding industries as the obvious place to settle. The South provided few jobs for free labor and the entry costs for land and slaves made joining the planter class out of reach for the typical immigrant. As the flow of immigrants swelled the population of the Free States outpacing that of the slave South, the size of the South's congressional representation *vis-à-vis* the North was

[25]A reflection on the economic development of the antebellum South by Gavin Wright takes a refreshing step back from the cliometrics to provide insightful historical and comparative context (Wright 2006).

threatened, thus her insistence on the extension of slavery into the West (Ransom and Sutch 2001).

The Debate over *Time on the Cross*

The rich outpouring of cliometric research on slavery in the 1960s and 1970s was further stimulated by the appearance in 1974 of *Time on the Cross* by Robert Fogel and Stanley Engerman. The authors were intentionally provocative. They announced their work as a radical interpretation of the economics of slavery, reporting the findings of "almost a decade and a half" of methodologically sophisticated research based on "new techniques and hitherto neglected sources" (Fogel and Engerman 1974: volume 1, pp. 4, 226). Chief among the startling findings was the authors' claim that the physical and psychological well-being of American slaves was much greater than previously believed. Slaves were provided not just adequate food, clothing, and shelter but a material standard of living which compared favorably with that of free American workers of the time. Slave owners used cash rewards and other positive incentives rather than physical punishment to motivate a high level of work effort. Humane treatment secured the willing cooperation of the slaves. Slave breeding and sexual exploitation were abolitionist "myths." The trade and traffic in human beings did not break many family units. Large plantations were more efficient than free farms of the North because gang labor and economies of scale made them super productive.

After more than four decades, it is difficult to recapture the shock that greeted this work. The surprised reaction was in part the consequence of the book's brash style and structural eccentricities. There were two volumes. The primary volume, subtitled *The Economics of American Negro Slavery*, was intended for the general reader and employed what Deirdre McCloskey would call a forensic style, a rhetoric more reminiscent of the prosecutor's closing argument than a typical historical narrative with careful footnotes and references to primary sources (McCloskey 1985). While it claimed to be rigorously scientific, Volume One lacked the careful methodological description, empirical precision, and caveats necessarily associated with the presentation of novel scientific findings. That material, when provided, was left to Volume Two, subtitled *Evidence and Methods, A Supplement*. Volume Two was a collection of "technical notes" filled with algebraic symbols, intricate cross-references, and terse descriptions of statistical results that would certainly be inaccessible to all but a minuscule few of its general readers. Only someone trained in economic theory, the logic and methods of statistical inference, and cliometric analysis could reliably put the two parts of *Time on the Cross* together.

In a sense, Fogel and Engerman were ahead of their time. They wished to convince a large audience of lay readers that "the view that black Americans were without culture, without achievement, and without development" was false (Fogel and Engerman 1974: volume 1, p. 258) while at the same time persuading their scientific colleagues of the validity of their findings. Their effort to simultaneously address two audiences proved to be a failure not only because the physical and

rhetorical separation of findings and meaning was seen as an attempt to overawe readers but because the inferences and many of the technical details had not been vetted through peer review.

The initial shock over the two volumes' intimidating format might have been a temporary distraction had the authors delivered what they promised – novel interpretations derived from reliable findings based on hard data, sophisticated mathematics, and computational technology with the power to digest massive amounts of numerical data. Subsequent scholarship, however, would reveal that the findings were not reliable. When quantitative historians with the necessary experience and credentials began the process of replicating – or attempting to replicate – the cliometrics, they could not, for the most part, do so.[26] The rejections were many and came from multiple quarters. Paul David, Herbert Gutman, Richard Sutch, Peter Temin, and Gavin Wright prepared a compilation of the cliometric corrections and logical critiques of *Time on the Cross*, which they published as a book, *Reckoning with Slavery,* complete with a concordance. The conclusion was blunt:

> We have attempted, collaboratively, to reproduce every important statistical manipulation, check every significant citation, reexamine every striking quotation, rethink every critical inference, and question every major conclusion in Fogel and Engerman's book. To our surprise and dismay, we have found that *Time on the Cross* is full of errors. The book embraces errors of mathematics, disregards standard principles of statistical inference, mis-cites sources, takes quotations out of context, distorts the views and findings of other historians and economists, and relies upon dubious and largely unexplicated models of market behavior, economic dynamics, socialization, sexual behavior, fertility determination, and genetics (to name some).
>
> No work of scholarship, and certainly no work which undertakes to cover so broad a canvas, is unblemished by some errors. *Time on the Cross,* however, is simply shot through with egregious errors. Even more dismaying is the consistent tendency in the mistakes we have uncovered: all seem to work in favor of the particular "radical reinterpretation" of the institution of slavery that has been put forth by Fogel and Engerman. When the faults are corrected and the evidence is re-examined, every striking assertion made in *Time on the Cross* is cast into doubt. The effect in many instances is to restore and reinforce more orthodox conclusions hitherto shared by conventional and quantitatively oriented students of the peculiar institution. (David et al. 1976: 339–340)

The objections were not confined to errors of factual detail. The interpretations presented in Volume One, the critics charged, did not follow logically from the cliometric analysis; Fogel and Engerman had warned readers that their findings (presented in Volume Two) and their interpretations (in Volume One) do not stand on the same level of reliability. "Interpretation sometimes involves additional data which are quite fragmentary and assumptions which, though they are plausible, cannot be verified at present. Hence, even when readers accept the validity of one or another of the principle findings, they may disagree with the significance that we attach to it" (1974: volume 1, p. 10). The objection of the critics, however, was not to the use of additional assumptions or fragmentary data to broaden a finding into a

[26] I was more than a witness of the verification effort. I was a participant (Sutch 1975b).

conclusion. Rather the charge was that the interpretations presented with great fanfare in Volume One would not be warranted even if the additional assumptions and fragmentary data are accepted and the cliometric findings in Volume Two were valid. The illogic of Volume One became a major issue.

The forensic style Fogel and Engerman adopted for Volume One intended to leave no room for disagreement, but to meet that standard the "corrections" of the traditional characterization of the slave economy must follow directly from the cliometric findings. Unlike a true forensic investigator who examines all the evidence and then builds a case, it appears to me that Fogel and Engerman decided upon their conclusions and then assembled plausible, supportive evidence after the fact. In the process, they left out the logical apparatus that should have connected the interpretation back to the finding.

The Relative Efficiency of Slavery

Fogel long maintained that the most significant contribution of his work on the economics of slavery was the novel demonstration that slavery was efficient; indeed, more efficient than free labor. According to the numbers Fogel and Engerman reported, Southern agriculture was 35% more efficient than Northern farming in 1859 and Southern slave farms were 28% more efficient than Southern free farms. *Time on the Cross* described this result as a paradox. Productive efficiency is an important component of American values and often counts as one of the higher virtues of a capitalist economy.[27] By challenging the long-held and seemingly secure belief that slavery was inefficient without a clarifying caveat, Fogel and Engerman needlessly created the "paradox" seemly suggesting that slavery was virtuous and inherently desirable. Of course, they did not think so. In an epilogue they explicitly denied they were trying to "sell slavery" (Fogel and Engerman 1974: volume 1, p. 258). They acknowledged that the "great power that slavery gave one group of men over another was, in and of itself, sinful." In brief and scattered observations throughout Volume One, they noted that enslaving blacks was also exploitative, expropriative, and racist (Fogel and Engerman 1974: volume 1, p. 159, 144, 153, 215).[28]

[27]The idea that efficiency is unambiguously desirable is a holdover from the Progressive Era. Historian Samuel Haber wrote, "efficient and good came closer to meaning the same thing in these years than in any other period of American history" (1964: ix). President Theodore Roosevelt helped cement this view in a message to Congress, "In this stage of the world's history to be fearless, to be just, and to be efficient are the three great requirements of national life" (1909). Efficiency also is often thought by neoclassical economists to be a consequence of the pursuit of profits by business owners in a competitive capitalist economy.

[28]Fogel later conceded he had not thought deeply about the moral issues "because they seemed so obvious" and that the scattered observations on the morality of slavery in *Time on the Cross* "did not add up to an adequate statement of the problem" (Fogel in Fogel et al. 1989–1992: primary volume 1, pp. 391–393). *Without Consent or Contract* considers the issue at length in an "afterword" [primary volume].

If slavery was more efficient than free labor – as Fogel and Engerman maintained – the real implication should have been allowing humans their own agency, as with free labor, comes at the cost of lost output and must be defended on noneconomic principles (Wright 1978).

On the efficiency of slavery as on other issues, Fogel and Engerman's critics questioned virtually every element of their approach – their definition of efficiency, their measurement of output and labor inputs, the evaluation of farm land, the econometrics of production function estimation, and many technical details of measurement and inference.[29] The multi-round exchange that took place is worth considering despite its failure to reach a general consensus because it illustrates the problem of audience in cliometric research. The prose in Volume One of *Time on the Cross* left the meaning of "efficiency" to the reader's sense of common usage. Yet there are several common meanings. A farm that produces a given physical output of cotton with a minimum of wasted expense can be said to be efficient. This concept is known as "technological efficiency."[30] An individual worker is said to be personally efficient when he or she performs a task in a well-organized and competent way. A productive process is efficient if it avoids the wasteful use of a particular resource, energy for example. This is "resource efficiency." Calculating any one of these concepts does not provide information on the other two.

By not clarifying their definition of efficiency for the lay reader, Fogel and Engerman allowed a major confusion to enter their exposition. The index of efficiency calculated in *Time on the Cross* corresponded to none of the three common-sense definitions, but was actually a fourth concept, "revenue efficiency." Southern farms produced cotton (and corn), while Northern farms produced no cotton (but plenty of corn). Since the outputs of the two regions were different, direct comparison of their physical efficiency is not possible. Fogel and Engerman skirted this apples-and-oranges difficulty by calculating the total value of the crops and livestock products reported by the two regions for the crops of 1859 by evaluating each product using a uniform national price. They measured productivity with a "geometric index of total factor productivity," defined as the ratio of the value of output to a weighted average of the inputs of labor (slave and free), land, and capital. They measured "relative efficiency" by the ratio of the total factor productivity of Southern farms to that of Northern farms (Fogel and Engerman 1974: volume 1, Fig. 42, pp. 192–193). Defined in this way, "productivity" is a measure of *revenue* in 1859, and calculating relative efficiency with this definition of product gives an index of

[29]The list of back-and-forth responses is long. Gavin Wright provides a concise summary of the main issues but remarked that "this issue [of the relative efficiency of slavery] has been scrutinized so extensively, and along so many dimensions, that an attempt at exhaustive review of the debate would be foolhardy" (2006: 94–121).

[30]Technological efficiency is inherently a physical concept. It is used, for example, to compare the performance of one firm relative to another firm when both produce the same identical product, say pencils, measured in physical terms, the number of pencils produced per day. The firm that produces more pencils given the same quantity of each input is said to be the more technologically efficient firm.

"revenue efficiency" (David and Temin 1974: 775–778). Despite what many readers of *Time on the Cross* might have assumed, revenue efficiency is not a measure of physical efficiency nor can it be a standard for assessing resource efficiency, nor can it be taken as a gauge of the workers' proficiency or diligence.

Moreover, in Fogel and Engerman's application, revenue efficiency overstates the superiority of slave plantation agriculture. The crop year 1859 saw unusually high yields of cotton, while at the same, time cotton prices remained relatively stable. Thus, the 1859 revenues from cotton production were extraordinary. Conditions in the North were not unusual; thus, using 1859 outputs introduced a bias into the calculation that favored Southern farms over Northern farms. In fact, "there is no other year which would put the South in a better light in terms of the Fogel-Engerman version of 'efficiency'" (Wright 1976: 313–316; 1979).

When they first presented a relative efficiency calculation in 1971, Fogel and Engerman conjectured that a finding that Southern farms were more efficient than Northern farms might be explained by the "superior entrepreneurship and managerial ability" of Southern farmers (Fogel and Engerman 1971a: 364–365). In *Time on the Cross* they went a step further. To superior management of planters and overseers, they added the "the superior quality of black labor." In *Time on the Cross*, these explanations for the supposed efficiency of slavery were no longer casual conjectures; they were presented as logical inferences from the calculations of revenue productivity (Fogel and Engerman 1974: volume 1, pp. 209–210). This new interpretation then became the central theme of their book, a motif they emphasized throughout from the introduction to the conclusion. The typical field hand "was harder-working and more efficient than his white counterpart" (p. 5). Slaves were "diligent and efficient workers" (p. 263). "All, or nearly all, of the advantage [of plantations] is attributable to the high quality of slave labor, for the main thrust of management was directed at improving the quality of labor" (p. 210).

This inference is not warranted. Paul David and Peter Temin were quick to point out that the relative "quality" of productive factors and the efficiency of alternative production process are two analytically distinct concepts. Even if we accept the validity of the calculations, the comparative analysis of total factor productivity indexes cannot shed light on the comparative performance of the managers and workers operating in different environments (David and Temin 1974, 1979). Fogel and Engerman later abandoned their claims about the quality and diligence of enslaved labor, but continued to defend their efficiency calculations (Fogel and Engerman 1980). They allowed their general audience for Volume One to believe that high revenue efficiency – which is the same as high profitability – was inherently a good thing, both morally acceptable and a felicitous consequence of the capitalistic nature of American slavery.

Fogel and Engerman's critics did not accept the validity of the productivity calculations and their complaints were not restricted to revenue as the measurement of output. In their view, *Time on the Cross* underestimated the labor input on slave farms (David and Temin 1974; Wright 1979). For their estimates of labor, Fogel and Engerman employed the Parker-Gallman sample, which had first been engaged to investigate the issue of self-sufficiency. The age and sex detail available for the slave

population in that database allowed Fogel and Engerman to make a number of adjustments to the labor force estimate not possible with the published census data used in the North-South comparisons (Fogel and Engerman 1974: volume 2). The full extent of these adjustments was not evident until Engerman and John Olsen outlined the "basic procedures" in 1992 (Engerman and Olson 1992). Their description is not entirely clear, but it raises a number of questions. Slave labor was measured in "hand-rating" equivalents (Fogel in Fogel et al. 1989–1992: primary volume 73–74 and Fig. 13; Engerman and Olson 1992: Table 24.1). Where these age and sex adjustments came from is not revealed.[31] In any case, Gavin Wright has shown that the hand-rating equivalents for women greatly underestimated their contribution of labor and thus inflated the total factor productivity of slave farms (Wright 2006).

Critics also questioned Fogel and Engerman's valuation of land. Originally, they measured the input of land by the total acreage in farms (improved plus unimproved) without an adjustment for differences across regions in natural fertility or land use (for crops or pasture) (Fogel and Engerman 1971a). In *Time on the Cross*, a refinement was made to measure land at its cash value (Fogel and Engerman 1974: volume 2). This would be a poor measure in any case since it does not take into account the different proportions of improved and unimproved acreages. Significantly, though, it would artificially depress the apparent efficiency of Northern farms because Northern land values presumably capitalized their greater proximity to governmentally subsidized railroads (Wright 2006; David and Temin 1974). In their response to these criticisms, Fogel and Engerman attempted to correct for the locational component of land rents with a revised set of calculations (1977). David and Temin pointed out that the corrections "entirely eliminated land inputs from their production function!" (David and Temin 1979: 216, punctuation as in the original). Fogel and Engerman responded by charging that it was David and Temin who had called for the correction that had eliminated land (Fogel and Engerman 1980). Actually, David and Temin had not suggested such a correction. Instead, they reasserted their original argument that "the conceptual apparatus [used by Fogel and Engerman] is inappropriate to the task of ascertaining the relative technical efficiency of free and slave agriculture" (1979: 216). By this point, the debate had become mired in the cross purpose of how to properly measure inputs for a conceptual apparatus that the two sides could not agree was appropriate. The message, which should be retained, as Thomas Haskell emphasized, is that Fogel and Engerman in the course of the exchange had abandoned the major theme of *Time on the Cross*, their claim that the efficiency calculations established the diligence and

[31]Children under 10 and slaves 70 and over were assumed to be unemployed. The hand-equivalent measures are further reduced by subtracting a fractional measure of the number of male slaves engaged in nonagricultural pursuits. This adjustment is probably based on a sample of probate records, but that is not explicitly stated]. Inexplicably, farms without male slaves or with "unusual sex ratios" were dropped from the sample. A farm was also dropped, also without explanation, if it was not self-sufficient in grain production. Both of these deletions removed small farms (1–15 slaves) almost exclusively and probably distorted the measured productivity of that size class.

willing cooperation of enslaved black labor in the production of agricultural crops (Haskell 1979).[32]

Economies of Scale and Gang Labor

Although they had dropped their claim about the relative quality and diligence of slave labor, Fogel and Engerman continued to insist the productivity advantage of slavery was real and that their measure of revenue efficiency was a meaningful gauge of that advantage. They next asserted the relative efficiency of slave farms was the consequence of economies of scale in slave agriculture, an advantage not available on free family operated Northern farms and not possible with fewer than 16 slaves. They used the Parker-Gallman sample to calculate revenue efficiency for farms of different sizes, using the number of slaves owned as the measure of scale (1974: volume 1, Fig. 43).

Gavin Wright pointed out again that Fogel and Engerman's revenue efficiency "is largely a measure of who happened to be growing cotton during the most extraordinary cotton year of the nineteenth century." And significantly, Wright's calculations revealed that when "the crop mix is held constant, there is no productivity advantage for slaves and there are no scale economies for slave plantations" (Wright 1976: 317). Fogel nevertheless continued to emphasize the productivity advantage of scale, and he further claimed that a break can be observed when farm size exceeded 15 slaves (or 5 prime male hands), indicating to him that the source of the gains was the use of the gang system for field work (Fogel in Fogel et al. 1989–1992, primary volume: 26–29; Fogel 2003: 29–32).

Fogel and Engerman first introduced the idea that the gang system was common, even universal, on large plantations in *Time on the Cross.* They described the system as one in which the slaves were divided into several gangs based on their physical capabilities. The gangs were then driven through the field, one following the other with the lead gang setting the pace. Fogel and Engerman described the gang system in operation during cultivation:

> Field hands were divided into two groups: the hoe gang and the plow gang. The hoe hands chopped out the weeds which surrounded the cotton plants as well as excessive sprouts of cotton plants. The plow gangs followed behind, stirring the soil near the rows of cotton plants and tossing it back around the plants. Thus the hoe and plow gangs each put the other under an assembly-line type of pressure. The hoeing had to be completed in time to permit the plow hands to carry out their tasks. At the same time the progress of the hoeing, which entailed lighter labor than plowing, set a pace for the plow gang. The drivers or overseers moved back and forth between the two gangs, exhorting and prodding each to keep up with the pace of the other, as well as inspecting the quality of the work. (Fogel and Engerman 1974: volume 1, p. 204)

[32]In a baffling retort to Haskel published 13 years later, Fogel and Engerman merely contradicted Haskel without addressing the fundamental point made by David and Temin that the quality of labor could not be inferred from total factor productivity calculations (Fogel and Engerman 1992a).

This vivid portrayal of field work drew freely upon their imagination; it is not based on contemporary evidence. Yet, Fogel and Engerman went on and proposed that assigning slaves to "highly disciplined, interdependent teams capable of maintaining a steady and intense rhythm of work," was the "crux of the superior efficiency of large-scale operations" (Fogel and Engerman 1974: volume 1, p. 204).

The claim that the gang system was commonly employed on farms with 16 or more slaves is nothing more than conjecture. Fogel and Engerman provided no evidence that the gang system was widely used in the late antebellum period. Immediately following the passage just reproduced, they quote and paraphrase Frederick Law Olmsted's several contemporary descriptions of "slave driving," leaving the impression that they referenced the gang system just described, but none of Olmsted's observations referred to driving gangs. In these passages, Olmsted was reporting on the necessity of using beatings or the threat of whipping to compel slaves to work hard. None of his informants worked slaves in gangs (Olmsted 1856: 84, 205–206, 372–373).[33] Subsequent research on the organization of plantations, and the supervision of slaves, has produced only scattered references to the gang system (Metzer 1975). Gavin Wright warns that plantations exceeding some threshold scale cannot be associated with any particular form of organization. He also builds a case that the gang system was passing out of use by 1840 (Wright 2006: 95–96). Alan Olmstead and Paul Rhode report that they have "seen almost no slave era testimony extolling the productivity advantages of the gang system (under any name) in any cotton production activity" (Olmstead and Rhode 2008b: 1152). They also report there is little support for images of "assembly-line" pressure or the metaphor of "factories in the field" (Olmstead and Rhode 2015).

While the debates on relative efficiency, economies of scale, and the gang system left the original interpretations presented in *Time on the Cross* in tatters, it has so far failed to achieve a broad consensus about the technical efficiency of large-scale slave plantations relative to small slave farms, free Southern farms, or Northern farms. There have been a number of contributions attempting to settle the matter employing a variety of statistical techniques (Cobb-Douglas, translog, stochastic frontier, and translog ray frontier production models) with a broad and bewildering scattering of conclusions (Schaefer and Schmitz 1979; Fogel and Engerman 1980; Field 1988; Grabowski and Pasurka 1989, 1991; Hofler and Folland 1991; Field-Hendry 1995; Toman 2005). Given the limitations of the census data, I am tempted to conclude

[33]Fogel and Engerman quote Olmsted's description of a hoe gang working on a Mississippi plantation as evidence of "slave teamwork, coordination, and intensity of effort" (Fogel and Engerman 1974: volume 1, p. 205). However, they neglected to mention that Olmsted intended his description to illustrate the necessity of the threat of whipping ("the whip was evidently in constant use") not the typical organization of tasks on a large plantation (Olmsted 1860: 81–82; Gutman and Sutch 1976a).

that efforts to reach a satisfactory conclusion about the relative physical efficiency of small versus large slave farms have reached a point of greatly diminished returns. Basically, the question posed was poorly framed from the outset. It too narrowly focused on technical efficiency, a concept that proved difficult if not impossible to measure.

Taking a broader view, slavery is surely inefficient. American slaves were overworked and kept illiterate, denying them the opportunity to reach their full human potential. An investment in literacy and education and the acquisition of human capital have well-documented high rates of return (Goldin 2016).

Bondage meant slaves were not free to allocate their talents and labor time to their best advantage, barring them from contributing to the arts and sciences, barring them from leadership roles in a democratic society. Any economic system that so outrageously misuses its human resources cannot be called efficient.

The critics of *Time on the Cross* prevailed on the broader points since they were able to show that the cliometric results presented in Volume Two do not support the interpretations presented in Volume One, even if the cliometric results are taken at face value. In his "retrospective meditation" on the slavery debates, Fogel ultimately conceded that the efficiency calculations he and Engerman performed did not establish the superior management of planters and the superior quality of black labor. He rejected both the belief he once held that "technological efficiency is inherently good" and the notions entertained by some readers of *Time on the Cross* "that productivity is necessarily virtuous." Economic forces do not "automatically select moral solutions" (Fogel 2003: 46–47, 69). What cliometricians can safely conclude from the debate over efficiency is that the world got cheap cotton at the expense of the education, intellectual development, inspiring ambitions, and personal security of its enslaved producers.

The Stability of the Black Family

Fogel and Engerman's insistence on the efficiency of slavery had misled them into trumpeting the superior quality of slave labor. Their insistence on the personal efficiency of black workers then pushed them to suggest enslaved people were willing collaborators in the production of staple crops. The slaves' presumed contentment with their status would be plausible, they argued, if one recognized that the black family was the basic unit of social organization under slavery. The security and stability of family life would have been essential to gain the slave's compliant acceptance. This chain of inferences allowed Fogel and Engerman to announce one of their principal corrections was to overturn the belief that sexual abuse of slave women had destroyed the black family (Fogel and Engerman 1974: volume 1).

Fogel and Engerman denied that sexual abuse was common. They did this, first, by declaring without evidence that slave breeding was "myth"[34] and then by denying that white masters and overseers "ravished black women frequently," justifying that claim with the assertion that "white men who desired illicit sex had a strong preference for white women." The only fact offered to support this generalization (and indeed the only evidence that supported the entire house of cards) was "the failure of Nashville's brothels to employ slave women" (Fogel and Engerman 1974: volume 1, pp. 133–135). The phrasing is unfortunate. As Martha Hoffman noted in her discussion of the moral issues raised by *Time on the Cross*, the word "failure" implied to some readers that slave women "should have been" prostitutes (1992: 600). But surely, leaving that point aside, the conclusion that Southern white men's preferences in these matters would protect black women is naïve in the extreme.[35]

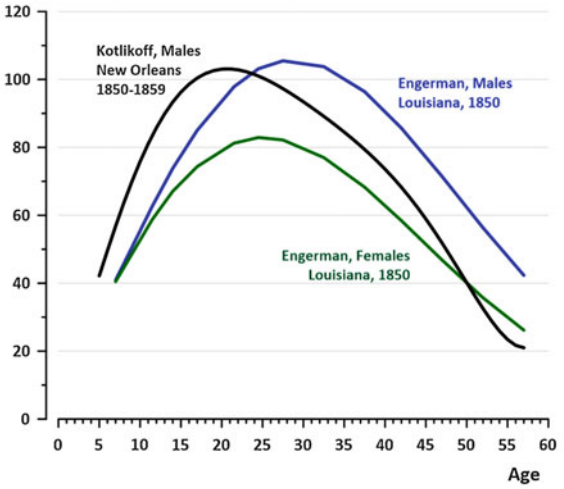
One of the first of Fogel and Engerman's claims in this chain of inferences to fall was the alleged "fact" that enslaved black women were not forced into prostitution. Yet, their source, David Kaser, had clearly noted that the 1860 census did not record the occupations of slaves (Kaser 1964). Even if Nashville had hundreds of them, they would not have appeared in the records. In the opening pages of *Time on the Cross*, Fogel and Engerman defended the robustness of the facts and findings they would report, saying "when persistent efforts to contradict the unexpected discoveries failed," they were forced to accept a "radical reinterpretation of American slavery" (Fogel and Engerman 1974: volume 1, p. 8). In the case of the Nashville prostitutes, the vetting could hardly have been persistent or even attentive.

After the Controversy

Replication is a normal part of the scientific method. If others cannot confirm reported findings, the new results are declared unproven and the research that produced them is deemed a failure. However, a focus exclusively on Fogel and Engerman's careless handling of their evidence and their arrogant claims of possessing a precise and powerful methodology unavailable to ordinary historians

[34]Fogel and Engerman ignored the evidence that I had presented, which I discuss elsewhere in this chapter (Sutch 1975a). They argued for a rejection of slave breeding with two conjectures: (1) that interference could have had little effect on the rate of population growth and (2) that high costs to slave breeding would have eroded away all profit in the business. Gutman and I countered these suppositions by pointing out that the first rested on the fact that the fertility of the slave population was close to the biological maximum. But those high levels might have been produced by breeding. They should have asked how much the fertility rate would have fallen had breeding been stopped. No evidence was offered to support the second argument other than the supposed loss of the slaves' willing cooperation in the production process (Gutman and Sutch 1976b).

[35]Fogel and Engerman claimed that the census data on the proportion of the slave children under 10 recorded as mulatto (estimated by Steckel at 10.1% in 1860) cannot be used as evidence of frequent impregnations of enslaved women by whites (Fogel and Engerman 1974: volume 1, pp. 131–133). However, the evidence they provided in volume 2 on this point is totally without scientific creditability (Gutman and Sutch 1976b). See Steckel (1980) and Malone (1992) for extended discussion of miscegenation.



Sources: *Males and Females, Louisiana, 1850*: Engerman, Sutch, and Wright 2006: series Bb217-Bb218.
Males, New Orleans, 1850-1860: Kotlikoff 1979: table 4.

Fig. 4 Age profile of slave values by sex, Louisiana 1850s. Index: 100 = Value of average male slave, 18 to 30

would miss an important lesson. *Time on the Cross* was the product of a system that rewards with lavish attention powerful findings that seemingly offer either an outright rejection or strong support for conclusions the public finds comfortable. Eager to make a splash, Fogel and Engerman let their enthusiasm overwhelm their precautions. Practicing cliometrics, however, does not absolve the practitioner from adhering to the strictures and norms of science. It must be emphasized on the other hand that Stanley Engerman and Robert Fogel generously answered questions and provided raw data to their critics and welcomed the exchange of results and interpretations (Fogel et al., *Technical Papers*, volume 1, 1989, p. xvi). That openness to criticism and their willingness and eagerness to engage in the debate have now become the hallmarks of first-class cliometrics.

While considering *Time on the Cross* an interpretive failure, I do not want to overlook some of the positive contributions made by Fogel and Engerman. Among these is the valuable archival research that assembled data on slave values by age, gender, and location developed from probate records. Figure 4 presents the age-value profiles for Louisiana and compares those to the male profile estimated by Laurence Kotlikoff from his analysis of the selling price of slaves in the New Orleans slave market.[36]

[36]Fogel and Engerman 1974: volume 1, pp. 72–78, figures 15, 16, and 18; volume 2: 24, 79–82. The data plotted were provided by Stanley Engerman and are reproduced in Engerman et al. 2006: volume 2, p. 373, tables Bb209–214 and Bb215–218). The age profile of the New Orleans sale prices for 1850–1859 is based on the coefficients of a sixth-degree polynomial reported by Laurence Kotlikoff 1979: Table 4.

An even more significant contribution was Fogel and Engerman's multidisciplinary approach to their subject. This was fairly novel at the time, and their example inspired others to pursue opportunities to push ahead on those fronts. This openness to the perspectives of other disciplines had a profound influence on the cliometric investigation of many topics beyond those raised by slavery in the years that followed. Two examples of uniting disciplines to address the economics of slavery illustrate the importance of this reorientation. Richard Steckel looked at the fertility of slave women using the tools of demography (Steckel 1977). His exploration into this issue led to the collection of data on the heights of teenage slave women to judge the age of menarche (Trussell and Steckel 1978; Steckel 1979, 1986a). That statistic is relevant to judging how early in a woman's life, biologically speaking, was her age at her first birth – how soon after puberty. Ultimately, this line of research produced the anthropometric revolution in cliometric research. Steckel and many others collected comparative data on human stature and weight to evaluate the well-being of disparate populations throughout history (for a review, consult Komlos and Alecke 1996; Craig 2016).

Kenneth Kiple and Virginia Kiple's exploration of the mortality of slave children and my effort to critique the analysis of the slave diet in *Time on the Cross* led us to explore the basics of nutritional science (Kiple and Kiple 1977; Sutch 1975b). There is now an extended literature on the biologic and biomedical history of the black population both during and after slavery (Kiple and King 1981; Steckel 1986a, b; Troesken 2004). At the same time, multidisciplinary studies of the nutritional and medical histories of many other populations have applied the methods to a wide range of diverse settings (Steckel and Floud 1997).

Time on the Cross advanced the general discussion of slavery by extending it to topics well beyond the viability of the institution and its impact on economic growth. Fogel and Engerman brought cliometric evidence and methods to assess the provision of food, shelter, and clothing; the slave's family life; and punishments, rewards, and expropriation. Their provocative opinions on these topics, even as they were overturned, accelerated the general pace of research on the economics, sociology, and demography of slavery. As Fogel remarked, the public debate "was a debate in which there were no losers" (2003: 32). The new research and the collection of new data by Fogel and Engerman's critics and defenders greatly enhanced the credibility and acceptance of cliometrics by economists. Yet, there was an unfortunate consequence as well.

This outpouring of new work had little apparent influence on the discussion of slavery among historians. Alan Olmstead and Paul Rhode noted this disconnect:

> In the past, historians and economists (sometimes working as a team) collectively advanced the understanding of slavery, southern development, and capitalism. There was a stimulating dialog. That intellectual exchange deteriorated in part because some economists produced increasingly technical work that was sometimes beyond the comprehension of many historians. Some historians were offended by some economists who overly flaunted their findings and methodologies. (Olmstead and Rhode 2018: 14)

Apparently, the bitterness of the debate over *Time on the Cross*, the far too many factual errors in the book, and the ostentatious, yet imprecise, formalism of Volume Two led some historians to simply ignore the entirety of cliometric literature. Recently, two historians of American slavery, Sven Beckert and Seth Rockman, dismiss their own neglect with a single sentence. "The economic history of slavery has labored in the shadows of the interpretative controversies surrounding . . . Time on the Cross" (2016: 10). Presumably, this excused them from critiquing the cliometric literature and freed them to contribute to and celebrate an alternative economic history of slavery and American economic development. Their loss (and ours) has become glaringly apparent in the recent discussion by cliometricians of what historians have come to call the "New History of Capitalism and Slavery" (Murray et al. 2015; Olmstead and Rhode 2018).

In 1989, Robert Fogel responded to critics of *Time on the Cross* and the "crackling atmosphere" of the debates. This effort took four volumes collectively titled *Without Consent or Contract: The Rise and Fall of American Slavery* (Fogel et al. 1989–1992). The primary volume, by Fogel alone, begins with a review of several selected issues addressed in the debates. His tone is less polemical and the message less audacious than in the 1974 volumes. The revised portrayal of the slave system is more nuanced, more complex, somewhat subtle, and less contentious (for a review, see Clark Nardinelli 1994). Endnotes were attached. The primary volume also presented evidence relating to slave societies outside of the United States and expanded its attention to address ideological, religious, moral, and political issues. Three companion volumes, which appeared in 1992, provide a mixture of brief research memos together with technical papers, many of which had already been published years before in the journal literature.[37] Judged by citations, none of the new material on antebellum slavery in the companion volumes attracted rebuttal from the critics or attention from others.[38]

Fogel announced in 1989 that he had written his last words on slavery with the publication of *Without Consent or Contract* (Fogel et al. 1989: 13). That all but ended the debate over *Time on the Cross*. Most of the participants, like Fogel, moved on to other topics.[39] This finality, unfortunately, left the resolution of the

[37]For some insight into the organization of the "loosely structured" research project that produced both *Time on the Cross* and *Without Consent or Contract*, see Fogel and Engerman's "General Introduction" to the *Technical Papers* (Fogel et al. 1992: volume 1).

[38]In the primary volume of *Without Consent or Contract*, Fogel continued to report total factor productivity calculations like those in *Time on the Cross*, but he reframed the conclusion to claim "the superior efficiency of the big plantations was due not merely to inherent advantages of the gang system but also to the concentration of above-average ability in the ownership of such farms" (Fogel in Fogel et al. 1989–1992: primary volume, figure 14). Gavin Wright responded by noting that "the ability of the ownership is no more directly observable in census data than the gang labor system itself" (Wright 2006: 96).

[39]There was some continuing work by economic historians to provide quantitative data that might inform the short-lived reparation movement to redress the injustice of slavery (America 1990).

debate – and particularly Fogel and Engerman's joint view – somewhat muddled.[40] That may have had the effect of discouraging further cliometric research. Apart from the continuing work with anthropometric and biometric evidence already mentioned, the next quarter century produced only a thin list of cliometric contributions on American slavery. There were, however, two exceptions to the general lack of new research that I shall come to shortly. Apart from these contributions, the scholarly interest in slavery by quantitative historians switched away from the late antebellum South to topics like the Atlantic slave trade, colonial slavery, and slavery in the Caribbean, Brazil, and Africa. It is not clear why this should be so. Perhaps, the practicing cliometricians had exhausted themselves, if not the topic. Perhaps a younger generation of cliometric scholars had learned that the sensitive issues involved had best be avoided given the topic's history of rather acrimonious debate.

Group Sales and Price Discounts in the Market for Slaves

A new issue dealt with after the controversy had died down was the puzzling price discounts for intact slave families when sold on the New Orleans market. Under Louisiana law, enslaved blacks were analogous to real estate, and the recording of deeds ensured the title to this form of property. One of the contributions of the Fogel and Engerman project had been the digitization of a large sample of the bill-of-sales recorded for slaves in New Orleans (Fogel and Engerman 1974: volume 2, Table B.1 (data set 9)). Over 5700 invoices covering the period 1804 to 1862 were recorded (less than 5% of the total number in the archives). Analysis of the structure of the sale prices by Laurence Kotlikoff revealed that a well-functioning ("rational") market existed. Kotlikoff noted that the hump shape of the age-price profile (see Fig. 4), the premium prices paid for males, and for warranties of good conduct "all point to careful, calculating transactors operating in a highly developed market in human beings." In a point that should not be passed over lightly, he also took note of the degrading physical inspections of the men and women by the "careful" and "calculating" potential buyers. Slaves were stripped of all clothing and closely examined to assess muscle development and to discover physical defects such as whipping scars (Kotlikoff 1979).

It was a surprise, however, that the sales records also revealed that very few family groups were sold. Of the 5,785 sales in the sample, only 40 involved a husband and wife sold together (in 22 of these cases, the couple was accompanied by their child). Another 94 were mothers sold with her accompanying child or children. Seventy-seven percent of the sales were slaves sold individually (Kotlikoff 1979: 513). There was little apparent regard for preserving enslaved families intact. More

[40]The primary volume of *Without Consent or Contract* was written by Fogel alone. Engerman writing jointly with Kenneth Sokoloff seems to have retreated from *Time on the Cross* and accepted the view that in the years before the Civil War, the South "lagged behind the North . . . in evolving a set of political institutions that were conducive to broad participation in the commercial economy" (Engerman and Sokoloff 2002: 61).

surprising was that, except for the 22 husband-wife-child combinations, slaves in family groups sold for deep discounts compared to their value if they had been sold separately. This suggests either that buyers in New Orleans did not place much value on the behavioral benefits in terms of demeanor and submission that Fogel and Engerman thought would protect family units, or that the families sold at a discount (predominately mother-child pairs) were in less than prime condition when they arrived in the Crescent City.

Jonathan Pritchett, together with several colleagues, undertook a re-examination of the New Orleans sales data to explore the puzzle. He joined with Herman Freudenberger to point out that the slaves sold in New Orleans had been selected by traders for transport to New Orleans because of their high value. They were better able to bear the cost of transport by ship. Not only were the slaves clustered in the prime ages (10–30), but they were in prime condition, noticeably taller than slaves transported by their owners. Presumably, they were stronger and in better health than average. This selection bias was particularly noticeable for children and adolescents (Pritchett and Freudenberger 1992). Pritchett and Richard Chamberlain confirmed this result by comparing the price of slaves in estate sales, which would include both high-valued and other slaves, with the prices in the New Orleans sales records. They found no evidence that adverse selection had put slaves with hidden defects on the New Orleans market (Pritchett and Chamberlain 1993). Charles Calomiris and Pritchett analyzed the deep discounts for mother-child pairs. Noting the relative rarity of such sales (74% of children, 4 to 13, shipped by traders were unaccompanied), they explored the possibility that the typical mother-child unit was not in prime condition. Pritchett and Freudenberger had earlier found that children shipped to New Orleans were taller than other children of the same age and gender, but they had not distinguished between those children who were shipped with their mother and those who were unaccompanied. Returning to the ship manifests and identifying likely mother-child pairs, Calomiris and Pritchett discovered that accompanied children were shorter than the unaccompanied by one and one-half inches (for children under 10), confirming the hypothesis that the discounts reflected perceived quality differences (Calomiris and Pritchett 2009: Table 4). The few observations of a mother sold together with a child in New Orleans were likely cases where the child was evidently poorly nourished or in otherwise poor health. The New Orleans data do not support the hypothesis that slaveholders valued and respected the human bonds of affection and mutual responsibility that define a family. When money was involved, many masters seemed willing to break families apart.[41]

[41]The legal historian, Thomas Russell, has assembled convincing evidence that Southerners believed that slaves sold separately would bring higher prices. Southern courts ruled that estate executors had a fiduciary duty to break up slave families in order to maximize the sales revenue (Russell 1996). His study of South Carolina suggests that court-supervised slave sales comprised one-half of all slave sales in that state (Russell 1993).

Biological Innovation and Southern Agricultural Development

The debates over the rates of antebellum economic growth in the South and North had established that a significant factor in the case of the South was the shift of population from the South Atlantic states to the East South Central states (as demonstrated by the comparison made in Table 3).[42] The advantage of the Western states had been attributed to the naturally high fertility of the soil: the black-belt soils of Alabama and the alluvial soils of the Mississippi basin. Alan Olmstead and Paul Rhode, in an important post-controversy contribution, insist that this summary ignores the role biological innovation played in Southern agricultural development. With the move onto virgin land in the West, planters needed to adapt the cotton plant to the local conditions. To this end, they experimented with different varieties and selected the most advantageous. Yields were increased, fiber quality was improved, and taller plants, which flowered and formed larger bolls further from the ground, were developed. Since the labor requirement for hand picking was the primary constraint on productivity per slave, the taller, more prolific varieties markedly increased the number of pounds of cotton that a worker could pick in a day. When superior varieties were perfected, they spread East and West and North and South, thus increasing productivity throughout the South (Olmstead and Rhode 2008a, b, 2011).

Olmstead and Rhode quantified the advances in productivity by collecting daily records of cotton picking. Plantation overseers recorded the weights of individual worker's pickings to monitor each slave's performance, reward high effort, and discipline slackers.[43] Their data collection totaled 704,800 individual daily reports covering 6,200 slaves working on 142 plantations. These data when plotted and averaged reveal a threefold improvement in productivity in the Western states between 1811 and 1860 and a doubling of productivity for the Atlantic Coast states. The time trends are plotted in Fig. 5.[44] A statistical analysis of the data by region revealed that the dynamics of Southern development were driven by the interaction of biological innovations with the Western movement to alluvial soils:

> Most of the new cotton varieties were invented in the West, and they were undoubtedly best suited for the area where they first evolved thereby enhancing the West's comparative advantage. The new varieties gradually were adapted to displace older varieties in a wide spectrum of geoclimatic conditions, but the new technologies were particularly suited for more fertile lands. The large plantation regions of the West had better soils to start with, but in a sense biological innovation made these soils even better. (Olmstead and Rhode 2008b: 1156–1157)

[42]Another factor was the extraordinary growth of the demand for cotton emphasized by Gavin Wright (1975)

[43]"Discipline" may be a euphemism here; slaves were not infrequently whipped for below par performance (Olmstead and Rhode 2008b: 1143).

[44]The four Atlantic coast states are Georgia, North and South Carolina, and Virginia. The seven Western states are Alabama, Arkansas, Florida, Louisiana, Mississippi, Tennessee, and Texas.

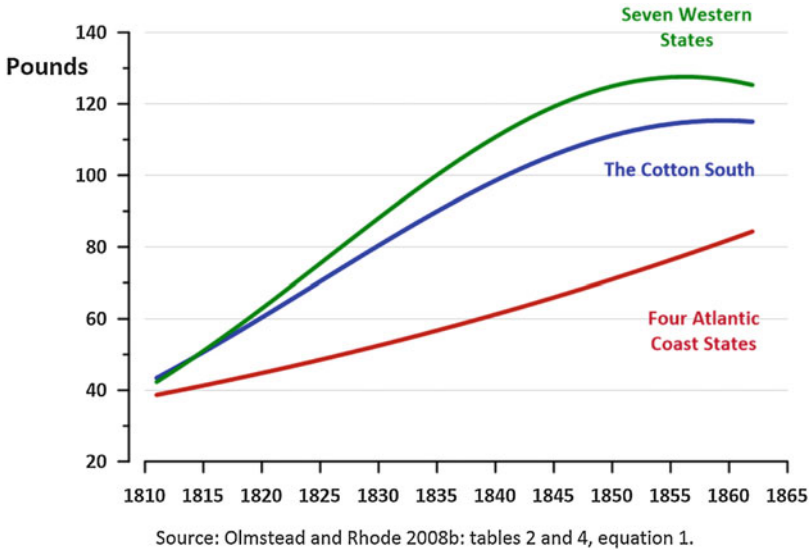


Fig. 5 Daily cotton picking rates, 1811–1862. Mean pounds harvested per worker per day

Olmstead and Rhode consulted an extensive set of primary sources to support and contextualize their statistical findings. Newspaper accounts, agricultural journals, letters from planters and overseers, and treatises on plant biology and scientific farming established that the biological advances were the work of slave-owning planters who molded the cotton plant to their geoclimatic conditions, soil types, and the evolving requirements of textile manufacturers. Earlier characterizations of slave owners as "uninventive" rested on the planters' failure to experiment with mechanical inventions. A broader view of invention challenges this view.

An Assessment

American slavery was a brutally cruel and highly exploitive system. It was also a racist institution. No whites were slaves and the presumptive status of every black person was slave.

Those facts are disturbing. They make it impossible to regard American history as one of unblemished freedom, equality, and democratic community. After a 60-year-long investigation of the economics of slavery by economists and cliometricians, it is also impossible to deny the complicity of America's economic system based on the inviolability of private property, strict contract enforcement, and an unregulated labor market in propelling and sustaining the slave regime.

The Southern planter was a capitalist – a capital owner, who made decisions to buy or sell slaves and to employ them at one task or another guided by a profit motive and the constant pressure of competition. The major contribution of

cliometrics to the economic history of slavery was to view the planter as a capitalist. Only the slave owner willing to employ his slaves in the most efficient manner would earn a return sufficient to justify their price. Achieving high output required that workers be engaged in intensively hard work for long hours under the constant threat of corporal punishment (as acknowledged by Fogel in Fogel et al. 1989–1992, primary volume: p. 34). Each slave owner had to replicate those features on his own plot of land to remain in business. If maintaining a profitable enterprise necessitated the disrespect of marital status and family bonds, few planters could afford more empathetic treatment. Those unwilling to continence the inhumanity of the system left the business to those who could. To those who became and remained masters, a blind-eyed belief in racial inferiority served to excuse and justify racial degradation and brutal treatment. The absence of contrary voices in the South allowed the racism to fester and intensify. Racism poisoned antebellum society and, as we know all too well, it still poisons American culture today. The true burden of slavery defies quantification.

Acknowledgments I thank Howard Bodenhorn, Susan Carter, Alexander Field, Michael Haupert, David Mitch, Jonathan Pritchett, Roger Ransom, Paul Rhode, Alan Olmstead, and Gavin Wright for encouragement and helpful suggestions.

References

Aitken HGJ (ed) (1971) Did slavery pay? Readings in the economics of black slavery in the United States. Houghton Mifflin, Boston

America RF (ed) (1990) The wealth of races: the present value of benefits from past injustices. Greenwood Press, New York

Anderson RV, Gallman RE (1977) Slaves as fixed capital: slave labor and southern economic development. J Am Hist 64(1):24–46

Atack J, Bateman F (1987) To their own soil: agriculture in the Antebellum North. Iowa State University Press, Ames

Bancroft F (1931) Slave-trading in the Old South. J. H. Furst

Bateman F, Weiss T (1976) Manufacturing in the Antebellum South. Res Econ Hist 1:1–44

Bateman F, Weiss T (1981) A deplorable scarcity: the failure of industrialization in the slave economy. University of North Carolina Press, Chapel Hill

Bateman F, Foust J, Weiss T (1974) The participation of planters in manufacturing in the Antebellum South. Agric Hist 48(2):277–297

Battalio RC, Kagel J (1970) The structure of Antebellum Southern agriculture: South Carolina, a case study. Agric Hist 44(1):25–37

Beckert S, Rockman S (2016) Slavery's capitalism: a new history of American Economic Development. University of Pennsylvania Press, Philadelphia

Bodenhorn H (2015) The color factor: the economics of African-American well-being in the nineteenth-century South. Oxford University Press, Oxford

Calomiris CW, Pritchett JB (2009) Preserving slave families for profit: traders' incentives and pricing in the New Orleans slave market. J Econ Hist 69(4):986–1011

Calomiris CW, Pritchett JB (2016) Betting on secession: quantifying political events surrounding slavery and the civil war. Am Econ Rev 106(1):1–23

Carter SB, Gartner SS, Haines MR, Olmstead AL, Sutch R, Wright G (eds) (2006) Historical statistics of the United States: earliest times to the present, Millennial Edition, Five volumes. Cambridge University Press, New York

Clayton R (2002) Cash for blood: the Baltimore to New Orleans domestic slave trade. Heritage Books, Bowie
Conrad AH, Meyer JR (1958) The economics of slavery in the Antebellum South. J Polit Econ 66(2):95–130. Reprinted in their book, The Economics of Slavery and Other Studies in Econometric History, Aldine, 1964: 43–92
Conrad AH, Dowd D, Engerman S, Ginzberg E, Kels C, Meyer JR, Scheiber HN, Sutch R (1967) Slavery as an obstacle to economic growth in the United States: a panel discussion. J Econ Hist 27(4):518–560
Craig LA (2016) Nutrition, the biological standard of living, and cliometrics. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer, pp 113–130
Crawford S (1992) The slave family: a view from the slave narratives. In: Goldin C, Rockoff H (eds) Strategic factors in nineteenth century American economic history: a volume to honor Robert W. Fogel. University of Chicago Press, Chicago, pp 331–350
David PA, Temin P (1974) Slavery: the progressive institution? J Econ Hist 34(3):729–783. Reprinted with changes in David et al [1976: 165–230]
David PA, Temin P (1979) Explaining the relative efficiency of slave agriculture in the Antebellum South: comment. Am Econ Rev 69(1):213–218
David PA, Gutman HG, Sutch R, Temin P, Wright G (1976) Reckoning with slavery: a critical study in the quantitative history of American Negro slavery. Oxford University Press, New York, pp 134–164
Diebolt C, Haupert M (2016) An introduction to the handbook of cliometrics. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer, Berlin, pp v–xiv
Dowd DF (1958) The economics of slavery in the Ante Bellum South: a comment. J Polit Econ 66(5):440–442
Easterlin RA (1961) Regional income trends, 1840–1950. In: Harris SE (ed) American Economic History. McGraw-Hill, New York, pp 525–547
Easterlin RA (1974) Farm production and incomes in old and new areas at mid-century. In: Klingaman DC, Vedder RK (eds) Essays in nineteenth century economic history: the Old North West. Ohio University Press, Athens, pp 77–117
Engerman SL (1967) The effects of slavery upon the Southern economy: a review of the recent debate. Explor Entrep Hist Second Series 4(2):71–97. Winter 1967
Engerman SL (1971) Some economic factors in Southern backwardness in the nineteenth century. In: Kain JF, Meyer JR (eds) Essays in regional economics. Harvard University Press, pp 279–306
Engerman SL (1975) Comments on the study of race and slavery. In: Engerman SL, Genovese ED (eds) Race and slavery in the Western Hemisphere: quantitative studies. Princeton University Press, Princeton, pp 495–530
Engerman SL, Olson JF (1992) Basic procedures for the computation of outputs and inputs from the Parker-Gallman sample, including a procedure for the elimination of defective observations. In: Fogel RW, Galantine RA, Manning RL (eds) Without consent of contract: evidence and methods. W. W. Norton, pp 205–209
Engerman SL, Sokoloff KL (2002) Factor endowments, inequality, and paths of development among new world economies. Economía 3(1):41–109. Fall 2002
Engerman SL, Sutch R, Wright G (2006) "Slavery," Chapter Bb in Carter et al., Historical statistics of the United States, vol 2. Cambridge University Press, pp 369–386
Evans R Jr (1962) The economics of American Negro slavery, 1830–1860, Aspects of labor economics. Princeton University Press, Princeton, pp 221–227
Field E (1988) The relative efficiency of slavery revisited: a Translog production function approach. Am Econ Rev 78(3):543–549
Field-Hendry E (1995) Application of a stochastic production frontier to slave agriculture: an extension. Appl Econ 27(4):363–368
Fishlow A (1964) Antebellum interregional trade reconsidered. Am Econ Rev 54(3):352–364
Fogel RW (1964) Discussion [A provisional view of the 'New Economic History']. Am Econ Rev 54(3):377–389

Fogel RW (1990) Interview conducted by Samuel H. Williamson and John S. Lyons. Newsl Cliometric Soc 5(3):20–29. July 1990: 3–8, Reprinted with changes in John S. Lyons, Louis P. Cain, and Samuel H. Williamson, editors. Reflections on the Cliometric Revolution: Conversations with Economic Historians, Routledge 2008: 332–353

Fogel RW (2003) The slavery debates, 1952–1990: a retrospective. Louisiana State University Press, Baton Rouge

Fogel RW, Engerman SL (1971a) The relative efficiency of slavery: a comparison of northern and southern agriculture in 1860. Explor Econ Hist 8(3):353–367. Spring 1971

Fogel RW, Engerman SL (1971b) The economics of slavery. In: Fogel RW, Engerman SL (eds) The reinterpretation of American economic history. Harper & Row, pp 311–341

Fogel RW, Engerman SL (1974) Time on the cross, volume 1, The economics of American Negro slavery; volume 2, evidence and methods – A Supplement. Little, Brown

Fogel RW, Engerman SL (1977) Explaining the relative efficiency of slave agriculture in the Antebellum South. Am Econ Rev 67(3):275–296

Fogel RW, Engerman SL (1980) Explaining the relative efficiency of slave agriculture in the Antebellum South: reply. Am Econ Rev 70(4):672–690

Fogel RW, Engerman SL (1992a) Reply to Haskel. In: Fogel RW, Engerman SL (eds) Without consent or contract: technical papers, vol 1. Markets and Production, W. W. Norton, p 293

Fogel RW, Engerman SL (1992b) The slave breeding thesis. In: Fogel RW, Engerman SL (eds) Without consent or contract: technical papers, volume 2, Conditions of slave life and the transition to freedom. W. W. Norton, pp 455–472

Fogel RW, et al. (1989–1992) Without consent or contract: the rise and fall of American slavery, four volumes: Robert W. Fogel, [primary volume] 1989; Robert W. Fogel, Ralph A. Galantine, and Richard L. Manning, editors, Evidence and Methods,1992; and Robert William Fogel and Stanley L. Engerman, editors, Markets and Production, Technical Papers, volume 1, 1992, and Conditions of Slave Life and the Transition to Freedom, Technical Papers, volume 2; W.W. Norton 1992

Foust JD, Swan DE (1970) Productivity and profitability of antebellum slave labor: a micro-approach. Agric Hist 44(1):39–62

Gallman RE (1966) Gross National Product in the United States, 1834–1909. In: Brady D (ed) Output, employment, and productivity in the United States after 1800. Columbia University Press, pp 3–90

Gallman RE (1970) Self-sufficiency in the cotton economy of the Antebellum South. Agric Hist 44(1):5–23

Goldin CD (1975) A model to explain the relative decline of urban slavery: empirical results. In: Engerman SL, Genovese ED (eds) Race and slavery in the Western Hemisphere: quantitative studies. Princeton University Press, pp 427–450

Goldin CD (1976) Urban slavery in the American South. 1820–1860: a quantitative history. University of Chicago Press, Chicago

Goldin CD (2016) Human capital. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer, pp 55–86

Govan TP (1942) Was plantation slavery profitable? J South Hist 8(4):515–536

Grabowski R, Pasurka C (1989) The relative efficiency of slave agriculture: an application of a stochastic frontier. Appl Econ 21(5):587–595

Grabowski R, Pasurka C (1991) The relative efficiency of slave agriculture: a reply. Appl Econ 23(5):869–870

Gray LC (1933) History of agriculture in the Southern United States to 1860, two volumes. Carnegie Institution of Washington

Gutman HG (1975) The world two cliometricians made: a review essay of F+E = T/C. J Negro Hist 60(1):53–227. Reprinted as Slavery and the Numbers Game: A Critique of Time on the Cross, University of Illinois Press

Gutman H, Sutch R (1976a) Sambo makes good, por were slaves imbued with the protestant work ethic?. In: David PA et al (eds) Reckoning with slavery. W.W. Norton, pp 55–93

Gutman H, Sutch R (1976b) Victorians all? The sexual mores and conduct of slaves and their masters. In: David PA et al (eds) Reckoning with slavery. W.W. Norton, pp 134–162
Haber S (1964) Efficiency and uplift: scientific management in the progressive era, 1890–1920. University of Chicago Press, Chicago
Haskell TL (1979) Explaining the relative efficiency of slave agriculture in the Antebellum South: a reply to Fogel and Engerman. Am Econ Rev 69(1):206–207
Hoffman MK (1992) Thoughts on the treatment of moral issues in time on the cross. In: Fogel RW, Galantine RA, Manning RL (eds) Without consent or contract: evidence and methods. W.W. Norton, pp 599–603
Hofler R, Folland S (1991) The relative efficiency of slave agriculture: a comment. Appl Econ 23(5):861–868
Johnson W (1999) Soul by soul: life inside the Antebellum slave market. Harvard University Press, Cambridge, MA
Kaser D (1964) Nashville's women of pleasure in 1860. Tenn Hist Q 23(4):379–382
Kiple KF, King VH (1981) Another dimension to the black diaspora: diet, disease, and racism. Cambridge University Press, Cambridge, UK
Kiple KF, Kiple V (1977) Slave child mortality: some nutritional answers to a perennial puzzle. J Soc Hist 10(3):284–309. Spring 1977
Komlos J, Alecke B (1996) The economics of Antebellum slave heights reconsidered. J Interdiscip Hist 26(3):437–457. Winter 1996
Kotlikoff LJ (1979) The structure of slave prices in New Orleans, 1804 to 1862. Econ Inq 17(4):496–517
Kravis IB (1972) The role of exports in nineteenth-century United States growth. Econ Dev Cult Chang 20(3):387–405
Lindert PH, Williamson JG (2016) Unequal gains: American growth and inequality since 1700. Princeton University Press, Princeton
Lindstrom D (1970) Southern dependence upon interregional grain supplies: a review of the trade flows, 1840–1860. Agric Hist 44(1):101–113
Malone AP (1992) Sweet chariot: slave family and household structure in nineteenth-century Louisiana. University of North Carolina Press, Carolina
McClelland PD (1997) Sowing modernity: America's first agricultural revolution. Cornell University Press, Ithaca
McCloskey D, [Deirdre] N (1985) The problem of audience in historical economics: rhetorical thoughts on a text by Robert Fogel. Hist Theory 24(1):1–22. Reprinted with revisions as "The Problem of Audience in Historical Economics: Robert Fogel as Rhetor," in McCloskey, The Rhetoric of Economics, University of Wisconsin Press, pp 113–137
McCloskey D, [Deirdre] N, Hersh GK Jr (1990) A bibliography of historical economics to 1980. Cambridge University Press, Cambridge, UK
Metzer J (1975) Rational management, modern business practices, and economies of scale in the Antebellum southern plantations. Explor Econ Hist 12(2):123–150
Meyer JR, Conrad AH (1957) Economic theory, statistical inference, and economic history. J Econ Hist 17(4):524–544
Modigliani F (1961) Long-run implications of alternative fiscal policies and the burden of the national debt. Econ J 71(284):730–755
Modigliani F (1966) The life cycle hypothesis of savings, the demand for wealth and the supply of capital. Soc Res 33(2):160–217
Murray JE, Olmstead AL, Logan TD, Pritchett JB, Rousseau PL (2015) Roundtable of reviews for the half has never been told: slavery and the making of American capitalism [by Edward E. Baptist]. J Econ Hist 75(3):919–931
Nardinelli C (1994) Fogel's farewell to slavery: a review essay. Hist Methods 27(3):133–139
North DC (1961) The economic growth of the United States, 1790–1860. Prentice-Hall, Englewood Cliffs

North DC (1963) Quantitative research in American economic history. Am Econ Rev 53(1): 128–130. Part 1 March 1963
Olmstead AL, Rhode PW (2008a) Creating abundance: biological innovation and American agricultural development. Cambridge University Press, New York
Olmstead AL, Rhode PW (2008b) Biological innovation and productivity growth in the antebellum cotton economy. J Econ Hist 68(4):1123–1171
Olmstead AL, Rhode PW (2011) Productivity growth and the regional dynamics of Antebellum Southern development. In: Rhode PW, Rosenbloom JL, Weiman DF (eds) Economic evolution and revolution in historical time. Stanford University Press, Stanford, pp 180–213
Olmstead AL, Rhode PW (2015) Were Antebellum cotton plantations factories in the field? In: Collins WJ, Margo RA (eds) Enterprising America: businesses, banks, and credit markets in historical perspective. University of Chicago Press, pp 245–276
Olmstead AL, Rhode PW (2018) Cotton, slavery, and the new history of capitalism. Explor Econ Hist 67(1):1–17
Olmsted FL (1856) A journey in the seaboard slave states; with remarks on their economy. Dix and Edwards, New York
Olmsted FL (1860) A journey in the back country. Mason Brothers, New York
Parker WN (1970a) Introduction: the cotton economy of the Antebellum South. In: Parker WN (ed) The structure of the cotton economy of the Antebellum South. Agricultural History Society, Washington, DC, pp 1–4
Parker WN (1970b) Slavery and southern economic development: an hypothesis and some evidence. Agric Hist 44(1):115–125
Parker WN, Klein JLV (1966) Productivity growth in grain production in the United States, 1840–60 and 1900–10. In: Brady DS (ed) Output, employment, and productivity in the United States after 1800. Columbia University Press, New York, pp 523–580
Phillips UB (1905) The economic cost of slaveholding in the cotton belt. Political Sci Q 20(2): 257–275
Phillips UB (1918) American Negro slavery: a survey of the supply, employment and control of Negro labor as determined by the plantation regime. Appleton, New York
Phillips UB (1929) Life and labor in the Old South. Little, Brown
Pritchett JB (2001) Quantitative estimates of the United States interregional slave trade, 1820–1860. J Econ Hist 61(2):467–475
Pritchett JB, Chamberlain RM (1993) Selection in the market for slaves: New Orleans, 1830–1860. Q J Econ 108(2):461–473
Pritchett JB, Freudenberger H (1992) A peculiar sample: the selection of slaves for the New Orleans market. J Econ Hist 52(1):109–127
Pritchett J, Hayes J (2016) The occupations of slaves sold in New Orleans: missing values, cheap talk, or informative advertising? Cliometrica 10(2):181–195
Ramsdell CW (1929) The natural limits of slavery expansion. Miss Val Hist Rev 16(2):151–171
Ransom RL, Sutch R (1977) One kind of freedom: the economic consequences of emancipation. Cambridge University Press, New York. First Edition 1977, Second Edition 2001
Ransom RL, Sutch R (1988) Capitalists without capital: the burden of slavery and the impact of emancipation. Agric Hist 62(3):133–160. Summer 1988
Ransom RL, Sutch R (2001) Conflicting visions: the American civil war as a revolutionary event. Res Econ Hist 20:249–301
Rhode PW, Sutch R (2006) Estimates of national product before 1929. In: Carter S et al (eds) Historical statistics of the United States: earliest times to the present. Millennial edition, vol 3. Cambridge University Press, pp 12–20
Roosevelt T (1909) Special message to the Senate and House of Representatives: January 22, 1909. Online at Gerhard Peters and John T. Woolley. The American Presidency Project. http://www.presidency.ucsb.edu/ws/?pid=69658
Russell TD (1993) South Carolina's largest slave auctioneering firm. Chicago-Kent Law Rev 68 (3):1241–1282

Russell TD (1996) Articles sell best singly: the disruption of slave families at court sales. Utah Law Rev 1996(4):1161–1209
Saraydar E (1964) A note on the profitability of ante bellum slavery. South Econ J 30(4):325–332
Saraydar E (1965) The profitability of ante bellum slavery – a reply [to Sutch]. South Econ J 31(4):377–383
Schaefer DF, Schmitz MD (1979) The relative efficiency of slave agriculture: a comment. Am Econ Rev 69(1):208–212
Schmidt LB (1939) Internal commerce and the development of national economy before 1860. J Polit Econ 47(6):798–822
Stampp KM (1956) The peculiar institution: slavery in the Ante-Bellum South. Vintage Books, New York
Stampp KM (1976) A humanistic perspective. In: David PA et al (ed) Reckoning with slavery: a critical study in the quantitative history of American Negro slavery. Oxford University Press, pp 1–30
Steckel RH (1977 [1985]) The economics of U.S. slave and southern white fertility, PhD dissertation. University of Chicago 1977, published by Garland Publishing, 1985
Steckel RH (1979) Slave height profiles from coastwise manifests. Explor Econ Hist 16:363–380
Steckel RH (1980) Miscegenation and the American slave schedules. J Interdiscip Hist 11(2): 251–263. Autumn 1980
Steckel RH (1986a) A peculiar population: the nutrition, health, and mortality of American slaves from childhood to maturity. J Econ Hist 46(3):721–741
Steckel RH (1986b) A dreadful childhood: the excess mortality of American slaves. Soc Sci Hist 10(4):427–465. Winter 1986
Steckel RH, Floud R (eds) (1997) Health and welfare during industrialization. University of Chicago Press, Chicago
Sublette N, Sublette C (2016) The American slave coast: a history of the slave-breeding industry. Lawrence Hill Books, Chicago
Sutch R (1965) The profitability of ante bellum slavery – revisited. South Econ J 31(4):365–377
Sutch R (1975a) The breeding of slaves for sale and the Westward expansion of slavery, 1850–1860. In: Engerman SL, Genovese ED (eds) Race and slavery in the Western Hemisphere: quantitative studies. Princeton University Press, pp 173–210
Sutch R (1975b) The treatment received by American slaves: a critical review of the evidence presented in time on the cross. Explor Econ Hist 12:335–435
Tadman M (1989) Speculators and slaves: masters, traders, and slaves in the Old South. University of Wisconsin Press, Madison
Temin P (1967) The causes of cotton-price fluctuations in the 1830's. Rev Econ Stat 49 (4):463–470
Temin P (1999 [2008]) Interview conducted by John Brown. Newsl Cliometric Soc 14(3):3–6, 41–45. Reprinted with changes in John S. Lyons, Louis P. Cain, and Samuel H. Williamson, editors. Reflections on the Cliometric Revolution: Conversations with Economic Historians, Routledge 2008: 421–435
Toman JT (2005) The gang system and comparative advantage. Explor Econ Hist 42(2):310–323
Troesken W (2004) Water, race, and disease. MIT Press, Cambridge, MA
Trussell J, Steckel RH (1978) The age of slaves at menarche and their first birth. J Interdiscip Hist 8(3):477–505. Winter 1978
Wanamaker MH (2014) Fertility and the price of children: evidence from slavery and slave emancipation. J Econ Hist 74(4):1045–1071
Whaples R (1991) A quantitative history of the journal of economic history and the cliometric revolution. J Econ Hist 51(2):289–301
Williams HA (2003) Self-taught: African American education in slavery and freedom. University of North Carolina Press, Chapel Hill
Woodman HD (1963) The profitability of slavery: a historical perennial. J South Hist 29(3): 303–325

Wright G (1970) Economic democracy' and the concentration of wealth in the cotton south, 1850–1860. Agric Hist 44(1):63–99
Wright G (1975) Slavery and the cotton boom. Explor Econ Hist 12(4):439–452
Wright G (1976) Prosperity, progress, and American slavery. In: David PA et al (eds) Reckoning with slavery. W.W. Norton, pp 302–336
Wright G (1978) The political economy of the cotton South: households, markets, and wealth in the nineteenth century. W. W. Norton, New York
Wright G (1979) The efficiency of slavery: another interpretation. Am Econ Rev 69(1):219–226
Wright G (1986) Old South, New South: revolutions in the Southern economy since the civil war. Basic Books, New York
Wright G (2006) Slavery and American economic development. Louisiana State University Press, Baton Rouge
Wright G (2013) Sharing the prize: the economics of the civil rights revolution in the American South. Harvard University Press, Cambridge, MA
Yasuba Y (1961) The profitability and viability of plantation slavery in the United States. Econ Stud Q 12(1):60–67

Institutions

Philip T. Hoffman

Contents

Abstract

Institutions clearly play a major role in economic growth and political development. But much more needs to be done to verify and to clarify their role, and to show that it is causal, and not the result of other factors. The necessary work will involve careful historical research, the assembly of large data sets, and careful econometrics and formal modeling. And it should also involve cooperation with other social scientists, from experimental economics to anthropology and political science.

Keywords

Institutions · Economic growth · Politics · Property rights · Political science · Culture · Glorious Revolution · Douglass North · Political economy

P. T. Hoffman (✉)
California Institute of Technology (CalTech), Pasadena, CA, USA
e-mail: pth@hss.caltech.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_42

Introduction

Institutions matter for economic outcomes, and they matter a lot. The contrast between North and South Korea makes that crystal clear. It is admittedly only an example, but it is a striking one, at least for believers in revealed preference. Residents of the North risk their lives to escape the penury and malnutrition in North Korea – and the political repression too – in order to live in the South, where the 2016 per-capita GDP (corrected for purchasing power) is about the same as that in Italy or New Zealand. The yawning gap in living standards is obviously not caused by cultural differences or major differences in the level of education. Rather, it is the result of autocratic political institutions in the North. No one wants institutions like that.[1]

If social scientists now concur that institutions have a big effect on economic outcomes, their agreement comes only after a long debate over the impact institutions have. The debate began in economic history, where Douglass North was the evangelist arguing for the importance of institutions. As his message spread through economics, political science, and the other social sciences, it ultimately convinced scholars that institutions shaped economic growth, the distribution of income, and political development. North's message in turn drew considerable attention to economic history, and it is no doubt the reason why his publications are widely cited: more frequently in fact than the work of Robert Fogel – the co-winner of the Nobel Prize with North – and even more often than the writings of Kenneth Arrow, one of the most influential Nobel Laureates in economics.[2]

The debate over institutions bears closer examination, because it has implications for future research. It turns on the question of measuring the economic effect of institutions – a difficult task since institutions themselves are endogenous – and it also involves the still unresolved issue of how institutions change. We will look at the debate, beginning with how institutions are defined, and pay close attention to the critics who argue that the role of institutions has been exaggerated. When their criticisms are on target, they reveal what future research needs to be done. So does recent work on institutions in fields outside of economic history – in behavioral and experimental economics, in evolutionary anthropology, and in political science. That future research is the last topic we will take up.

[1]For further evidence, see the case study in Acemoglu et al. 2005.

[2]North's most cited article (according to Google Scholar consulted on April 12, 2018) was his 1991 essay on institutions (North 1991), which had 54,008 citations. Fogel's most cited work was his coauthored book on slavery, *Time on the Cross* (Fogel and Engerman 1995), which had 1806 citations. As for Arrow, his most referenced publication was his book on his impossibility theorem (Arrow 2012), with 17,983 citations. The citation counts here include references to earlier editions of *Time on the Cross* and Arrow's book.

What Are Institutions?

Discussion of institutions abounds in the social sciences, but often it is not clear precisely what an institution is. In sociology, an institution is typically a "complex social form" that reproduces itself – for instance, "governments, the family, ..., business corporations, and legal systems" (Miller 2014). That definition includes groups and organizations such as the family, the government, or corporations, alongside the laws and courts that make up the legal system. North's definition is narrower and more precise:

> Institutions are the humanly devised constraints that structure political, economic and social interaction. They consist of both informal constraints (sanctions, taboos, customs, traditions, and codes of conduct), and formal rules. (constitutions, laws, property rights) (North 1991, 97)

So for North, institutions are human constraints and they are distinct from organizations such as the government or the family. As in sociology, North's institutions change slowly. In his view, humans devise institutions "to create order and reduce uncertainty in exchange" but their effects need not be beneficial, for they can push economic change toward "growth, stagnation, or decline" (North 1991, 97).

North came to this definition via his research for two innovative coauthored books, one on the role of institutions in the economic history of the United States (Davis et al. 1971) and the other on the part they played in medieval and early modern Europe (North and Thomas 1973). Both works argued that economic theory could not explain economic growth and that the explanation had to involve "institutional arrangements." Both also pointed to how new institutional arrangements could foster increased efficiency or come to the rescue when markets failed. For example, the creation of a new corporate form – the corporation – would allow firms to realize economies of scale. Similarly, the establishment of insurance companies would help individuals cope with risk when markets were incomplete (North and Thomas 1973; Davis and North).

The unspoken assumption in this work from the early 1970s was that individuals create institutions to foster economic growth or move the economy toward a Pareto superior outcome when the first welfare theorem of economic theory fails to apply. Sociologists have criticized that assumption (Granovetter 1992), as have economic historians (Ogilvie 2007). But North's ultimate definition of institutions – humanly devised constraints – explicitly abandons any such premise (North 1991). For North, institutions could easily harm the economy, and subsequent research in economic history has made abundantly clear the damage bad institutions could do.[3] North's

[3]Guilds in medieval and early modern Europe provide a striking example of the damage institutions could do, by enforcing monopolies and limiting the supply of skilled workers. That is Sheilagh Ogilvie's argument (Ogilvie 2004), in response to earlier claims that guilds helped accumulate human capital (Epstein 1998). One could make a similar case for the effect of the institutions that shaped the development of health insurance in the United States, including laws that allowed the use

definition is also compatible with an analysis of social structure, which sociologists believe economists neglect. The social structure in North's view consists of organizations, such as firms, families, political coalitions, or social groups. They in turn can be studied using tools from political economy, the economics of the family, or the economics of networks.

The actual problem with North's definition is not any assumption that institutions are in any sense optimal or welfare improving. Rather, it is the risk of glossing over the question of how the humanly devised constraints are enforced. North himself recognizes the importance of enforcing rules and he acknowledges that enforcement may not be effective (North 1991). But his brief definition omits explicit mention of enforcement.

The problem of enforcement comes to the fore in Avner Greif's work on institutions, especially his 2006 book (Greif 1989, 1993, 2006a). For Greif, the key question is explaining why people follow rules. A law may outlaw theft, but why do people obey it? Saying that the threat of punishment makes people obey is insufficient, for what ensures that the police will catch the thieves and courts jail them? After all, there are clear examples – even in an orderly democracy such as the 1950s United States – of thieves operating with impunity at least in some places (Hoffman 2006). To explain that sort of outcome, Greif argues that a satisfactory definition of institutions has to incorporate beliefs and norms and organizations alongside the rules. A person will not steal, for example, if he has internalized the norm against theft or believes that he is likely to be caught and thrown in jail. That belief may in turn depend on the effectiveness of organizations (the police and courts) or on working of other norms and beliefs (for instance, the expectation that the judges and the police will not accept bribes). So for Greif, institutions end up being a system of rules, beliefs, norms, and organizations that generate regular social behavior (Greif 2006a). In practice, one could perhaps reduce his definition to rules with a means of enforcement (as for instance in Hoffman et al. 2000), but specifying why people heed the rules is essential, and the answer will involve beliefs, norms, or organizations.

The Effect of Institutions

There is abundant and persuasive qualitative evidence that institutions matter in North's early work on institutions (Davis et al. 1971; North and Thomas 1973) and in Greif's book (Greif 2006a). But critics who have doubts about the role of institutions want quantitative evidence.

Such evidence exists, and it too is persuasive, particularly microeconomic evidence concerning a portion of the economy or a particular locality. Perhaps the most convincing evidence comes from Jean-Laurent Rosenthal's study of drainage and

of fringe benefits to get around wartime wage controls and continued tax exemptions for employer provided health care (Thomasson 2003). They led to heavier spending on health care than in other developed economies for health outcomes that were often inferior.

irrigation in France (Rosenthal 1990, 1992). Before the French Revolution, there were a number of irrigation projects in southern France that were never undertaken even though they would have raised agricultural productivity and earned a return higher than that available from alternative investments. The same was true for drainage projects in northern France. The problem was not a lack of potential entrepreneurs, a failure of credit markets, or technological obstacles, because the technology (essentially unskilled labor with picks and shovels, plus some skilled work by surveyors, masons, and carpenters) was well known. Rather it was institutions. Property rights overlapped, and because the court system lacked a clear central authority that could deliver a final legal decision, property owners had an incentive to sue in order to hold water control projects up for ransom.

Profitable water control projects therefore failed, or were never started, until the French Revolution and Napoleon completely reformed the court system and property rights. The reforms halted the pattern of unending litigation over water projects and left the ultimate decision to a centralized authority with powers of eminent domain. The result was a burst of irrigation and drainage in France after 1820, even though the projects themselves were less profitable than they would have been before the Revolution's outbreak in 1789.

The obstacle blocking irrigation and drainage projects was institutional – in particular, overlapping property rights, a legal system with no finality, and no effective power of eminent domain. The root of the problem was the lack of a sovereign legislative or judicial authority that could wield effective powers of eminent domain and either resolve problems of overlapping property rights or keep them from arising in the first place. Institutions of that sort were politically impossible in pre-revolutionary France, even under a supposedly absolute monarchy. They would have violated the basic political equilibrium between the kings and elites, which rested on divided fiscal, legal, and political authority. Even changing rules of evidence in a way that would have favored drainage projects was impossible because it risked diminishing the king's tax income.

How much did that the institutional failure cost the French economy? In areas that would have benefitted the effect was large: in southern France, for example, the eighteenth-century irrigation projects would have boosted the total factor productivity of agriculture by 30–40% (Rosenthal 1992). Drainage might have had a similar effect. That is of course only agriculture, but similar problems likely prevented private entrepreneurs from building the sort of toll roads and transportation canals that proliferated in eighteenth-century England and helped the economy industrialize (Bogart 2005a, b, 2011; Bogart and Richardson 2011). The same problem plagued most of Western Europe (Epstein 2002). Only England escaped, because of its early political centralization.

Institutional failure also took a heavy toll on the textile industry in nineteenth-century Brazil and Mexico. The industry was one with no economies of scale, yet it was heavily concentrated in Mexico and Brazil. The reason was restricted access to capital markets in Mexico and Brazil. In Mexico, legal obstacles kept banks from forming and limited sales of equity to firms with connections to political leaders. Similar hurdles obstructed access to capital markets in Brazil, although they did not

last as long. The only textile firms that could form were therefore the one that could tap the wealth of the owners' families or play upon political connections. The number of such firms was small and the textile industry was therefore concentrated, in contrast to the United States, where capital markets were open and textiles firms were small (Haber 1991).

Institutions can impose costs even in wealthy democracies such as the United States. Consider, for instance, property rights to land in the United States. They are defined using two different surveying systems. One, the rectangular system, was based on a standardized rectangular grid and was imposed on new states by a 1785 Federal Land Law. Before then another system, the metes and bounds system, which is based on idiosyncratic descriptions of parcels of land, prevailed in most of the states. The state of Ohio ended up with both systems because part of Ohio was already using the metes and bounds for exogenous reasons before Ohio became a state in 1803. A careful analysis of adjacent areas employing the two systems shows that the rectangular system raised property values by up to 20–30%. Conflict over property rights was far less common with the rectangular system too. The long run consequences were large: population densities and land use patterns in areas that were otherwise similar ended up being radically different, depending on the surveying system (Libecap and Lueck 2011). The differences, in other words, seem to have been caused by the institutions that defined property rights.

The Impact of Institutions on Economic Growth

In these three examples, institutions clearly have an effect, but only on part of the economy – for instance, French agriculture in regions that would have benefitted from drainage or irrigation. What about the ramifications for the economy as a whole or for economic growth?

A seminal article by North and Barry Weingast suggested how institutions could affect economic growth for an entire economy, using the example of England's 1688 Glorious Revolution (North and Weingast 1989). By obliging the monarchy to get Parliament's assent to its actions, the Glorious Revolution constrained what English monarchs could do. The new constraint – a new institution – made credible the monarch's promise to repay loans, and as a result, government borrowing jumped by an order of magnitude. But the effects of the new institution extended well beyond the government debt market. It stimulated private capital markets too and paved the way for the British Industrial Revolution by securing all property rights, not just those of the government's debtors.

Although North and Weingast left the proof of this last assertion to other researchers, the institution that stimulated economic growth was clear. It was the constraint on a powerful ruler that made all property rights secure. Broader econometric evidence to support such a claim – evidence from well beyond seventeenth-century England – was then provided by a pair of influential articles, by Acemoglu et al. (2001, 2002). They examined countries that had once been European colonies, countries that range from ones that are rich today (as in the case of the United States)

to ones that are poor (for instance Haiti). By using evidence on urbanization rates and population densities as a proxy for real per-capita income in the past, they traced the difference in incomes today back to institutions established when the countries were colonized. Where the native population was small and incomes were initially low, Europeans settled in large numbers and established institutions that encouraged investment. Where there were a larger number of natives and incomes were initially high, the Europeans created institutions designed to extract rents, not encourage investment. The crucial feature of the institutions that encouraged investment was that they protected property rights by constraining rulers, and those institutions explain why once poor colonies like the United States ended up rich today, even when their low initial income and other factors affecting economic growth are taken into account.

Another seminal argument traced important differences in institutions back to yet another ultimate cause – factor endowments at the time of colonization (Engerman and Sokoloff 1997, Engerman et al. 2012). Like Acemoglu, Johnson, and Robinson, Engerman and Sokoloff examine colonies, this time in the Americas only. In some American colonies, factor endowments created extreme inequality; for instance, when natives could be forced to work in silver or mercury mines, as in parts of Spanish America, or when slaves could be imported to grow sugar, as in Haiti. In other colonies – Canada and what is now the northeastern United States – sugar could not be grown, and native populations were too small to exploit for forced labor. Inequality among a population that soon became overwhelmingly European was then relatively small.

Where inequality was extreme, the European settlers fashioned institutions that benefitted the wealthy elite and limited the poor's political influence and their economic opportunity. For example, the elite might require voters to be literate and then limit public funding for education. If inequality was high, the poor would be unable to borrow to finance their children's education and as a result, they would be disenfranchised. In addition, education levels would be low, and economic growth would suffer.

A final example of the impact institutions could have on economies as a whole comes from Nathan Nunn's work on the consequences of the African slave trade (Nunn 2008). Countries where large numbers of slaves were seized in the past turn out to be significantly poorer today, even when other factors affecting per-capita incomes are controlled for. A plausible instrumental variable suggests that the relationship was causal, and historical evidence points to two ways for the slave trade to have harmed the economy. The first was that it destroyed nascent states in Africa. The second was that it heightened ethnic fractionalization and hindered the formation of broader ethnic groups. The result was a prevailing norm of distrust of other ethnic groups in areas where slaves had been captured, and modern survey research supports that claim (Nunn and Wantchekon 2011). The norm of distrust was an institution that worked against long distance trade. It also hampered the formation of states, as did the destruction of early African states. So areas where slaves were taken would end up with weak states that could not protect trade or property rights or settle cross ethnic disputes, and they would end up much poorer.

Criticisms of the Claims for Institutions

The work on the effect of institutions is widely cited, but it has also been criticized, particularly when it comes to the impact on economies as a whole. The studies of particular regions or sectors of the economy are hard to attack. For example, once the French Revolution changed institutions, successful irrigation and drainage projects proliferated, even though rates of return had dropped. It is difficult to argue with that sort of evidence.

Assertions about the whole economy or about long-run economic growth, however, are more vulnerable to criticism, and some of the critics are right on the mark. Others, though, fail to hit the target.

Let us start with criticisms that fail. A number of them aim at North's and Weingast's claim about the Glorious Revolution. Greg Clark (1996), for instance, contested their argument that the Glorious Revolution made private capital markets secure. His evidence came from the rates of return on land and from loans backed by real estate as collateral. The returns did not rise during periods of political turmoil, as they presumably would if property rights were insecure, and although returns were dropping in England, the decline did not accelerate after the Glorious Revolution, as they would if North and Weingast were correct. For Clark, the implication is that property rights were not secured by the Glorious Revolution; they had in fact long been secure.

Upon closer inspection, though, Clark's evidence falls apart. His evidence comes from the safest capital market, the market for land and loans backed by real estate. That market had little or nothing to with the capital markets North and Weingast are concerned with, the markets for government debt and investment in private companies. More importantly enforcement of the relevant contracts in Clark's capital market – tenancy and mortgage contracts – had been settled in the Middle Ages. It was a matter for lawyers and common law courts, not the monarch, who had no incentive to interfere (Cox 2016). Clark's evidence, in short, is simply not relevant to the Glorious Revolution.

There is evidence about private rates of return, though, that is harder to dismiss, and evidence too that property rights in capital markets were not secured overnight in 1688 (Quinn 2001; Sussman and Yafeh 2006; Stasavage 2003, 2007; Pincus and Robinson 2011; Murphy 2013). Secure property rights in capital markets required the lobbying by creditors and a powerful Whig Party in Parliament, all of which took time to achieve. Gary Cox (2016) makes sense of the whole process in what is perhaps the ultimate word on the political economy of the Glorious Revolution. The issue, as in North and Weingast, was making sovereign promises credible. Doing that involved giving Parliament ultimate authority over the promises. That entailed establishing Parliament's exclusive right to determine who could arrange sales of these promises and limiting the authority of ministers and other royal officials so that they could not interfere or evade Parliamentary control. These changes took time, but they raised Britain's tax revenues dramatically (Dincecco 2009, 2011) and had a similar effect on her ability to borrow via long term debt.

That the Glorious Revolution boosted Britain's tax revenues (even when one controls for other characteristics of the British economy) also casts doubt on Robert Allen's criticism of North and Weingast (Allen 2009). Having gathered data on European economies from the late Middle Ages to the nineteenth century, Allen estimated a system of four linear equations to explain urbanization, real wages, agricultural productivity, and the amount of rural industry. Among his explanatory variables was an indicator variable that took the value 1 when the executive authority in a given economy (typically a monarch) was constrained by representative institutions, as in Britain after the Glorious Revolution. When Allen used the coefficients from the regression to gauge the effect of representative institutions, he found it had only a minimal effect on urbanization and real wages – both reasonable proxies for per-capita GDP that play a major role in Allen's explanation of the Industrial Revolution.

One problem with this test is that an indicator variable that suddenly takes the value 1 after a certain date (for instance after 1688 in Britain) does not capture all the changes involved in making sovereign promises credible, at least from what we know of the Glorious Revolution. Cox's research (2016) suggests that the process was not instantaneous elsewhere in Europe either. There is another serious problem with Allen's test as well. One of the variables that does have a big impact on urbanization and real wages is per-capita intercontinental trade. But Britain won a lion's share of intercontinental trade in the eighteenth century because of military victories, victories against France in particular. It won those wars because it could levy much heavier taxes than France – heavier even as a fraction of GDP. And it could impose such heavy taxes because of the political changes that gave ultimate power to Parliament; representative institutions clearly increased states' capacity to tax (Hoffman and Norberg 2002; Dincecco 2009, 2011; Hoffman 2015). Intercontinental trade and other explanatory variables in Allen's regressions are endogenous and clearly influenced by institutions. So institutions are one of the ultimate causes behind Britain's high wages and extensive urbanization, and hence part of Allen's explanation for the British Industrial Revolution.

Those are the unsuccessful criticisms of the claims about institutions. What about the successful ones? Some of the successful ones are simply telling case studies. North and Weingast, for instance, argued for a connection between public debt markets and capital markets in general. If a monarch's promises to repay are credible, then the rights of investors in private capital markets are secure too, and private capital markets will prosper. In the eighteenth century, though, France had a thriving market for private debt, even though its monarchy defaulted repeatedly (Hoffman et al. 2000). And nineteenth-century Brazil's constitutional monarchy borrowed extensively at home and abroad without defaulting. Nonetheless, the private capital market in Brazil failed to thrive (Summerhill 2015). These two counterexamples cast doubt on the connection that North and Weingast emphasized. Having a sovereign who is a credible borrower is, in short, neither a necessary nor a sufficient condition for a flourishing private capital market.

There are reasons to criticize other work on institutions as well. One problem is a tendency to generalize hastily from limited data. A clear example comes from the

allegations about the causal role of the so-called European Marriage Pattern in stimulating economic growth. In parts of Western Europe before the Industrial Revolution, families were nuclear, celibacy rates were high, and women married late. These demographic patterns – named the European Marriage Pattern – have been invoked to explain why Europe industrialized early.[4] The virtue of the argument was that it fit an important model of economic growth and demographic change: unified growth theory (Galor and Weil 1999; Galor 2005). Yet although the model was fine, the evidence used to link it to Europe's industrialization does not stand up to scrutiny (Dennison and Ogilvie 2014). Dennison and Ogilvie's analysis of data from 39 European countries shows that the clearest examples of the European marriage pattern are in fact correlated with economic stagnation, not growth. Economic historians had, in short, generalized from a tiny number of cases – a mistake that led them astray.

Another problem is a recurrent one: gaps in the evidence connecting an institutional cause in the past and an economic outcome today. Research that connects institutions to modern outcomes typically relies on regressions where one of the explanatory variables is a variable measured in the distant past. It might be a characteristic of past institutions or a variable whose past value shapes the sort of institutions that eventually emerge in a country. In Acemoglu et al. (2002), for example, low native population densities at the time of colonization made it easier for Europeans to settle in new colonies; the settlers would then lobby for institutions that protected their property rights. Higher native population densities made it easier to take over tax systems of existing native states or to force the natives to work. The colonists would then design institutions to extract resources from the colony, not to protect property rights. A regression of current per-capita income on correlates of economic growth and past population density should therefore yield a negative coefficient for past population density, because of the effect the past density had on the development of institutions.

The trouble is the lack of data to follow the institutions over time.[5] Without such data, it is impossible to tell whether the bad institutions, once they were created because of high population densities, ended up lasting past the end of the colonial regime and on into independence, even though newly independent colonies often received new constitutions modeled after those of wealthy democracies. And

[4]Voigtländer and Voth 2006, 2013; De Moor and van Zanden 2010; Foreman-Peck 2011 are the relevant papers here. All were criticized by Dennison and Ogilvie for basing an argument on a generalization that does not hold up to scrutiny. Dennison and Ogilvie also have concerns about Greif (2006b) and Greif and Tabellini (2010), which also mention the nuclear family. But Greif (2006b) links the nuclear family to the corporation, a more limited and defensible claim, and Greif and Tabellini (2010) takes up a different topic – the contrasting way in which cooperation was achieved in China (via clans) and Europe (via cities).

[5]There are other worries as well, both for this particular example and for other articles linking past institutions to modern outcomes. First of all, the data (typically urbanization or population densities) is often suspect outside of Europe. Second, the regressions typically involve instruments because institutions are endogenous; the instruments, though, may be problematic as well. See, for example, Albouy 2012.

institutions in ex colonies did change after independence, in a way that might be consistent with a different argument, namely, that increases in human capital caused both economic growth and improvement in institutions (Glaeser et al. 2004).

The criticisms here are not peculiar to the influential 2002 paper by Acemoglu et al. Other work on institutions is vulnerable to similar attacks. To take an important example, consider Dell's impressive article (2010) on the effects of forced labor in colonial Latin America. She aimed to show what effect a bad institution – a colonial forced labor requirement – had on modern outcomes, and also to explain why this effect persisted. To do so, she contrasted outcomes in an area of Bolivia and Peru that had to provide labor for silver and mercury mines with an adjacent area that was exempt from the labor exactions. By using a regression discontinuity design and colonial evidence that the two areas were similar before just the labor requirement was imposed, she isolated the effect of the forced labor requirement, which was large: it lowered household consumption by 25%.

As to why the effect persisted after the forced labor requirement was abolished in 1812, Dell points to a colonial policy that restricted the formation of *haciendas* (large rural estates worked with forced labor) in areas subject to the labor requirement. The policy aimed to minimize competition with the authorities who sought coerced native labor for the mines. In the areas exempt from the labor requirement, elites formed haciendas and got secure property rights. They also lobbied for public goods such as roads. By contrast, in the areas that had to furnish mine labor, there was no one to lobby for roads, and the prevailing property rights were traditional communal land tenures that were abolished after the labor requirements ended. Property rights were therefore insecure. The result would be that the area with haciendas and rich elite would end up with higher incomes – a sharp contrast to Engerman and Sokoloff's argument.

The virtue of Dell's article is that it does lay out an explanation for why bad institutions persisted. But there are still gaps in the quantitative data, in particular between the end of the labor requirement in 1812 and the modern era. Before 1812, there were penalties for moving, but once the penalties disappeared in 1812 why did people not abandon the region with no roads and insecure property rights? Movement of individuals is the bane of regression discontinuity design when the discontinuity follows a geographic boundary. If only the poor or the uneducated stayed behind in the labor requirement area, then we might have a very different explanation for its low income today. Only more historical data could tell for sure and determine whether her argument holds up.

Another example of a gap in the data between a current outcome and a cause in the past comes from the influential literature on legal origins (La Porta et al. 1997, 1998). As La Porta and his coauthors have demonstrated, the size of financial markets today is correlated with the sort of legal system a country uses, and the correlation persists even after controlling for per-capita income and other variables. Legal systems based on British common law, which prevails in Britain and in former British colonies, such as the United States, favor the development of financial markets, apparently because they do more to protect investors. The other great family of legal systems – civil law, which is used in continental Europe, in the

ex-colonies of continental European powers, and in countries such as Japan – is not as effective at defending investors, and it leads to smaller financial markets. These legal systems were determined in the past – for instance, during colonization – but their effects survive today.

Implicit here is the assumption that the effect of the legal systems persist, so that if the civil law weakens protection for investors, it should do so in the past as well as today. But Aldo Musacchio's (2008, 2009) careful historical study of one civil law country – Brazil – suggests that is not true. He gathered data on investor protection in the past to fill in the gap between the current outcome and the past cause and found that Brazil did defend investors in the late nineteenth and early twentieth centuries, contrary to what the legal origins literature assumes. Brazilian creditors had strong legal rights, and the Brazilian bond market grew. Brazilian shareholders ended up being protected too, through corporate bylaws; Brazil's stock market boomed too. And Brazil, as Musacchio and Rajan and Zingales (2003) have shown, was not an exception. Other civil law countries had thriving financial markets in 1900 too, suggesting at the very least that other causes have to intervene to explain why the protection that is possible with the civil law ended up failing after 1900.

What is needed here is better historical research to fill in the gaps between the distant causes and the current outcomes – more research, for instance, on institutional change and political economy of colonies after independence. For several topics that research is already underway – for instance, on business organizations that have been linked both to legal systems and to economic development. It has been argued that the corporation is, as an institution, the superior form of business organization for economic development and that the common law favored the establishment of corporations. Timothy Guinnane and his coauthors (Guinnane et al. 2007) demonstrate that the historical reality is more complex. The corporation did have virtues but it also had disadvantages, and another type of organization – the private limited liability company – was often better for small and medium sized firms. But the common law, so their evidence suggests, actually worked against creating private limited liability companies.

We also need better models of how political institutions foster economic growth. Both North and Weingast (1989) and Acemoglu et al. (2001, 2002) argue that institutions fostered growth, but the way that happens needs to be fleshed out, with more historical research and with models that can explain how institutions protect property rights or promote economic growth – ideally models that can be generalized.

Cox (2016) models what happened after the Glorious Revolution in Britain, and his model does generalize to the rest of western Europe, as his historical research demonstrates. North et al. (2009) lay out a more general conceptual framework to understand both political and economic development. In their framework, the fundamental problem before 1800 is controlling violence, and the solution to this problem is the same in most societies. It consists of having violence controlled by a powerful coalition of elites held together by economic privileges that elites enjoy and that they would lose if coalition members split apart and began fighting with one another. Economic development requires both a political and an economic transition,

and the political transition is fundamental. It happens when the elite is opened up to both political and economic competition. Their framework could certainly be formalized and tested with quantitative data.

Acemoglu et al. (2005) have devised a similar argument about both political and economic development. For them, the fundamental causes of economic growth are political institutions and the distribution of resources. Political institutions determine de jure political power (the political power enshrined in laws and constitutions); the distribution of resources determines de facto political power (the unwritten power of the wealthy). Together, these two forms of power give rise to both current day economic institutions and future political institutions. The economic institutions then dictate economic performance today and the future distribution of resources. They illustrate this verbal framework with case studies and references to research that supports their claims or elaborates on them. Their framework could be formalized too, and the research they refer to shows how, with Acemoglu and Robinson (2005) being a particularly good example.

To improve our understanding of how institutions promote growth, researchers should also move beyond the fixation on protecting property rights, because growth may demand more than just safety for one's possessions and investments. Personal security is important too, as is access to a predictable and unbiased means of settling disputes. To assure all these things, the state will have to be powerful enough to settle disputes and protect lives and property; it should also be able to intervene when property rights overlap (for instance, for the construction of infrastructure) by using eminent domain. All of these powers will necessitate a state strong enough to impose taxation, and yet the state will have to wield all these powers without threating property or people, and it will have to use its powers in a way that encourages cooperation. So far the only study that begins to explain in detail how all this happens is Cox's (2016) study of the Glorious Revolution.

Explaining Institutional Change

One issue common to most of these criticisms is the question of explaining institutional change. If bad institutions in the past yield poverty or autocracy today, what has kept those institutions from changing? If good institutions arise and then produce wealth and democracy, what gave rise to the good institutions in the first place? What, in short makes institutions change and what locks them into place, even when they make people worse off?

For Avner Greif (2006a), the tools needed to answer these questions are game theory and careful historical analysis. Both are necessary: the game theory because institutions involve interactions among individuals where beliefs and regularities of behavior are important; the detailed historical research because institutions arise, change, and survive in particular societies, where historical details wield enormous power. The importance of history makes it difficult to generalize about institutional change. Only careful historical research will explain why institutions are born or persist even when the consequences are bad. But some institutions, Greif argues

(Greif 2006a) will be self-reinforcing in the sense that more individuals will find themselves in situations where they want to behave in the regular way that is associated with the institution. That sort of institution will persist; in a similar way, other institutions can actually undermine themselves and disappear.

Greif himself analyzes specific examples of how institutions arise, persist, or fall – institutions such as the community responsibility for debt that was common in medieval Europe (Greif 2006a). Community responsibility made whole groups responsible for an individual member's bad debts, particularly when long distance trade was involved. All Flemish merchants in England might be liable, for instance, if one of them had left England without paying his creditors; similar rules could apply even to merchants from other towns in the same country. But by facilitating interactions between an increasing number of communities, the institution of community responsibility lost its effectiveness. It became harder to verify that a merchant really was a member of a known community and easier to falsify a community affiliation. The institution dug its own grave and disappeared in the late thirteenth century.

Another example comes from eighteenth and early nineteenth-century France, where mortgages all paid one interest rate – 5%. Eighteenth and nineteenth-century France had a thriving mortgage market (Hoffman et al. 2019). In 1740, it allowed a third of French families to borrow, and in 1840 it was mobilizing as much credit for mortgages (relative to GDP) as the United States banking system did in the 1950s. But for most of the two centuries, the market was not cleared by prices. The 5% rate was common to all mortgages.

That regularity of behavior was an institution, one that had arisen in the 1660s, when the government lowered the maximum legal interest rate from 6.3% to 5% on what was the most common mortgage loan at the time – perpetual annuities. With the 6.3% legal limit, interest rates on the annuities had varied between 5% and 6.3%. But once the 5% maximum was in force, interest rate variation disappeared for two reasons. One, clearly, was the 5% maximum rate. The other was a serious problem of asymmetric information. Lenders did not necessarily know what collateral was worth or whether borrowers would repay their loans. In such a market, price competition may disappear, and credit will be rationed on the basis of creditworthiness and the value of collateral. Borrowers who had good collateral and seemed likely to repay would get loans at 5%; those who did not measure up would not. In France, there were mortgage brokers who had information about collateral and creditworthiness, and they solved the information problem by matching up lenders with reliable borrowers, often by referring a trustworthy borrower to another broker. That made all the lending possible. And they had no incentive to deviate from the 5% equilibrium rate because it would raise questions about the reliability of their borrowers and their referrals.

The institution was, in short, self-reinforcing, and it only disappeared at the end of the nineteenth century for several reasons – all of them exogenous. First, the government created a mortgage bank, granted it a monopoly on the issue of mortgage backed securities, and gave those securities implicit government backing. The mortgage bank could draw upon publicly available information about collateral and past mortgages that grew more plentiful in the nineteenth century, so it could do without the mortgage brokers' information, siphon off the best borrowers, and then

offer them loans at below 5% by raising money via its government-backed securities. Second, an agricultural depression in the 1880s cut demand for mortgages in the countryside and strengthened the hand of the best borrowers in cities, who could demand interest rates below 5% from mortgage brokers. Third, by 1899, returns on liquid government bonds had fallen to 3.5%; the low rate on government bonds made riskier 5% loans to less creditworthy borrowers attractive investments. Fourth, in the early twentieth century, the government began to intervene directly in the mortgage market to provide subsidized mortgages at below 5% to selected borrowers. All four exogenous factors made the 5% equilibrium unravel, and the institution disappeared.

Religion can also give rise to institutions by establishing organizations, rules and norms of behavior, and beliefs about how others will act. Recent work on the Islamic world offers two examples. For Timur Kuran (2012), Islamic commercial law blocked the establishment of large joint stock companies, making it harder for Muslims to take advantage of economies of scale. The law could not simply be changed, because it was the accidental result of inheritance rules spelled out in the Koran. Jared Rubin (2017) has pointed to a different institutional contrast between western Europe and the Islamic world. In his view, Islamic rulers tended to rely more heavily on religious authorities to establish their legitimacy than Christian rulers did. The reasons, again, were largely accidental: from the outset, Christianity distinguished between secular and religious powers, and the Reformation ended up diminishing the independent political power of the Christian clergy, even in Catholic lands. The consequences, Rubin argues, were large: religious authorities were generally conservative and they prohibited new ideas or technologies that threaten their authority. In the Islamic world, for instance, they banned the printing press. In the West, they simply had less power to block innovation.

New directions

There is a great deal of research on institutions left to do. Among the urgent tasks is to answer the critics of the claims for institutions. That will entail extending the analysis of the long-run impact of institutions by formalizing models of why they persist or change, and by undertaking careful historical research to fill in the gaps between the institution established in the past and the current day outcome. Both types of research are necessary to measure the true causal role of institutions and to understand why the effects of institutions are not diluted by other forces. Why bad colonial institutions survive past independence is one topic that obviously deserves study. Another is the persistence of institutions with bad outcomes, such as the United States health care system.

A third promising avenue for research is to take up the issue of culture. Economic historians such as Joel Mokyr and Avner Greif have long argued for the role that culture plays in changing institutions or locking them into place (Mokyr 2017; Greif 2006a; Greif and Tabellini 2010). Culture may also shed light on one of the fundamentals in the working of institutions: making sense of why people cooperate

when pure self-interest would seem to rule out such behavior. People may, for instance, contribute to a public good when shirking would seem to be the dominant strategy, and they may do so not just in laboratory experiments but in the real world when their lives are at stake. Contributing to the public good may in turn establish norms or beliefs about behavior – an institution.[6] One clear example comes from Dora Costa and Matthew Kahn's illuminating analysis of desertions during the Civil War: despite high mortality rates and minimal penalties for desertion, many Union army soldiers did not desert. Why they risked their lives – even though others did desert – was linked to the social homogeneity of a unit, which created a norm of honorable service and a fear of shame that helped soldiers to serve honorably (Costa and Kahn 2003a, b, 2010).

Research on culture in behavioral and experimental economics offers tools for understanding this sort of behavior; so does parallel research in cultural anthropology (Bowles and Gintis 2011; Boyd and Richerson 1988; Henrich et al. 2001; Camerer 2011). By using this literature and related work in experimental economics, Hoffman (2015) has explained how warrior leaders came to dominate Western Europe after the collapse of the Roman Empire and how these leaders could wage war despite having no system of permanent taxation – in other words, how the institution of "feudalism" originated. In a different vein, Marco Casari has employed his own work in experimental economics (Casari and Plott 2003) and his own detailed historical research (Casari 2007; Casari and Tagliapietra 2018) to lay bare the origins and working of the decentralized institutions that for centuries governed communal property rights to forests and pasture in communities in northern Italy. Users with rights could, at their own expense, report infringements on the commons, and collect a fine if the infringement was unjustified. The working of the institutions required group decisions, and the groups worked best when they remained a bit below 200 individuals. When they grew too large, they tended to divide to facilitate future decision making.

Besides culture and behavioral economics, cliometricians would also profit by cooperating with researchers in political science. The relevant political scientists are trained in economics and use econometrics and models from economics. They also do original historical research, some of it with large, original data sets. Cox (2016) is a model here of combining models with careful historical research. Other examples include David Stasavage's excellent study (2011) of how the size of states affected the development of representative institutions and the ability of states to issue sovereign debt.

As for political scientists' relevant work with large historical data sets, Stasavage has also done important research on how progressive taxation derived not from the extension of the suffrage to lower income voters (an institutional change that models of the median voter would emphasize), but rather from mass mobilization warfare (Scheve and Stasavage 2010). An even more exciting political science project with large historical data sets is Steve Haber's research (see Elis et al. 2017) linking political and economic development to natural endowments. The connection

[6]For historical and modern examples of this sort of behavior and a discussion of the relevant literature in experimental and behavioral economics and cultural anthropology, see Hoffman 2015.

suggested by their research runs through institutions, but it was not deterministic. The development of institutions favoring democracy and economic growth would be more likely with the right endowments (a climate that produced large storable crop surplus without large scale floods or droughts; level terrain and access to water transport) at the right time (circa 1800). Democracy and high incomes would then be more likely too, but they would not be inevitable. The relationship would be complex and probabilistic.

Conclusion

Despite all the criticism, institutions clearly play a major role in economic growth and political development. But much more needs to be done to verify and to clarify their role, and to show that it is causal, and not the result of other factors. The necessary work will involve careful historical research, the assembly of large data sets, and careful econometrics and formal modeling. And it should also involve cooperation with other social scientists, from experimental economics to anthropology and political science.

Cross-References

- ▶ Analytic Narratives
- ▶ Cliometrics of Growth
- ▶ Douglass North and Cliometrics
- ▶ Early Capital Markets
- ▶ Economic History and Economic Development: New Economic History in Retrospect and Prospect
- ▶ History of Cliometrics
- ▶ Path Dependence
- ▶ Political Economy
- ▶ The Industrial Revolution: A Cliometric Perspective

References

Acemoglu D, Robinson JA (2005) Economic origins of dictatorship and democracy. Cambridge University Press, Cambridge

Acemoglu D, Johnson S, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. Am Econ Rev 91(5):1369–1401

Acemoglu D, Johnson S, Robinson JA (2002) Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q J Econ 117(4):1231–1294

Acemoglu D, Johnson S, Robinson JA (2005) Institutions as a fundamental cause of long-run growth. In: Aghion P, Durlauf SN (eds) Handbook of economic growth, vol 1. Elsevier, Amsterdam/New York, pp 385–472

Albouy DY (2012) The colonial origins of comparative development: an empirical investigation: comment. Am Econ Rev 102(6):3059–3076
Allen RC (2009) The British industrial revolution in global perspective, vol 1. Cambridge University Press, Cambridge
Arrow KJ (2012) Social choice and individual values, vol 12. Yale University Press, New Haven
Bogart D (2005a) Turnpike trusts and the transportation revolution in 18th century England. Explor Econ Hist 42(4):479–508
Bogart D (2005b) Did turnpike trusts increase transportation investment in eighteenth-century England? J Econ Hist 65(2):439–468
Bogart D (2011) Did the Glorious Revolution contribute to the transport revolution? Evidence from investment in roads and rivers. Econ Hist Rev 64(4):1073–1112
Bogart D, Richardson G (2011) Property rights and parliament in industrializing Britain. J Law Econ 54(2):241–274
Bowles S, Gintis H (2011) A cooperative species: human reciprocity and its evolution. Princeton University Press, Princeton
Boyd R, Richerson PJ (1988) Culture and the evolutionary process. University of Chicago press, Chicago
Camerer CF (2011) Behavioral game theory: experiments in strategic interaction. Princeton University Press, Princeton
Casari M (2007) Emergence of endogenous legal institutions: property rights and community governance in the Italian alps. J Econ Hist 67(1):191–226
Casari M, Plott CR (2003) Decentralized management of common property resources: experiments with a centuries-old institution. J Econ Behav Organ 51(2):217–247
Casari M, Tagliapietra C (2018) Group size in social-ecological systems. Proc Natl Acad Sci. February 22:201713496. https://doi.org/10.1073/pnas.1713496115. Consulted 28 Apr 2018
Clark G (1996) The political foundations of modern economic growth: England, 1540–1800. J Interdiscip Hist 26(4):563–588
Costa DL, Kahn ME (2003a) Cowards and heroes: group loyalty in the American Civil War. Q J Econ 118(2):519–548
Costa DL, Kahn ME (2003b) Civic engagement and community heterogeneity: an economist's perspective. Perspect Polit 1(1):103–111
Costa DL, Kahn ME (2010) Heroes and cowards: the social face of war. Princeton University Press, Princeton
Cox GW (2016) Marketing sovereign promises: monopoly brokerage and the growth of the English state. Cambridge University Press, New York
Davis LE, North DC, Smorodin C (1971) Institutional change and American economic growth. Cambridge University Press, Cambridge
De Moor T, Van Zanden JL (2010) Girl power: the European marriage pattern and labour markets in the North Sea region in the late medieval and early modern period. Econ Hist Rev 63(1):1–33
Dell M (2010) The persistent effects of Peru's mining mita. Econometrica 78(6):1863–1903
Dennison T, Ogilvie S (2014) Does the European marriage pattern explain economic growth? J Econ Hist 74(3):651–693
Dincecco M (2009) Fiscal centralization, limited government, and public revenues in Europe, 1650–1913. J Econ Hist 69(1):48–103
Dincecco M (2011) Political transformations and public finances: Europe, 1650–1913. Cambridge University Press, Cambridge
Elis R, Haber S, Horrillo J (2017) Climate, geography, and the evolution of economic and political systems. In: Paper delivered at Barnard College
Engerman SL, Sokoloff KL (1997) Factor endowments, institutions, and differential paths of growth among new world economies. In: Haber S (ed) How Latin America fell behind. Stanford University Press, Palo Alto, pp 260–304
Engerman SL, Sokoloff KL et al (2012) Economic development in the Americas since 1500. Cambridge University Press, Cambridge
Epstein SR (1998) Craft guilds, apprenticeship, and technological change in preindustrial Europe. J Econ Hist 58(3):684–713

Epstein SR (2002) Freedom and growth: the rise of states and markets in Europe, 1300–1750, vol 17. Routledge, Abingdon
Fogel RW, Engerman SL (1995) Time on the cross: the economics of American Negro slavery, vol 1. WW Norton & Company, New York
Foreman-Peck J (2011) The Western European marriage pattern and economic development. Explor Econ Hist 48(2):292–309
Galor O (2005) From stagnation to growth: unified growth theory. In: Aghion P, Durlauf SN (eds) Handbook of economic growth, vol 1. Elsevier, Amsterdam/New York, pp 171–293
Galor O, Weil DN (1999) From Malthusian stagnation to modern growth. Am Econ Rev 89(2):150–154
Glaeser EL, La Porta R, Lopez-de-Silanes F, Shleifer A (2004) Do institutions cause growth? J Econ Growth 9(3):271–303
Granovetter M (1992) Economic institutions as social constructions: a framework for analysis. Acta Sociol 35(1):3–11
Greif A (1989) Reputation and coalitions in medieval trade: evidence on the Maghribi traders. J Econ Hist 49(4):857–882
Greif A (1993) Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. Am Econ Rev 83:525–548
Greif A (2006a) Institutions and the path to the modern economy: lessons from medieval trade. Cambridge University Press, New York
Greif A (2006b) Family structure, institutions, and growth: the origins and implications of western corporations. Am Econ Rev 96(2):308–312
Greif A, Tabellini G (2010) Cultural and institutional bifurcation: China and Europe compared. Am Econ Rev 100(2):135–140
Guinnane T, Harris R, Lamoreaux NR, Rosenthal JL (2007) Putting the Corporation in its Place. Enterp Soc 8(3):687–729
Haber SH (1991) Industrial concentration and the capital markets: a comparative study of Brazil, Mexico, and the United States, 1830–1930. J Econ Hist 51(3):559–580
Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, McElreath R (2001) In search of homo economicus: behavioral experiments in 15 small-scale societies. Am Econ Rev 91(2):73–78
Hoffman PT (2006) Institutions and the Path to the Modern Economy: lessons from Medieval Trade: review of Avner Greif. Institutions and the Path to the Modern Economy. EH.net (August). Consulted 14 April 2018
Hoffman PT (2015) Why did Europe conquer the world? Princeton University Press, Princeton
Hoffman PT, Norberg K (2002) Fiscal crises, liberty, and representative government, 1450–1789. Stanford University Press, Palo Alto
Hoffman PT, Postel-Vinay G, Rosenthal JL (2000) Priceless markets: the political economy of credit in Paris, 1660–1870. University of Chicago Press, Chicago
Hoffman PT, Postel-Vinay G, Rosenthal JL (2019) Dark matter credit: the development of peer-to-peer lending and banking in France. Princeton University Press, Princeton
Kuran T (2012) The long divergence: how Islamic law held back the Middle East. Princeton University Press, Princeton
La Porta R, Lopez-de-Silanes F, Shleifer A, Vishny RW (1997) Legal determinants of external finance. J Financ 52(3):1131–1150
La Porta RL, Lopez-de-Silanes F, Shleifer A, Vishny RW (1998) Law and finance. J Polit Econ 106(6):1113–1155
Libecap GD, Lueck D (2011) The demarcation of land and the role of coordinating property institutions. J Polit Econ 119(3):426–467
Miller S (2014) Social Institutions. In: Zalta EN (ed) The Stanford encyclopedia of philosophy, (Winter 2014 edn). https://plato.stanford.edu/archives/win2014/entries/social-institutions/. Consulted 13 Apr, 2018
Mokyr J (2017) A culture of growth: the origins of the modern economy. Princeton University Press, Princeton
Murphy AL (2013) Demanding 'credible commitment': public reactions to the failures of the early financial revolution. Econ Hist Rev 66(1):178–197

Musacchio A (2008) Can civil law countries get good institutions? Lessons from the history of creditor rights and bond markets in Brazil. J Econ Hist 68(1):80–108
Musacchio A (2009) Experiments in financial democracy: corporate governance and financial development in Brazil, 1882–1950. Cambridge University Press, Cambridge
North DC (1991) Institutions. J Econ Perspect 5(1):97–112
North DC, Thomas RP (1973) The rise of the western world: a new economic history. Cambridge University Press, New York
North DC, Weingast BR (1989) Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. J Econ Hist 49(4):803–832
North DC, Wallis JJ, Weingast BR (2009) Violence and social orders: a conceptual framework for interpreting recorded human history. Cambridge University Press, New York
Nunn N (2008) The long-term effects of Africa's slave trades. Q J Econ 123(1):139–176
Nunn N, Wantchekon L (2011) The Slave trade and the origins of Mistrust in Africa. Am Econ Rev 101(7):3221–3252
Ogilvie S (2004) Guilds, efficiency, and social capital: evidence from German proto-industry. Econ Hist Rev 57(2):286–333
Ogilvie S (2007) 'Whatever is, is right'? Economic institutions in pre-industrial Europe. Econ Hist Rev 60(4):649–684
Pincus SC, Robinson JA (2011) What really happened during the glorious revolution? Working paper w17206. National Bureau of Economic Research
Quinn S (2001) The Glorious Revolution's effect on English private finance: a microhistory, 1680–1705. J Econ Hist 61(3):593–615
Rajan RG, Zingales L (2003) The great reversals: the politics of financial development in the twentieth century. J Financ Econ 69(1):5–50
Rosenthal JL (1990) The development of irrigation in Provence, 1700–1860: the French Revolution and economic growth. J Econ Hist 50(3):615–638
Rosenthal JL (1992) The fruits of revolution: property rights, litigation and French agriculture, 1700–1860. Cambridge University Press, Cambridge
Rubin J (2017) Rulers, religion, and riches: why the West got rich and the Middle East did not. Cambridge University Press, New York
Scheve K, Stasavage D (2010) The conscription of wealth: mass warfare and the demand for progressive taxation. Int Organ 64(4):529–561
Stasavage D (2003) Public debt and the Birth of the democratic state: France and Great Britain 1688–1789. Cambridge University Press, Cambridge
Stasavage D (2007) Partisan politics and public debt: the importance of the 'Whig Supremacy'for Britain's financial revolution. Eur Rev Econ Hist 11(1):123–153
Stasavage D (2011) States of credit: size, power and the development of European polities. Princeton University Press, Princeton
Summerhill WR (2015) Inglorious revolution: political institutions, sovereign debt, and financial underdevelopment in imperial Brazil. Yale University Press, New Haven
Sussman N, Yafeh Y (2006) Institutional reforms, financial development and sovereign debt: Britain 1690–1790. J Econ Hist 66(4):906–935
Thomasson MA (2003) The importance of group coverage: how tax policy shaped US health insurance. Am Econ Rev 93(4):1373–1384
Voigtländer N, Voth HJ (2006) Why England? Demographic factors, structural change and physical capital accumulation during the Industrial Revolution. J Econ Growth 11(4):319–361
Voigtländer N, Voth HJ (2013) How the West "Invented" fertility restriction. Am Econ Rev 103(6):2227–2264

Political Economy

Mark Koyama

Contents

This chapter was completed while I was a W. Glenn Campbell and Rita Ricardo-Campbell Fellow at the Hoover Institution.

M. Koyama (✉)
Department of Economics, George Mason University, Fairfax, VA, USA
e-mail: mkoyama2@gmu.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_54

Abstract

This chapter surveys research on political economy in economic history. It discusses the integration of the public choice/political economy approaches with economic history. It provides a thematic survey of topics such as the origins of the state, different regime types, labor coercion, warfare, religion, and state capacity. The chapter also provides detailed illustrations of how economic historians have investigated specific historical episodes such as the Glorious Revolution, French Revolution, the consequences of European empires, and the rise of democracy.

Keywords

Political economy · Public choice · State capacity · Conflict

Introduction

Economic historians have always been interested in political economy. However, for a long time, political economy was resistant to cliometric approaches. In the work of Frederick Lane (1958) and John Hicks (1969), economic theory was a metaphor to illuminate topics such as conflict, war, and state development. This approach was insightful, but it stopped short of the aim of cliometrics: to derive predictions from theory that can be tested using historical evidence.

To summarize progress since Lane and Hicks, I first consider the relationship between cliometrics and political economy, before providing a thematic survey of how cliometrics has advanced our understanding of state formation, warfare, public debt, the state, and religion. The final part of the chapter considers some case studies, including the Glorious Revolution, French Revolution, the political economy of empire, and the rise of democracy.

Early cliometric work focused on employing neoclassical economics to understanding historical questions. These included the productivity of the slave economy in the American South or contributions to US growth in the late nineteenth century. This work is well represented by the research which won Robert Fogel and Douglass North their Nobel prizes.[1]

At the same time as economic history was revolutionized by cliometrics, the field of political economy was reborn in the hands of public choice scholars in Chicago, Rochester, and Virginia. This work was, by and large, abstract, theoretical, and ahistorical, as exemplified by classics such as *The Calculus of Consent* (Buchanan and Tullock 1962), *The Theory of Political Coalitions* (Riker 1962), *The Logic of Collective Action* (Olson 1965), and *The Theory of Economic Regulation* (Stigler 1971). Empirical work inspired by Buchanan, Tullock, Riker, Olson, and Stigler

[1]The Nobel Prize was awarded to Fogel and North for pioneering cliometrics, specifically, "research that combines economic theory, quantitative methods, hypothesis testing, counterfactual alternatives and traditional techniques of economic history, to explain economic growth and decline" (*The Prize in Economics 1993 – Press Release* 1993).

followed, but, initially focused on elections, voting systems, and regulation, it was somewhat removed from the concerns of most economic historians.

Public choice contrasts with conventional public finance exemplified by Musgrave (1959). The pioneers of postwar public finance explicitly differentiated it from older work which "proceeded in a historical and institutional context" (Musgrave 1959, v). Public finance began with the assumption that the optimal policy can be implemented by a social planner. It is normative rather than positive, focusing on what policies should be followed to maximize a social welfare function.

Public finance does not explicitly model government agents as rational actors. Such institutional-free analysis had little role for history. Public choice, in contrast, imposes behavioral symmetry. Its predictions are made on the basis of the self-interested behavior of politicians, bureaucrats, and voters. In the 1980s and 1990s, this public choice approach integrated into mainstream political science and economics, acquiring the label, political economy.[2]

The Introduction of Political Economy into Cliometrics

The influence of North, and the new institutional economics, was crucial to the integration of political economy into economic history.[3] North helped introduce cliometric approaches to economic history, both in his own research and through his editorship of the *Journal of Economic History*. Then, beginning in the late 1960s and early 1970s, North began to focus on institutional economics. Initially, he worked under the assumption that institutions were efficient (North and Thomas 1973). Recognition that this was not necessarily the case turned North's attentions toward political economy (see North 1981).

Through his status as a leading cliometrician, North bridged the historical literature and public choice and political economy scholarship. His influence did not stem from the fact that his writings were particularly quantitative. They were not. North framed his arguments in a way that was attractive to more quantitatively orientated scholars and suggested ways that it *could* be tested. North's work inspired other scholars to build on his arguments or to challenge them.

Following North, interest in political economy topics among economic historians blossomed. Many strands of research employ the Northian emphasis on "rules of the game" and take into account the self-interested behavior of politicians, elites, voters, and bureaucrats. Engerman and Sokoloff (1994, 2000) traced the connection between factor endowments, institutions, and differential economic development in the Americas. Stephen Haber and coauthors applied institutional arguments to the complex political economy environment that is Latin America (Haber 1997; Haber

[2]The political economy label is particularly associated with the works of Persson and Tabellini (2000), Alberto Alesina, and Dani Rodrik.

[3]See discussion in Diebolt and Haupert (2018) of the impact of Fogel and North on economic history.

et al. 2003). North himself revised his institutional framework throughout his career; the final iteration was the natural states to open-access framework of North et al. (2009).

The first generation of research relied on a combination of theory and analytical narratives (Bates and Lien 1985; Levi 1988; North and Weingast 1989; Greif et al. 1994; Bates et al. 1998). Since the seminal attempt to measure the impact of colonial institutions (Acemoglu et al. 2001), this line of research has fused with empirical research that attempts to identify the impact of institutions.[4]

Two approaches to research design have become particularly influential, instrumental variables (IV) and regression discontinuity design (RDD), in part because economic historians can rarely run experiments and often lack the panel data required for a difference-in-differences approach (DID).

Melissa Dell's (2010) influential study of the long-run effects of Peruvian Mita, for example, popularized the use of RDD. Recent papers that employ IV approaches include Dippel (2014) who instruments for the forced integration of Native American communities by using data on historical mining rushes.

Not all topics in political economy are amenable to causal methods. In politics few things are truly exogenous. Instruments that satisfy the exclusion restriction are difficult to come by. Political borders often arise endogenously making RDD designs challenging. Therefore, it is important that economic historians studying political economy employ mixed methods. Formal models (as in the analytical narrative tradition), qualitative evidence, and descriptive econometrics all have a place in improving our understanding of the interaction of politics and economics. In political economy, perhaps more than in other fields, understanding the historical context and the data generating process remains of paramount importance.

In his presidential address to the Economic History Association, Hoffman (2015, 305) argued that "we still know too little about what determines the laws, regulations, and policies that states adopt or what goods and services they provided … Worse yet, we do not even understand how states arise in the first place or how they gain the ability to tax." I will take a different view and argue that progress has been made in understanding both the rise of states and the development of modern welfare states. To substantiate this, the next section provides a thematic approach to major cliometric contributions to political economy.

[4]The current state of the field is influenced by Acemoglu et al. (2001, 2005a) who took Northian arguments and tested them econometrically using innovative empirical methods. Following the success of Acemoglu et al. (2001), this approach has bloomed both within economic history and in the related fields of growth economics, development and political economy. Other important publications by Acemoglu and Robinson and their coauthors include Acemoglu et al. (2005b); Acemoglu (2006).

A Thematic Overview

Origins of the State

In contrast to the assumption of a benevolent social planner maximizing a social welfare function, political economy begins by considering states as they actually are – i.e., organizations that are made up of the same individuals who populate private firms. Rather than taking the economy as its starting point (and treating it as an autonomous segment of society that can be analyzed independently of politics), political economy concerns itself with the interaction between the state and the economy. A natural starting point is the origins of the state.

Speculation about the emergence of states goes back millennia. But while it is possible to begin with Aristotle or Hobbes, modern social scientific analysis can be traced to Oppenheim (1922) and Carneiro (1970). In economics, Brennan and Buchanan (1980) developed a simple model of a revenue-maximizing leviathan state. These ideas were combined in Mancur Olson's stationary bandit model (1993). For Olson, organized states arise from violence and disorder. The presence of competing warlords or roving bandits causes poverty, as each warlord robs without taking into account the damage he causes. However, if a single bandit establishes a monopoly of violence, such a stationary bandit will have a more encompassing interest in the society that he now "governs." Olson's story is both a parable and a model with testable predictions.

A key problem for such a stationary bandit is that mobile populations cannot be compelled to pay taxes; they can move to avoid fiscal extraction. Allen (1997) applied these ideas to ancient Egypt. Egypt developed a stable and long-lasting state because its geography made it easy for rulers to corral tax payers. Geographical circumscription was possible earlier than elsewhere because agriculture first developed along the fertile Nile valley, and this area of lush farmland was surrounded by inhospitable desert. Early statehood brought stability, but may have made the average Egyptian worse off, as the surplus created by agriculture could be extracted by the political elite.

For both Olson and Carneiro, agriculture was a precondition for statehood because farming produces a storable food surplus. Mayshar et al. (2017), however, note that a slow rise in productivity alone could not have generated a surplus, since in a Malthusian environment, population size adjusts to prevent the creation of such a surplus. Only where output was transparent and storable could states emerge. Mayshar et al. (2017) contrast the pattern of state development between Egypt and ancient Mesopotamia. Nile agriculture relies on flooding. How much the Nile flooded was public information, transparent to the Pharaoh. Hence the Egyptian state could extract all the surplus. Egyptian peasants were tenant-serfs; elites were dependent on the Pharaoh. In Mesopotamia, however, agricultural productivity, while transparent to local elites, was opaque to a centralized state. Hence, authority tended to be organized at the city level. Consistent with this prediction, states were less stable in Mesopotamia than in Egypt.

City-States and Republics

The earliest large states that we know about were autocracies. Wittfogel (1957) ascribed the prevalence and resilience of autocracy in much of Asia to the demands of irrigation-based agriculture. He argued that such systems of irrigation required strong, centralized state control. Where agriculture was heavily dependent on river water, rather than being rainfed, strong and centralized autocratic states were likely to arise, such as the Pharaohs of Egypt, the Assyrian, Babylonian, Persian Emperors of Iraq and Iran, and the Emperors of China. Wittfogel's original hypothesis has been widely criticized. Recent empirical work suggests some support for it. Bentzen et al. (2017) find that irrigation-based agriculture is more likely to be associated with autocratic rule. Elis et al. (2018) argue that there are optimal climatic conditions for the emergence of participatory democracy.

An alternative political organization to autocracies was the city-state. The best known city-states were the *poleis* of classical Greece. In comparison to the ancient empires, these were egalitarian. They permitted self-government among strata of the population, either wealthy landowners in oligarchies or free males in democracies.

To understand how democracies emerged in ancient Greece, Fleck and Hanssen (2006) develop a model based on a time inconsistency problem. According to their account, democracy came about when and where there were productive opportunities for investment for the demos.[5] Democratic rights ensured that these investments would not be expropriated by a ruler or an elite. Another hypothesis for why democracy emerged in ancient Greece is that ecological conditions produced a relatively egalitarian distribution of income and a large population of smallholder farmers. The development of the hoplite military formation also empowered these smallholders and necessitated their political representation. In Athens, the rise of a navy was crucial in cementing the power of the demos (Kyriazis and Zouboulakis 2004).

Classical Greek *polis* witnessed an economic, cultural, and intellectual flourishing (Ober 2015). But they were limited in size and scope – unable to integrate outsiders – and in a state of constant hostility with one another. Inter-polis conflict was responsible for the catastrophic Peloponnesian War. The rise of Rome led to the establishment of an imperial monarchy, and monarchy was the main way subsequent European states organized themselves into the modern period.

The medieval period saw the revival of the city-state, especially in Italy and the Low Countries (Pirenne 1925). These city-states evolved institutions that sought to limit social conflict, as analyzed in the Genoese case by Greif (1998, 2006). But like their ancient counterparts, most medieval city-states were limited in size.

Medieval city-states are often viewed as economic success stories (Cipolla 1976). Their institutions were indeed conducive to initial economic growth, but as Stasavage (2014) documents, the constitutions of medieval city-states became

[5]Fleck and Hanssen (2013) discuss how the institution of tyranny – stable autocratic rule – helped to pave the way for democratization.

more oligarchic over time. As this occurred, the growth advantage enjoyed by city-states declined. Using city population as a proxy for economic development, Stasavage finds that overall autonomous city-states did not grow faster than other cities on average. City-states that had been independent for less than 200 years grew faster; then, after 200 years of independence, they grew more slowly than other cities.

Stasavage interprets this in terms of a model of oligarchies proposed by Acemoglu (2008). In this model, as oligarchic societies represent the interests of producers, they tend to protect property rights and to impose low taxes. But they also erect barriers to entry which can result in sluggish growth in the long run. In contrast, democracies impose higher taxes, redistribute more, but also tend to allow free entry. Oligarchic societies, therefore, tend to experience initially rapid growth followed by stagnation.

An example of this process is Venice, where the prosperity associated with trade made merchants politically powerful. They ended the hereditary rule of the Doge in 1032, which became an elected office, and, in 1172, established a Grand Council to further constrain executive authority. Tracing family names of the Council members, Puga and Trefler (2014) show that there was considerable mobility into and out of the Council, suggesting that membership of the political elite was fluid. However, this mobility threatened the profits of established merchants. This led to a period known as the Serrata ("closure") in which the established merchants used their power in the Council to pass laws to close entrance. A law in 1297 restricted new entrants in the Council, and one in 1323 restricted entry into long-distance trade. This speaks to the general trend: city-states were associated with impressive but transitory prosperity.

Medieval States and Feudal Institutions

An alternative to large empires or city-states were territorial states. Huning and Wahl (2016) argue that more observable agricultural output is associated with the emergence of larger territorial states. Testing this model using the Holy Roman Empire, they find higher observability leads to larger and longer-lasting states.

The larger territorial states that emerged in medieval Europe were feudal. Feudalism has a bad reputation among economic historians. It represented the dispersal of coercive power to such an extent that some historians deny that feudal polities were "states" (Strayer 1965). Blaydes and Chaney (2013) find that feudalism led to greater political stability in Western Europe. Under feudalism, rulers relied on powerful nobles for military support, and in return for their support, they granted these nobles privileges and political power. These bargains helped to solidify the nascent polities of Western Europe. After 900 CE, Blaydes and Chaney (2013) find a divergence in the duration of Christian and Muslim rulers, in favor of the former (Fig. 1).

Greater economic opportunities and constraints on the executive accompanied the increased stability of European monarchies. This more stable environment promoted urbanization and a Commercial Revolution. Feudalism ensured that monarchical

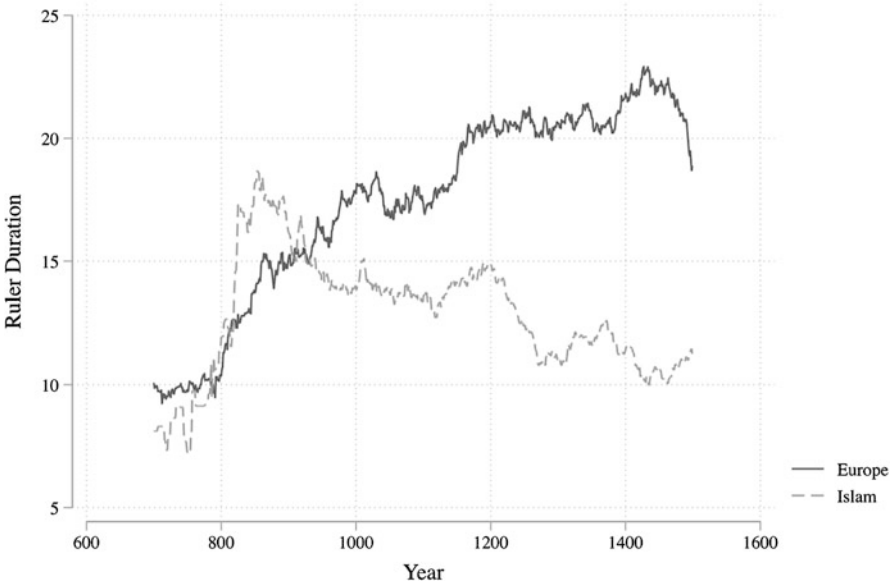


Fig. 1 100-year moving average for ruler duration in Europe and the Islamic world. (Data from Blaydes and Chaney 2013)

abuses were self-limiting. Political elites had the military power to stand up to overbearing monarchs (Blaydes and Chaney 2013). Perhaps the most well-known instance of this was the ability of the English barons to force King John to sign the Magna Carta (Koyama 2016).

Salter and Young (2018) argue that medieval polities were successful because they aligned the incentives of landowning elites. Political rights in medieval societies were bundled with property rights. According to this view, medieval lords were incentivized to pursue policies that were beneficial to development because they had economic property rights in their realms. And unlike satraps or appointed governors in centralized empires, they had the right to bargain with their sovereigns and to hold them to account.

For Blaydes and Chaney (2013), the rise of feudalism has implications for the divergence between Europe and the Middle East. In contrast to Western Europe, Islamic states came to rely on slave soldiers. Landlords were alienated from political power. Levels of political stability in these two regions of the world thus diverged centuries prior to the divergence in per capita income (Blaydes 2017).

Acharya and Lee (2018) point to the role of economic development in the formation of the European state system. They argue that when trade and commerce are underdeveloped, and hence the value of governance is low, there is no incentive for territorial states to emerge. In this world, there will be pockets of state authority but no territoriality. Territoriality emergences when there are overlapping markets for protection.

Labor Coercion

Feudal societies relied on labor coercion. North and Thomas (1971) hypothesized that serfdom arose as a quasi-voluntary institutional response to the frequent conflict and invasions that characterized the early Middle Ages. Domar (1970) argued that serfdom emerged where and when labor was scarce. This prediction can help to explain the emergence of serfdom in early medieval Europe and in Eastern Europe after c. 1600. Brenner (1976) noted that a purely demographic model was insufficient to explain the *decline* of serfdom in Western Europe following the Black Death; this phenomenon required studying political power and class relations. Labor scarcity increased the bargaining power of laborers contributing to a crisis of surplus extraction, in Marxian terminology.

These explanations can be reconciled. Wolitzky and Acemoglu (2011) build a principal-agent model to study the relationship between labor scarcity, outside options, and labor coercion. In this framework, coercion and effort are complements. Hence, when labor is scarce, there is a stronger incentive to employ coercion. However, labor scarcity also improves workers' outside options, which reduces the incentive to use coercion.

What were the economic consequences of labor coercion? Studying early modern Bohemia, Klein and Ogilvie (2016) show how coercive labor market institutions shaped economic incentives. In particular, under serfdom, landlords suppressed activities from which they could not extract rent. Close to urban centers, landlords were more likely to enforce coercive restrictions on labor. These findings suggest that the power of landlords to coerce labor in Central and Eastern Europe helps to explain their relative economic underdevelopment relative to Western Europe in the centuries leading up to the Industrial Revolution.

Serfdom persisted in Eastern Europe until the nineteenth century. Even in England, master and servant laws continued to be used to keep wages low in Industrial Revolution England (Naidu and Yuchtman 2013). Ashraf et al. (2017) explain the decline of serfdom in the early nineteenth-century Prussia in terms of the economic incentives facing landowners. They hypothesize that as skilled labor became more important in the production process, elites had an incentive to emancipate serfs in order to encourage them to invest in broad-based human capital. They find empirical support for this hypothesis using data on county-level emancipations from the nineteenth-century Prussia.

Conflict and Consensus

Politics involves competition over resources. Economics views efficiency as its central concern; the allocation of resources is determined by the price mechanism according to their highest-valued use. Political economy is concerned about the distribution of resources, in the absence of prices and in the presence of political power. Ogilvie (2007) discusses how paying attention to political power undermines the conclusion that whatever institutions exist are efficient. Institutions result from

sociopolitical conflicts over resources, and the absence of a political Coase Theorem means that conflicts often result in inefficiency (Acemoglu 2003). Existing institutions need not be efficient.[6]

Institutions shaped by those with political power may be inimical to economic growth. The importance of political power can explain the attention that economic historians have devoted to understanding institutions that constrain the powers of rulers and represent the interests of non-elites. In European history, the most important such institutions are the parliaments, which emerged in the twelfth and thirteenth centuries.

De Long and Shleifer (1993) provide evidence that representative institutions were associated with economic growth in medieval Europe. Van Zanden et al. (2012) measure the frequency that parliaments or estates met across Europe. They document the rise and spread of parliaments and then their decline after 1500, outside England and the Netherlands. The underlying argument is that parliaments enabled merchants and the owners of capital to limit predation by rulers or landowning elites.

While parliaments contained rulers, recent research stresses that parliaments emerged out of councils that rulers called (Congleton 2010). Rulers recognized that parliaments could increase their ability to raise taxes and legitimize their rule. Boucoyannis (2015) discusses how the power of English kings, like Edward I, to compel attendance in Parliament ensured that it was a representative body. This perspective suggests that parliaments enabled, as well as constrained, premodern states.

A complementary hypothesis is offered by Leon (2018), who argues that the gradual expansion of the elite in medieval England helped lay the conditions for the peaceful transition to democracy. He develops a model in which the king expands the size of the elite to gain support against barons. Once this elite reaches a certain size, it becomes cheaper for the king to compensate them with rights rather than pay them directly, and this can make the rise of more representative institutions self-reinforcing.

The view that representative institutions were necessarily good for economic development has been challenged. Taking the examples of Poland and Wüttemberg, Ogilvie and Carus (2014) argue that parliaments which only represented landed interests tended to grant legal monopolies to elites at the expense of broad-based economic growth. The Dutch Republic too, they argue, stagnated after 1670 due to entrenched power of established business interests.

Under what conditions, then, are parliaments likely to produce conditions favorable to economic development? Acemoglu et al. (2005b) study how the rise of Atlantic trade interacted with institutional developments in northwestern Europe. According to this argument, the opportunities for trade and commerce that opened up

[6]For a contrary perspective on institutions, see Doug Allen (2011) or Peter Leeson (2017). My understanding of their argument is that existing power relationships should be viewed as constraints. Hence existing institutions can be viewed as efficient relative to the appropriately defined set of constraints.

after 1500 strengthened the nascent merchant class in societies like England and the Dutch Republic where the state was not too strong, but reinforced the positions of absolutist monarchs in Spain and Portugal.[7]

Another perspective is provided by Cox (2017b) who argues that independent city-states and national parliaments together provided the economic liberty that unleashed faster urban growth in the premodern period. Before 1100 inter-city urban growth rates were uncorrelated, but after 1100 urban growth rates in Western Europe – but not elsewhere – began to move together. In other words, economic liberty promoted growth clusters. Cox argues that political fragmentation could lead to competitive pressures, which could force rulers to charge lower tax rates on commerce and trade. The effects of political fragmentation were largest, he suggests, in the presence of parliaments in which merchants were represented. But what factors led to the representation of merchants in parliaments? And what ensured that merchants did not use their power to erect barriers to entry?

Warfare

Military competition played an important role in ensuring institutional innovation and openness. States like Poland and Wüttemberg that erected barriers to economic development and did not invest in state-building efforts were eventually subordinated by their rivals.

The role of warfare in state formation was stressed by Hintze (1906) and Tilly (1975, 1990). Economic historians have shown how frequent warfare led to investments in state capacity and to improvements in military technology. Hoffman (2011, 2015) documents how after 1500 technological innovations brought down the unit costs of guns and cannons in Western Europe. The military sector was one of the most innovative in the European economy in the period before the industrial revolution. Handguns, in particular, became both more effective and cheaper, lowering the costs of employing violence.

A theoretical perspective is required to understand why warfare promoted state development in some instances and state collapse in others. Gennaioli and Voth (2015) build a model that integrates insights from the Military Revolution literature (Parker 1976, 1988). In their model, the incentive to invest in fiscal capacity depends on the chances a state has of defeating its rivals. Increased military competition provides incentives for some (initially more homogenous) states to invest in standardizing their fiscal systems, in order to invest in more capital-intensive means of waging war. But other states (which may be initially more heterogeneous and hence have higher costs of centralizing) may not find this worthwhile and hence may end

[7]Economic historians have rightly criticized the coding and the depiction of Spain and France as governed by overly powerful absolutist monarchs as out of date. But this research showed what was possible with historical data.

up losing out as a result of this intensified military competition. This model is consistent with the observation that the pressures of military competition led the most advanced and centralized European states to invest even more in state capacity, while it led to the destruction of others.

War and state-building went together in Europe's most advanced states. What role did intensified military competition play in explaining patterns of economic development in Europe? Dincecco and Prado (2012) estimate the impact of fiscal capacity on modern development. Instrumenting for fiscal capacity with causalities in premodern wars, they argue that this relationship reflects the effect of premodern fiscal innovations on current fiscal institutions, and these current institutions have a positive impact on economic growth.

Another possible channel was via urbanization. Dincecco and Onorato (2016) argue that frequent warfare in medieval and early modern Europe led to urbanization and economic development. Cities acted as safe harbors from conflict. Conflict exposure was associated with between a 6% and a 11% increase in urban population over the course of a century. Dincecco and Onorato (2017) argue that this effect has persisted and explains regional level variation in economic development in Europe today.

Other authors push back on this bellicose hypothesis, however. For one, frequent warfare did not promote urbanization, economic development, or the rise of more inclusive institutions in other parts of the world (Centeno 1997). When population density is low, warfare may promote slave raiding rather than state-building (Herbst 2000). Warfare alone seems insufficient as an explanation.

Patterns of Political Fragmentation and Political Centralization

One reason why military intensification led to investment in fiscal capacity in Europe was the competitive European state system (Fig. 2). What then accounts for this competitive state system?[8]

Geography was an important factor. Diamond (1997) provided geographical hypotheses for why Europe tended to be politically fragmented, while China tended toward unification. Europe's mountain ranges and fractured coastline promoted fragmentation. Hoffman (2015) criticized this hypothesis on the grounds that China was more, not less, mountainous than Western Europe. Recent research does suggest that European geography helps to explain its political fragmentation (Fernandez-Villaverde et al. 2019).

Certainly, geographical factors intersect with political economic considerations. Ko et al. (2018) build on the research of historians of Central Asia and theoretical models of state size such as (Alesina and Spolaore 1997). They argue

[8]A literature extending back to Montesquieu and Hume argues that Europe's political fragmentation was key to its eventual rise and to modern economic growth (Baechler 1975; Jones 1981; Hall 1985; Rosenberg and Birdzell 1986).

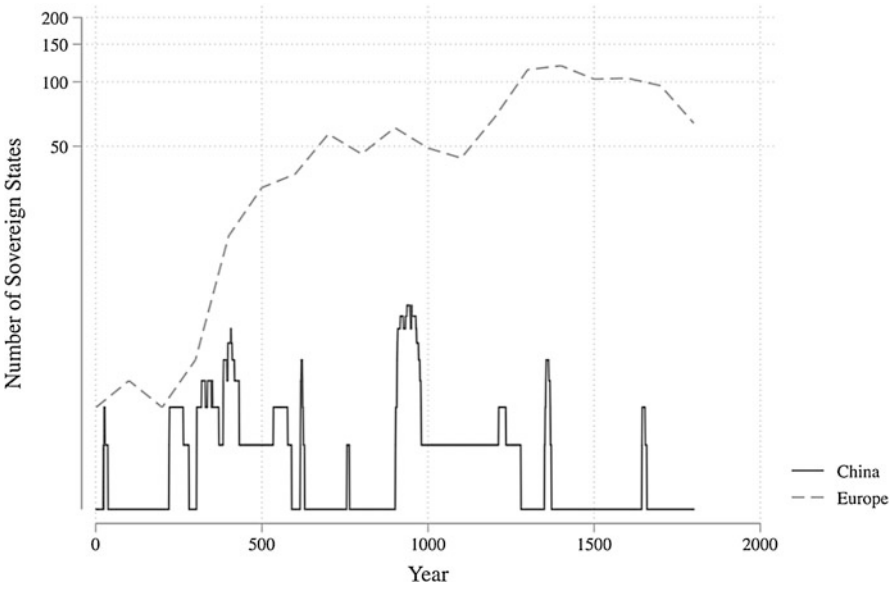


Fig. 2 Number of states in China and Europe, AD 0–1800. (Adapted from Ko et al. 2018)

that in comparison to China, Europe faced invasion threats from multiple directions, impeding the growth of a single European-wide hegemon. In contrast, China faced a single threat from the nomadic steppe. As a result, throughout Chinese history, a strong state tended to emerge in Northern China, and this state tended to be strong enough to establish an empire over the rest of the region. Ko et al. (2018) test their model using a combination of historical evidence and time-series analysis of the impact of nomadic invasions on political centralization in China.

What about the consequences of political fragmentation? Karayalcin (2008) introduced a model that predicts that fragmentation should have led to lower taxes in Europe. This is not what we observe, however. Taxes were higher in late medieval and early modern Europe than in either the Ottoman Empire or China. Ko et al. (2018) show that it might be in the interest of rulers of large empires to levy comparatively low taxes to reduce the probability of rebellion, but that this incentive is absent or much weaker in a competitive state system.

One insight from research is that political fragmentation was not an unambiguously positive factor in Europe's economic development. It imposed static costs and resulted in the multiplication of trade barriers and in an over investment in military power that was itself a cause of internecine warfare. In this respect, more centralized empires such as Qing China provided a potentially more fruitful environment for Smithian economic growth (as argued by Bin Wong (1997) and Rosenthal and Wong (2011)). There were, however, important *dynamic* benefits associated with political fragmentation and competition.

State Finances

One example of the institutional innovations associated with warfare is in the area of state finances. The Italian city-states were almost continuously at war. Venice and Florence pioneered institutional innovations like public debt, which enabled them to compete with much larger states and raise ever greater revenues to pay for their mercenary armies. The annual revenue of Florence in the fourteenth century varied between 250,000 and 350,000 florins, but historians estimate the direct cost of the 3-year war between Florence and the Papacy in 1375–1378 exceeded 2.5 million florins (Caferro 2008). To meet this shortfall, Italian city-states developed impersonal systems of public debt and permanent systems of taxation (Epstein 2000).

Epstein (2000) documents that throughout the late medieval and early modern periods, city-states and republics paid less interest on their debts than did territorial monarchies. Stasavage (2011, 2016) shows that representative institutions enabled city-states to borrow earlier and at lower cost. In city-states, the holders of capital were represented in government, ensuring that the promise to repay was credible. Figure 3, adapted from Stasavage (2011), is consistent with this argument (see also Chilosi 2014). Only in the seventeenth century did larger states, first the Dutch Republic and then England, develop institutions, such as the Bank of Amsterdam and the Bank of England, which replicated the success of smaller city-states (see section "The Glorious Revolution").

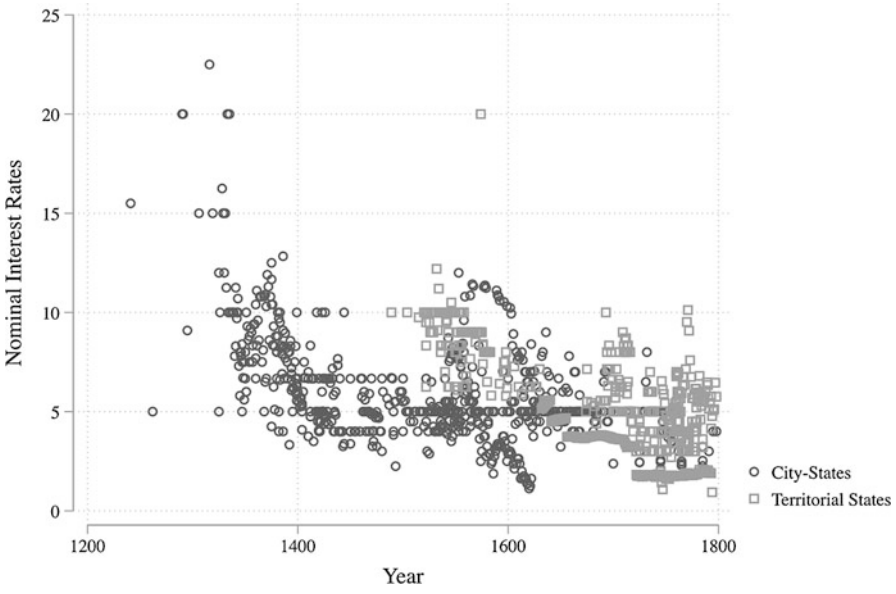


Fig. 3 Interest rates in city-states and territorial states. (Data from Stasavage 2016)

Religion

Despite the important role religion occupied in the studies of Weber (1922, 1930) and Tawney (1926), religion was neglected by cliometric historians until recently. A recent literature specifically focuses on the intersection of religion and political economy.

Kuran (2010) drew attention to the important role Islam played in shaping economic opportunities in the Middle East. He shows that Islam infused both the law and politics of Middle Eastern countries with important consequences for the formation of long-lasting business organizations and corporate institutions independent of the state, such as universities. Recent studies of the Reformation have moved away from studying the direct effects of religious change on economic growth toward discussing the institutional consequences of the Reformation (see Becker et al. 2016; Cantoni et al. 2018).

One area where the study of religion overlaps most significantly with political economy is in the relationship between the Church and the state. Throughout history, religion has played a key role in legitimizing political authority (Coşgel et al. 2012; Greif and Rubin 2015; Rubin 2017). Coşgel and Miceli (2009), for instance, develop a model of how a religious authority can lower the costs of taxation for the political authorities, the logic of which can be applied to numerous settings.

Chaney (2013) uses variation in the flooding of the River Nile in medieval Egypt to explore the relationship between religious and secular authorities. He finds that Nile shocks increased the political power of religious leaders. Nile shocks resulted in great potential for unrest, and this improved the bargaining power of religious authorities with respect to secular leaders. To neglect religion is to fail to understand many aspects of premodern political economy. In a recent contribution, Coşgel et al. (2018) compile a dataset of state religions going back to the year 1000 CE, which will prove usual for scholars working on these questions.

Rubin (2017) builds on these arguments to develop a novel account of the divergence that opened up between northwestern Europe and the Middle East after 1500. The greater power of religion to legitimize political authorities in the Middle East constrained the development of institutions and the adoption of new technologies such as the printing press. The comparative weakness of the religious authorities in northwestern Europe, however, set in motion institutional developments that helped limit government and provide the institutional preconditions for sustained growth.

Finally, Johnson and Koyama (2019) examine the rise of religious freedom from the medieval ages to the twentieth century. Given the centrality of religion to providing political legitimacy and enforcing institutional arrangements, they argue that the importance of religion to premodern polities meant that full religious freedom was inconceivable. Drawing attention to the various institutional arrangements premodern states used to deal with religious diversity, they point out their economic costs and the fragility of the resulting "conditional toleration" that resulted. In sum, religion cannot be ignored in studying the political economy of premodern states.

State Capacity

State capacity refers to the ability of the states to enforce property rights, implement policies, and provide public goods (Besley and Persson 2011; Johnson and Koyama 2017). It can be broken down into administrative capacity, legal capacity, and fiscal capacity. One challenge has been obtaining accurate measures of these concepts. Fiscal data now exists for a number of European countries from around 1650 onward (Dincecco 2009; Karaman and Pamuk 2013). But we lack comparable measures of other aspects of state behavior.

Preindustrial states were usually small and fiscally and legally fragmented. Fiscally fragmented regimes suffered from local tax free-riding. Fiscal centralization thus enabled states to increase revenues and often complemented the expansion of markets and the division of labor. This made it easier for responsible governments to follow sound fiscal policies and thus lowered credit risk. However, there was always the chance that rulers would waste the new funds on reckless wars. Thus, it was only the conjunction of limitations on the discretionary authority of government with centralized fiscal systems that enabled states to borrow at low cost.

Fiscal centralization was accompanied by legal centralization. Old Regime France was legally fragmented: customary law and the interpretation of Roman and canon law varied from place to place. During the seventeenth to eighteenth centuries, there was an effort to achieve some legal centralization. Johnson and Koyama (2014b) document how legal centralization and investment in state capacity was accompanied by a decline in witchcraft trials, in which obtaining convictions required a departure of standard legal procedures. Crettez et al. (2019) model the transition from legal decentralization to legal centralization in terms of a trade-off between allowing each region to choose its own idiosyncratic legal system and imposing legal centralization. The former allows a region to adopt rules closer to the preferences of its inhabitants; the latter eliminates legal disparities across regions. They apply this model to explain why the French Revolution was accompanied by a sharp move toward legal centralization.

Centralized fiscal systems were established comparably late in European history. Most European states achieved fiscal centralization and limited governments only after 1800. Dincecco (2009) dates the establishment of a centralized tax system in France and the Netherlands to the French Revolution. Centralized and limited regimes were only established in the Netherlands after 1848, in France after 1870, and in Spain after 1876. Only once European states established centralized tax systems, and limited government, did yields on public debt fall.

This account is consistent with the data provided by Karaman and Pamuk (2013), which depict marked increases in fiscal capacity in England, France, the Dutch Republic, Prussia, and Habsburg Austria after 1700, but no corresponding increase in fiscal capacity in the Ottoman Empire, Poland-Lithuania, Russia, or Sweden. Much of this increase in state capacity was associated with an intensification of interstate conflict. Karaman and Pamuk (2013) find that in states with representative institutions, such as parliaments, this increase in fiscal capacity was most pronounced. In contrast, in absolutist states, formal tax revenues were low. This was true of both the Ottoman Empire and Qing China. Due to a combination of economic growth and investment in

state capacity and military technology, by the mid-nineteenth century a large gap in state power had opened up between Western European states and the states of Asia.

Why did Asian states fail to keep up with Western Europe? Fiscal centralization was a gradual process in Europe. State tax collection emerged out of private tax farming. Tax farmers organized in order to discipline and bind monarchs who had an incentive to expropriate tax collects when in need of revenue (Johnson 2006; Johnson and Koyama 2014a). Balla and Johnson (2009) discuss how the inability of tax farmers to organize in the Ottoman Empire meant that they were unable to constrain the Ottoman state. To explain why tax rates were low in Qing China, Sng (2014) argues that a severe principal-agent problem meant tax collectors could embezzle revenues and extort taxpayers. In response, the ruler had to keep taxes low to minimize the threat of revolt. Ma and Rubin (2019) build on this argument by highlighting the commitment problem autocrats face vis-a-vis their own tax collectors. In times of crisis or war, autocratic rules cannot credibly commit not to expropriate tax collectors and administrators. One way to overcome this problem is to pay tax collectors very low formal salaries but also to turn a blind eye to the private exactions of tax collectors by not investing in monitoring technology. This results in an equilibrium where formal tax collected by the center is low and informal corruption is high. But it precludes investment in administrative capacity.

What accounts for the relative failure of Qing China to deal with new political threats after 1800? As Ma and Rubin (2019) discuss, whether states invested in fiscal capacity was a political decision. Koyama et al. (2017) study patterns of state development in East Asia after 1850. They examine why Japan embarked on political centralization and reforms following the intrusion of geopolitical threats from the Western powers after 1850, but why in China political authority was decentralized and modernization efforts failed. They point to how geographical size constrained the options of political actors. In China, the conjunction of threats from both the north and the south necessitated decentralization, and, as a consequence, the Chinese state lacked the capacity to ensure that modernization could be forced through successfully.

In summary, recent work in economic history has greatly improved our understanding of the political economy of preindustrial states. As a result of this body of research, which draws on the work of economists, political scientists, and historians, we have better explanations of the determinants and consequences of regime type. We also have new theories about the causes and consequences of political centralization and fragmentation. Our understanding of the relationship between religious and secular authority in the premodern period has also been greatly enhanced. The next section considers several important areas of research in political economy.

Case Studies

The Glorious Revolution

An important case study is England in the centuries leading up to the Industrial Revolution. The Glorious Revolution represents a widely studied critical juncture in English institutional history (North and Weingast 1989; Pincus 2009; Pincus and

Robinson 2014; Acemoglu and Robinson 2012). It involved England in a global war against France and led to the foundation of the Bank of England in 1694. By making England a constitutional monarchy, it laid the foundations of the party system and cabinet government (Stasavage 2002, 2003; Cox 2016).

The roots of this institutional transformation go further back in English history. One critical juncture was the English Civil War (1642–1651). Building on a hypothesis advanced by Brenner (1993) on the role merchants played in the first English revolution, Jha (2015) uses novel data on individual MPs to explore how emerging economic opportunities overseas helped the formation of a coalition in favor of constraining the crown. He finds that MPs with financial interests in overseas trade were more likely to support Parliament in the conflict with the crown. Shares aligned the interests of non-merchants, who otherwise would have lacked exposure to lucrative trading opportunities and hence expropriation risks overseas, with merchants. Ownership of shares in overseas trade shifted the views of non-merchants, helping to consolidate support for reformers in Parliament.

Fiscal centralization also began before the Glorious Revolution. The previous "bureaucratic patchwork of authorities subject to uniform surveillance or direction" (Brewer 1988, 91) and private tax farming began to change during the Civil War. Parliament raised new taxes, including the excise tax, to fund a new professional army and navy (O'Brien 2011). Taxes and expenditure fell in the 1660s, but lessons about how to modernize England's fiscal system were learned. Tax farming of the customs and then the excise was slowly dismantled, and the Treasury was given oversight over revenues (Johnson and Koyama 2014a).

The increase in taxes after 1688 was important, however. There was greater reliance on indirect taxation – in particular the excise – collected by a professional bureaucracy. The customs, excise, and land tax accounted for 90% of revenue. From 1720 onward, these revenues enabled the British state to secure its debt at low rates of interest – a key development in the formation of a modern state.

Other aspects of state administration remained dominated by patronage and cronyism. Cox (2017a) highlights the importance of the civil list in the political economy of the Hanoverian state. After 1689, while Parliament controlled taxation and the military budget, civil positions remained controlled by the crown. This resulted in a commitment problem. Because MPs did not have control over the civil administration, they were reluctant to fund it, impeding the development of a civilian bureaucracy. Reform only came in 1831 with the Civil List Act, which established full parliamentary control over the civilian budget and made it feasible for the state to invest in a modern bureaucracy.

North and Weingast (1989) argued that the credibility of the English monarchy to repay its debts after 1688 translated into more secure property rights in general. While this claim has not survived subsequent scrutiny, in other respects there was an improvement in institutional quality after 1688. It became easier to reorder property rights in order to exploit new investment opportunities. Parliament became a forum where land could be reallocated toward more productive uses (Bogart and Richardson 2009, 2011). As a consequence, investment in road and river transport dramatically improved, with important consequences for subsequent economic growth (Bogart 2011). Prior to the

Glorious Revolution, estate bills often failed due to political conflict. As a consequence, investment in road and river transport dramatically improved, with important consequences for subsequent economic growth (Bogart 2011). Dimitruk (2018) finds that political and constitutional changes after 1688 resolved or changed the nature of many of the conflicts in government. After the king became dependent on Parliament, sudden closures of Parliament were less likely.

The English/British case study is illuminating because it points to the importance of doing serious historical work rather than relying on proxies for institutional quality, as in common in many areas of economics. For instance, according to standard measures, institutional quality was constant in England between 1689 and 1830. Maden and Murtin (2017), for instance, claim that institutions did not drive British growth because they were unchanged in this period. But this conclusion is unwarranted; this is precisely the period that economic historians have found marked improvement in actual institutional performance.

Interest groups continued to push for rent-seeking legislation and to block potentially productive improvements. North et al. (2009) use the term "open access" to describe modern, liberal, market economies in which entry is not controlled by political elites. One sign that eighteenth-century England was not yet an open-access economy is that political parties made a difference for investments in infrastructure. Bogart (2018) finds that periods of Whig dominance were associated with investments in productive infrastructure, particularly river projects.

Parliament did become less rent-seeking over time. Early eighteenth-century parliaments were notoriously venal, passing numerous acts that benefitted specific interests – most notably the Calico Act of 1721 – at the expense of the larger public. Members of Parliament were expected to look out for their own material interests and pursued what to modern eyes looks like corruption and venality (Root 1991). Over the course of the eighteenth century, this changed. Mokyr and Nye (2007, 58) note that "[p]urely redistributional actions, however, began losing their appeal. Many special interest groups' legislated privileges, monopolies, exclusions, limitations on labor mobility, occupational choice, and technological innovation found themselves on the defensive as the 18th century wore on."

Consider the position taken by Edmund Burke as an MP for Bristol. Bristol was a port that benefited from mercantilist policies, notably protective tariffs on Irish goods. In 1778, however, Burke argued in favor of free trade with Ireland. This motion was successful, but Burke was attacked by his constituents. He was unwilling to be bought off by his prosperous constituents and instead argued for the policies that he believed would benefit the country as a whole (see Prior 1878).

This process of institutional change can be understood through Olson's distinction between decentralized and centralized rent-seeking. Britain moved from decentralized to centralized rent-seeking (Ekelund and Tollison 1981). This was associated with a lot of competition over rents in Parliament, but the elimination or streamlining of local rent-seeking. By the nineteenth century, it became evident that the growth of the British economy as a whole was the surest way for elites to gain materially and to retain their status. The British landowning elite were not replaced; they came to benefit from the growth of a commercialized economy.

An important historical topic that has not yet received attention from cliometricians is the retrenchment of the fiscal-military state undertaken after 1815. While authors such as O'Brien (2011), Vries (2015), and Ashworth (2017) make strong claims about the economic benefits of the mercantilist policies pursued by the British state, the period in which economic growth was strongest and which saw substantive gains in real wages for ordinary works was the period during which the fiscal-military state was being dismantled by Peel and Gladstone. Gladstone's success in reducing the fiscal burden of the state was remarkable. But this period of British history has yet to receive as much attention from economic historians as that before 1815.

The Political Economy of Empire

Concern with the economic costs of empire extends back to Adam Smith, a sharp critic of British imperialism. An extensive scholarship studies the costs and benefits of colonial empire. Historians influenced by either Marxism or world-systems theory place tremendous explanatory weight on imperialism to explain sustained economic growth in Europe and North America.

This work was largely based on impressionist evidence, however, and the Hobson-Leninist theory of imperialism. No doubt Atlantic and colonial trade stimulated urbanization and economic growth, particularly in cities like Bristol, Liverpool, Bordeaux, and Glasgow. But these accounts are apt to neglect the importance of domestic over international commerce. As a proportion of GDP, trade with the imperial periphery was small (O'Brien 1982). Imperialism is ancient; sustained economic growth is not. Moreover, the development of overseas empires, first in Asia and then Brazil, did not prompt industrialization in Portugal. Spain was "cursed" by the abundance of silver, obtained during its imperial heyday (Drelichman 2005).[9]

Nevertheless, in the case of Britain, O'Brien and Escosura (1998) argue that the "mercantile and mercantilist matrix dominated by colonialization and commerce with continents beyond Europe" was important though not necessarily crucial for Britain's long-run economic success. As the argument cannot rely on the share of trade in GDP, which O'Brien himself established was small, it hinges on the importance of overseas markets as a source of demand for British products and on increasing returns associated with the Atlantic trading nexus. Allen (2009), for instance, pointed out a role for colonization in boosting British wages, especially in trading entrepôts like London. These high real wages then provided the incentive for labor-saving technological change. Findlay and O'Rourke (2007) also make the

[9]Note that the most recent research on the Spanish economy, for instance, points to domestic factors such as the absence of integrated markets or a standardized fiscal system (Grafe 2012; Álvarez Nogal and de la Escosura 2013). Many of the bolder claims made on the behalf of the importance of empire to the origins of growth in Europe are ably dismissed by McCloskey (2010).

case that colonial trade played a critical role in the British Industrial Revolution. But, without a model or explicit theory, it is all but impossible to quantify these external benefits or linkages; hence this is a topic that needs further research.

The impact of colonial institutions on colonized counties, in contrast, was of critical importance. This was because it not only changed the economy of colonized countries, but their political economy as well. Building on Sokoloff and Engerman (2000) and Acemoglu et al. (2001), a large literature finds that extractive colonial institutions can explain the persistence of poverty in developing countries today. Acemoglu et al. (2001, 2002) argued that the colonial period saw a reversal in economic fortunes. The most developed parts of the Americas and Africa attracted the attentions of European colonial powers and, as a result, were among the least developed parts of the world by 1900.

In a seminal contribution, Nunn (2008) showed that the transatlantic slave trade left a lasting negative impact on economic development in Africa. The slave trade also lowered population density, impeded the formation of states, and left a legacy of mistrust (Nunn and Wantchekon 2011). Other research has found long-lasting negative effects of the colonial rubber industry in the Congo (Lowes and Montero 2018). There was some investment in colonial empires, particularly in the form of railroads (see Jedwab and Moradi 2016). But in general the burden of empire was borne by the colonized (see Huillery 2014).

The impact of colonial rule on African society should not lead one to believe that Africa was a tabula rasa prior to the arrival of the Europeans. Gennaioli and Rainer (2007) and Michalopoulos and Papaioannou (2013, 2016) point to the importance of pre-colonial African states. Michalopoulos and Papaioannou (2013) find that the complexity of pre-colonial institutions is associated with economic development as measured by light density. This confirms the hypothesis of political scientists that a history of statehood is important in explaining outcomes today.[10]

The political economy of colonial India has also been subjected to cliometric scrutiny. Banerjee and Iyer (2005) find that under British ruler landlords given proprietary rights in land were unequal in the post-colonial period, spent less on development and public goods, and have experienced less of a decline in poverty. Iyer (2010) exploits a British colonial policy, whereby the territories of rulers of native states who died without a natural heir were absorbed into the British Empire, in order to generate exogenous variation in whether an Indian state received direct or indirect colonial rule. She finds that direct British rule was associated with less public goods provision in the post-colonial period: low availability of schools, health centers, and roads. This translates into higher levels of poverty and infant mortality. At the same time, British railroad investments played a crucial role in decreasing trade costs and raising incomes in colonial India (Donaldson 2018) and reduced the incidence of famine (Burgess and Donaldson 2017).

[10]This leads one to ask what determines patterns of pre-colonial state development in Africa. According to Fenske (2014), pre-colonial African states emerged in more ecologically diverse environments where the returns to trade were greater.

In India, the impact of colonial rule interacted with preexisting state structures and distributions of political and social power. Chaudhary (2009) finds that primary education in British India was particularly low in areas that were religiously diverse and which had greater caste differences. Chaudhary and Rubin (2016) find that literacy was lower in Muslim states than Hindu states. They interpret this finding in the context of a model in which rulers have a greater incentive to provide public goods where a higher proportion of the population is coreligionists.

Besides this discussion of imperialism, the state, and economic growth, other aspects of the political economy of European and American empires have also come under scrutiny by cliometricians. Mitchener and Weidenmier (2005) provide evidence that the implicit American empire in Latin America, associated with Theodore Roosevelt's Corollary to the Monroe Doctrine in 1904, brought economic benefits, as reflected in the price of Latin American sovereign bonds. Roosevelt made the US's threat to intervene in South America credible. US hegemony, they suggest, provided the public goods of peace and security in the region. Ferguson and Schularick (2006) similarly find evidence for a positive empire effect, which enabled countries on the poor periphery to benefit from cheaper capital. The mechanisms behind the empire effect were institutional: a common legal framework protected investor rights (Ferguson and Schularick 2006). Formal and informal empire reduced default risk.

Looking at the period between 1870 and 1914, Mitchener and Weidenmier (2008) find that belonging to an empire doubled trade relative to countries that were not empires or colonies. Important channels for this effect include trade policy and transaction costs. Empires were trading blocks, within which there was often a fixed exchange and free trade with the metropole.[11] Another approach is by Accominotti et al. (2010). Looking within the British Empire, they note that whereas the white dominions were able to benefit from the lower borrowing costs associated with empire, the crown colonies were not able to do so, as their fiscal policy was determined in London. Crown dependencies had very small levels of government debt, and hence did not benefit from the empire, as the white dominions were able to.

The Consequences of the French Revolution

While the French Revolution is the subject of a vast historical literature, it has not been intensively studied by economic historians. Until recently, research focused on the importance of government debt in prompting the crisis of 1789, the inflation that followed the Revolution, and the role that the Revolution played in the nineteenth-century economic divergence that took place between Britain and France.

Recent scholarship explore the political economy consequences of the Revolution. Acemoglu et al. (2011) discuss how the invading French armies dismantled the institutions of the Old Regime that had prevailed for centuries, abolishing serfdom

[11]These articles have come under criticism for not taking into account the full cost of empire (see Coyne and Davies 2007).

and guilds, emancipating Jews, and replacing existing elites. They find that cities in territories that happened to be invaded by the French then grew more rapidly in the following century.

What about the economic consequences of the Revolution in France itself? Rosenthal (1992) argued that Old Regime property rights impeded investment in irrigation and that the Revolution permitted the reallocation of these rights. The French Revolution was also characterized by the massive redistribution of Church property. Finley et al. (2017) exploit the extensive spatial variation in confiscations of Church property to investigate the importance of the initial allocation of property rights for the success of institutional reform. Church property was confiscated and redistributed by auction. Combining disaggregated data on revolutionary confiscations of Church lands with data from agricultural surveys between 1841 and 1929, Finley et al. (2017) find that in regions where more Church land was auctioned off, land inequality was higher in the nineteenth century. This wealth imbalance was associated with higher levels of agricultural productivity and agricultural investments by the mid-nineteenth century. The effects of revolutionary land redistribution on agricultural productivity declined over the nineteenth century as other areas gradually overcame the transaction costs associated with reallocating feudal property rights.

Political Repression

Another side of the state that is worthy of consideration is political repression. Premodern states frequently purged political enemies and scapegoated religious or ethnic minorities – the most frequent victims of this in European history were Jews (Anderson et al. 2017). But in general, they lacked the capacity to conduct political repression on a large scale and for prolonged periods. Society-wide repression and violence occurred, but it was highly destabilizing. Johnson and Koyama (2019) discuss how the extremely costly disruptions associated with the Reformation led European states to reduce their reliance on religion as a source of political legitimacy.

One important premodern example of sustained political repression was the Spanish Inquisition. Rather than being subject to the Pope, the Spanish inquisition was a tool of the Spanish monarchy. Vidal-Robert (2013) argues that the Inquisition was used to suppress domestic opposition when the crown was committed to overseas war. Using data from Catalonia, Vidal-Robert (2014) finds that inquisitional trials reduced population growth during the early modern period. Even today Vidal-Roberts finds people living in areas with historically more intense levels of inquisitorial activity are more likely to think that new technologies will harm them.

Qing China conducted a fierce suppression of "'word crimes" through so-called literary inquisition. Xue and Koyama (2017) study the long-lasting consequences of this political repression. The persecution of intellectuals for their speech and writing led to a fall in social capital – as measured by the formation of local charities – in the decades following a persecution. Over the longer run, areas affected by the literary inquisitions display lower levels of trust, greater political apathy, and worse provision of basic education during episodes of decentralization.

The most widely studied modern episodes of political repression are Nazi Germany and the Soviet Union (Gregory et al. 2011; Harrison 2013). Nazi Germany provides an important setting for investigating the consequences of political repression. More than 1000 academics lost their jobs between 1933 and 1934 because they were of non-Aryan descent (Waldinger 2012). Many were at the pinnacle of their relative fields and would go on to win Noble prizes. The emigration of prominent scientists had both a negative direct effect on scientific output and a negative indirect effect via collaborations and partnerships, worse outcomes for PhD students, and other peer effects.

Revolution, Democracy, Public Goods

Prior to the modern era, states provided few services beyond defense and law and order. The rise of states that provide a broad range of public goods and insurance services is a phenomenon of the last two centuries.

Lindert (2004) documents the gradual rise of public provision from the late eighteenth century onward. Large-scale investment in public education began in Prussia. The French state created a system of secular and compulsory education after 1870. In Britain, investments in public education were slower than elsewhere in Europe. In America, investments in schooling were more decentralized, but by the early twentieth century, the high school movement was giving ordinary Americans access to the best broad-based system of education in the world (Katz and Goldin 2008).

Social insurance and pensions were likewise introduced in Prussia first before spreading to other developed countries. In the United Kingdom, the foundations of the welfare state were laid by liberal government before World War I. In the United States the size and activism of the federal government remained constrained until the Great Depression, but state governments played an important role in providing public goods at the local level.

These investments were driven by political considerations. One influential way to think about these developments is through the threat of revolution. In Acemoglu and Robinson (2000, 2006) the masses threaten revolution to obtain redistribution. If this threat is permanent, elites have an incentive to extend the franchise to make credible their promise to redistribute in the future. Aidt and Franck (2015) find that mobilization, in the form of the Swing Riots, increased the vote share of pro-reform politicians in Parliament before the Great Reform Act.[12]

[12]Dower et al. (2018) address the relationship between the threat of revolution and the emergence of representative institutions using data from Russia during the Great Reforms that abolished serfdom. They find that peasants received less representation in local assemblies (zemstvo) in districts that experienced more frequent peasant unrest in the years preceding 1864. This result is consistent with Acemoglu and Robinson (2000), who predict that political reforms are most likely to be offered when the poor posed only a temporary threat to the establish order. When the poor pose a permanent threat, however, democratization is no longer the sole way the elite can credibly commit to future redistribution.

The revolutionary threat hypothesis is not the only explanation for the transition to democracy. Political scientists have focused on inter-elite bargains and the role played by conservative political parties (Dasgupta and Ziblatt 2015; Ziblatt 2017). Galor and Moav (2006) discuss how the growing importance of human capital in the economy undermined the incentives of workers to overthrow the capitalist system. Galor et al. (2009) argue that while landowning elites had an incentive to oppose the granting of compulsory education, industrialists could actually benefit from this, as the basic skills learnt at schools were complements to industrial production. This helps explain why a political base supportive of compulsory education emerged in the late nineteenth century.

Not all societies experienced a smooth transition to democracy in the nineteenth and twentieth centuries. Carvalho and Dippel (2018) study transitions in power from white planters to colored merchants in the Caribbean. Even though colored merchants were more politically accountable to the citizenry, political outcomes did not improve as much as one might expect. They identify three mechanisms by which an "iron law of oligarchy" persists even in an electoral setting. The legacy of oligarchy, ethnic fractionalization, and weak institutions play an important role in accounting for the relative underdevelopment of many countries in sub-Saharan Africa and the Middle East today as discussed in section "The Political Economy of Empire."

The rise of democracy did not necessarily have a monotonic effect on public good provision. Chapman (2016) argues that the expansion of the franchise in Britain was initially associated with increased public goods spending. However, as voting rights were granted to poorer citizens, spending on public goods and infrastructure actually decreased.

The United States established democratic institutions before any major Western European nation but lagged behind in terms of government investment in public goods. Troesken (2015) discusses the trade-offs faced by authorities in the United States in deciding whether to invest in public health improvements. The decentralized federalist structure of the United States allowed individual states to adopt different public health policies. This facilitated jurisdictional sorting and allowed states to respond to local epidemics. But it was not suitable for coordinating responses to diseases that cross state boundaries. Using the example of small pox, Troesken shows that federalism impeded the introduction of measures such as vaccination against small pox that were crucial for lowering mortality rates. He argues that this same federalist structure ensured greater protection for individual liberty and spurred economic growth, but it came at a cost in terms of public health.

Concluding Comments

Cliometrics is based on the application of economic methods to historical questions. Economic theory can clarify questions and crystalize them in the form of testable predictions that can be taken to the historical evidence. As this evidence is often quantitative, cliometric economic historians have been at the forefront of introducing more formal econometric and statistical methods to the study of history.

Political economy also involves the application of economic models to nonstandard settings. In its modern form, it also stresses the importance of using economic theory to guide the empirical analysis of political questions. Its practitioners likewise need to have command over both economic theory and econometric techniques and knowledge of institutional details and particularities. Cliometrics and modern political economy can thus be viewed as two closely related and complementary fields within economics.

In addition to those topics discussed in this chapter, many other subjects could have been touched upon. Studies of the development of banking institutions in the nineteenth-century United States, for instance, increasingly consider questions of political economy. To what extent did local elites dominate and control banking institutions and to what extent were these institutions characterized by limited versus open access (Bodenhorn 2017)? The topic of trade and tariffs also raises questions of power and the extent to which particular interests are represented politically (Irwin 2017). This chapter has surveyed some prominent topics in the literature from the emergence of the state, the rise of state capacity, the political economy of empire, and the consequences of revolution. Political economy questions intrude on almost all aspects of economic history, leaving much fertile ground for future research.

Cross-References

- ▶ Cliometric Approaches to War
- ▶ Institutions
- ▶ Merchant Empires

References

Accominotti O, Flandreau M, Rezzik R, Zumer F (2010) Black man's burden, white man's welfare: control, devolution and development in the British empire, 1880–1914. Eur Rev Econ Hist 14 (1):47–70

Acemoglu D (2003) Why not a political Coase theorem? Social conflict, commitment and politics. J Comp Econ 31(4):620–652

Acemoglu D (2006) A simple model of inefficient institutions. Scan J Econ 108:515–546

Acemoglu D (2008) Oligarchic versus democratic societies. J Eur Econ Assoc 6(1):1–44

Acemoglu D, Robinson JA (2000) Why did the west extend the franchise? Democracy, inequality, and growth in historical perspective. Q J Econ 115(4):1167–1199

Acemoglu D, Robinson JA (2006) The economic origins of dictatorship and democracy. Cambridge University Press, Cambridge, UK

Acemoglu D, Robinson JA (2012) Why nations fail. Crown Business, New York

Acemoglu D, Johnson S, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. Am Econ Rev 91(5):1369–1401

Acemoglu D, Johnson S, Robinson JA (2002) Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q J Econ 117(4):1231–1294

Acemoglu D, Johnson S, Robinson JA (2005a) Institutions as a fundamental cause of long-run growth. In: Aghion P, Durlauf S (eds) Handbook of economic growth, Vol. 1 of Handbook of economic growth. Elsevier, Amsterdam, pp 385–472. chapter 6
Acemoglu D, Johnson S, Robinson J (2005b) The rise of Europe: Atlantic trade, institutional change, and economic growth. Am Econ Rev 95(3):546–579
Acemoglu D, Cantoni D, Johnson S, Robinson JA (2011) The consequences of radical reform: the French revolution. Am Econ Rev 101(7):3286–3307
Acharya A, Lee A (2018) Economic foundations of the territorial state system. Am J Polit Sci 62 (4):954–966
Aidt TS, Franck R (2015) Democratization under the threat of revolution: evidence from the Great Reform Act of 1832. Econometrica 83:505–547
Alesina A, Spolaore E (1997) On the number and size of nations. Q J Econ 112(4):1027–1056
Allen RC (1997) Agriculture and the origins of the state in ancient Egypt. Explor Econ Hist 34 (2):135–154
Allen RC (2009) The British industrial revolution in a global perspective. Oxford University Press, Oxford
Allen DW (2011) The institutional revolution. Chicago University Press, Chicago
Álvarez Nogal C, de la Escosura LP (2013) The rise and fall of Spain (1270–1850). Econ Hist Rev 66(1):1–37
Anderson RW, Johnson ND, Koyama M (2017) Jewish persecutions and weather shocks 1100–1800. Econ J 127(602):924–958
Ashraf QH, Cinnirella F, Galor O, Gershman B, Hornung E (2017) Capital-skill complementarity and the emergence of labor emancipation. Department of Economics Working Papers 2017–03, Department of Economics, Williams College
Ashworth WJ (2017) The industrial revolution: the state, knowledge, and global trade. Bloomsbury Academic, London
Baechler J (1975) The origins of capitalism. Basil Blackwell, Oxford
Balla E, Johnson ND (2009) Fiscal crisis and institutional change in the Ottoman empire and France. J Econ Hist 69(03):809–845
Banerjee A, Iyer L (2005) History, institutions, and economic performance: the legacy of colonial land tenure systems in India. Am Econ Rev 95(4):1190–1213
Bates RH, Donald Lien D-H (1985) A note on taxation, development, and representative government. Polit Soc 14(1):53–70
Bates RH, Greif A, Levi M, Rosenthal J-L, Weingast BR (eds) (1998) Analytic narratives. Princeton University Press, Princeton
Becker SO, Pfaff S, Rubin J (2016) Causes and consequences of the Protestant Reformation. Explor Econ Hist 62:1–25
Bentzen JS, Kaarsen N, Wingender AM (2017) Irrigation and autocracy. J Eur Econ Assoc 15(1):1–53
Besley T, Persson T (2011) Pillars of prosperity. Princeton University Press, Princeton
Bin Wong R (1997) China transformed : historical change and the limits of European experience. Cornell University Press, Ithaca
Blaydes L (2017) State building in the Middle East. Annu Rev Polit Sci 20:487–504
Blaydes L, Chaney E (2013) The feudal revolution and Europe's rise: political divergence of the Christian and Muslim worlds before 1500 CE. Am Polit Sci Rev 107(1):16–34
Bodenhorn H (2017) Opening access: banks and politics in New York from the Revolution to the Civil War. Unpublished manuscript
Bogart D (2011) Did the Glorious Revolution contribute to the transport revolution? Evidence from investment in roads and rivers. Econ Hist Rev 64(4):1073–1112
Bogart D (2018) Party connections, interest groups and the slow diffusion of infrastructure: evidence from Britain's first transport revolution. Econ J 128(609):541–575
Bogart D, Richardson G (2009) Making property productive: reorganizing rights to real and equitable estates in Britain, 1660–1830. Eur Rev Econ Hist 13(01):3–30

Bogart D, Richardson G (2011) Property rights and parliament in industrializing Britain. J Law Econ 54(2):241–274
Boucoyannis D (2015) No taxation of elites, no representation: state capacity and the origins of representation. Polit Soc 4(3):303–332
Brennan G, Buchanan JM (1980) The power to tax. Liberty Fund, Indianapolis
Brenner R (1976) Agrarian class structure and economic development in pre-industrial Europe. Past Present 70(1):30–75
Brenner R (1993) Merchants and revolution. Princeton University Press, Princeton
Brewer J (1988) The sinews of power. Harvard University Press, Cambridge, MA
Buchanan JM, Tullock G (1962) The calculus of consent. University of Michigan Press, Michigan
Burgess R, Donaldson D (2017) Railroads and the demise of famine in colonial India. Working paper
Caferro W (2008) Warfare and economy in Renaissance Italy, 1350–1450. J Interdiscip Hist 39 (2):167–2009
Cantoni D, Dittmar J, Yuchtman N (2018) Religious competition and reallocation: the political economy of secularization in the Protestant reformation. Q J Econ 133(4):2037–2096
Carneiro RL (1970) A theory of the origin of the state. Science 169(3947):733–738
Carvalho J-P, Dippel C (2018) Elite identity and political accountability: a tale of ten islands. Unpublished manuscript
Centeno MA (1997) Blood and debt: war and taxation in nineteenth-century Latin America. Am J Sociol 102(6):1565–1605
Chaney E (2013) Revolt on the Nile: economic shocks, religion and political power. Econometrica 81(5):2033–2053
Chapman J (2016) Extension of the franchise and government expenditure on public goods: evidence from nineteenth century England. Mimeo
Chaudhary L (2009) Determinants of primary schooling in British India. J Econ Hist 69(1):269–302
Chaudhary L, Rubin J (2016) Religious identity and the provision of public goods: evidence from the Indian princely states. J Comp Econ 44(3):461–483
Chilosi D (2014) Risky institutions: political regimes and the cost of public borrowing in early modern Italy. J Econ Hist 74(03):887–915
Cipolla CM (1976) Before the industrial revolution. Methuen and Co, London
Congleton R (2010) Perfecting parliament: constitutional reform, liberalism, and the rise of Western democracy. Cambridge University Press, Cambridge, UK
Coşgel MM, Miceli TJ (2009) State and religion. J Comp Econ 37(3):402–416
Coşgel MM, Miceli TJ, Rubin J (2012) The political economy of mass printing: legitimacy and technological change in the Ottoman empire. J Comp Econ 40(3):357–371
Coşgel M, Histen M, Miceli TJ, Yildirim S (2018) State and religion over time. J Comp Econ 46 (1):20–34
Cox GW (2016) Marketing sovereign promises: monopoly brokerage and the growth of the English state. Cambridge University Press, Cambridge, UK
Cox G (2017a) The developmental traps left by the Glorious Revolution. Mimeo
Cox GW (2017b) Political institutions, economic liberty, and the great divergence. J Econ Hist 77 (3):724–755
Coyne C, Davies S (2007) Empire: public goods and bads. Econ J Watch 4(1):3–45
Crettez B, Deffains B, Musy O (2019) Legal centralization: a Tocquevillian view. J Legal Stud (forthcoming)
Dasgupta A, Ziblatt D (2015) How did Britain democratize? Views from the sovereign bond market. J Econ Hist 75(1):1–29
De Long JB, Shleifer A (1993) Princes and merchants: European city growth before the industrial revolution. J Law Econ 36(2):671–702
Dell M (2010) The persistent effects of Peru's mining mita. Econometrica Econ Soc 78 (6):1863–1903
Diamond J (1997) Guns, germs, and steel. W.W. Norton, New York

Diebolt C, Haupert M (2018) A cliometric counterfactual: what if there had been neither Fogel nor North? Cliometrica 12(3):407–434
Dimitruk K (2018) "I intend therefore to prorogue": the effects of political conflict and the Glorious Revolution in parliament, 1660–1702. Eur Rev Econ Hist 22(3):261–297
Dincecco M (2009) Fiscal centralization, limited government, and public revenues in Europe, 1650–1913. J Econ Hist 69(1):48–103
Dincecco M, Onorato MG (2016) Military conflict and the rise of urban Europe. J Econ Growth 21 (30):259–282
Dincecco M, Onorato MG (2017) From warfare to welfare. Cambridge University Press, Cambridge, UK
Dincecco M, Prado M (2012) Warfare, fiscal capacity, and performance. J Econ Growth 17 (3):171–203
Dippel C (2014) Forced coexistence and economic development: evidence from native American reservations. Econometrica 82(6):2131–2165
Domar ED (1970) The causes of slavery or serfdom: a hypothesis. J Econ Hist 30(1):18–32
Donaldson D (2018) Railroads of the Raj: estimating the impact of transportation infrastructure. Am Econ Rev 108(4–5):899–934
Dower PC, Finkel E, Gehlbach S, Nafziger S (2018) Collective action and representation in autocracies: evidence from Russia's great reforms. Am Polit Sci Rev 112(1):125–147
Drelichman M (2005) All that glitters: precious metals, rent seeking and the decline of Spain. Eur Rev Econ Hist 9(03):313–336
Ekelund RB, Tollison RD (1981) Mercantilism as a rent-seeking society. Texas A & M University Press, College Station
Elis R, Haber S, Horrillo J (2018) The ecological origins of economic and political systems. Manuscript
Engerman SL, Sokoloff KL (1994) Factor endowments: institutions, and differential paths of growth among new world economies: a view from economic historians of the United States. Working Paper 66, National Bureau of Economic Research
Epstein SR (2000) Freedom and growth, the rise of states and markets in Europe, 1300–1700. Routledge, London
Fenske J (2014) Ecology, trade, and states in pre-colonial Africa. J Eur Econ Assoc 12(3):612–640
Ferguson N, Schularick M (2006) The empire effect: the determinants of country risk in the first age of globalization, 1880–1913. J Econ Hist 66(2):283–312
Fernandez-Villaverde J, Koyama M, Lin Y, Sng T-H (2019) Testing the fractured-land hypothesis: did geography drive Eurasia's political divergence? Working paper
Findlay R, O'Rourke KH (2007) Power and plenty. Princeton University Press, Princeton
Finley T, Franck R, Johnson ND (2017) The effects of land redistribution: evidence from the French Revolution. Working paper
Fleck RK, Andrew Hanssen F (2006) The origins of democracy: a model with application to ancient Greece. J Law Econ 49(1):115–146
Fleck RK, Hanssen FA (2013) How tyranny paved the way to democracy: the democratic transition in ancient Greece. J Law Econ 56(2):389–416
Galor O, Moav O (2006) Das human-kapital: a theory of the demise of the class structure. Rev Econ Stud 73(1):85–117
Galor O, Moav O, Vollrath D (2009) Inequality in landownership, the emergence of human-capital promoting institutions, and the great divergence. Rev Econ Stud 76(1):143–179
Gennaioli N, Rainer I (2007) The modern impact of precolonial centralization in Africa. J Econ Growth 12(3):185–234
Gennaioli N, Voth H-J (2015) State capacity and military conflict. Rev Econ Stud 82(4):1409–1448
Grafe R (2012) Distant tyranny: markets, power, and backwardness in Spain, 1650–1800. Princeton Economic History of the Western World, Princeton University Press
Gregory PR, Schröder PJH, Sonin K (2011) Rational dictators and the killing of innocents: data from Stalin's archives. J Comp Econ 39(1):34–42

Greif A (1998) Self-enforcing political systems and economic growth: late medieval Genoa. Princeton University Press, Princeton, pp 23–64. chapter 1
Greif A (2006) Institutions and the path to the modern economy. Cambridge University Press, Cambridge, UK
Greif A, Rubin J (2015) Endogenous political legitimacy: the English Reformation and the institutional foundations of limited government. Memo
Greif A, Milgrom P, Weingast BR (1994) Coordination, commitment, and enforcement: the case of the merchant guild. J Polit Econ 102(4):745–776
Haber S (ed) (1997) How Latin America fell behind: essays on the economic histories of Brazil and Mexico, 1800–1914. Stanford University Press, Palo Alto
Haber S, Razo A, Maurer N (2003) The politics of property rights. Cambridge University Press, Cambridge, UK
Hall JA (1985) Power and liberties. Penguin Books, London
Harrison M (2013) Accounting for secrets. J Econ Hist 73(04):1017–1049
Herbst J (2000) States and power in Africa: comparative lessons in authority and control. Princeton University Press, Princeton
Hicks J (1969) A theory of economic history. Oxford University Press, Oxford, UK
Hintze O (1906/1975) Military organization and the organization of the state. In: Gilbert F (ed) The historical essays of Otto Hintze. Oxford University Press, Oxford, pp 178–215
Hoffman PT (2011) Prices, the military revolution, and western Europe's comparative advantage in violence. Econ Hist Rev 64(1):39–59
Hoffman PT (2015) What do states do? Politics and economic history. J Econ Hist 75:303–332
Huillery E (2014) The black man's burden: the cost of colonization of French West Africa. J Econ Hist 74(01):1–38
Huning TR, Wahl F (2016) You reap what you know: observability of soil quality, and political fragmentation. BEHL working paper WP2015-05
Irwin DA (2017) Clashing over commerce: a history of US trade policy. University of Chicago Press, Chicago
Iyer L (2010) Direct versus indirect colonial rule in India: long-term consequences. Rev Econ Stat 92(4):693–713
Jedwab R, Moradi A (2016) The permanent effects of transportation revolutions in poor countries: evidence from Africa. Rev Econ Stat 98(2):268–284
Jha S (2015) Financial asset holdings and political attitudes: evidence from revolutionary England. Q J Econ 130(3):1485–1545
Johnson ND (2006) The cost of credibility: The company of general farms and fiscal stagnation in eighteenth century France. Essays Econ Bus Hist 24:16–28
Johnson ND, Koyama M (2014a) Tax farming and the origins of state capacity in England and France. Explor Econ Hist 51(1):1–20
Johnson ND, Koyama M (2014b) Taxes, lawyers, and the decline of witch trials in France. J Law Econ 57:77–112
Johnson ND, Koyama M (2017) States and economic growth: capacity and constraints. Explor Econ Hist 64(2):1–2
Johnson ND, Koyama M (2019) Persecution & toleration: the long road to religious freedom. Cambridge University Press, Cambridge, UK
Jones EL (1981/2003) The European miracle, 3rd edn. Cambridge University Press, Cambridge, UK
Karaman K, Pamuk S¸e (2013) Different paths to the modern state in Europe: the interaction between warfare, economic structure and political regime. Am Polit Sci Rev 107(3):603–626
Karayalcin C (2008) Divided we stand united we fall: the Hume-North-Jones mechanism for the rise of Europe. Int Econ Rev 49:973–997
Katz LF, Goldin C (2008) The race between education and technology. Harvard University Press, Cambridge, MA

Klein A, Ogilvie S (2016) Occupational structure in the Czech lands under the second serfdom. Econ Hist Rev 69(2):493–521
Ko CY, Koyama M, Sng T-H (2018) Unified China and divided Europe. Int Econ Rev 59 (1):285–327
Koyama M (2016) The long transition from a natural state to a liberal economic order. Int Rev Law Econ 47(1):29–39
Koyama M, Moriguchi C, Sng T-H (2017) Geopolitics and Asia's little divergence: state building in China and Japan after 1850. J Econ Behav Organ 155:178–204
Kuran T (2010) The long divergence. Princeton University Press, Princeton
Kyriazis NC, Zouboulakis MS (2004) Democracy, sea power and institutional change: an economic analysis of the Athenian naval law. Eur J Law Econ 17(1):117–132
Lane FC (1958) Economic consequences of organized violence. J Econ Hist 18(4):401–417
Leeson PT (2017) WTF. Stanford University Press, Stanford
Leon G (2018) Feudalism, collaboration and path dependence in England's political development. Br J Polit Sci (forthcoming) https://www.cambridge.org/core/journals/british-journal-of-political-science/article/feudalism-collaboration-and-path-dependence-in-englands-political-development/745D699250AD08C3CC4A963CBD51C2A7
Levi M (1988) Of rule and revenue. University of California Press, London
Lindert PH (2004) Growing public: social spending and economic growth since the eighteenth century. Cambridge University Press, Cambridge, UK
Lowes S, Montero E (2018) Blood rubber. Unpublished manuscript
Ma D, Rubin J (2019) The paradox of power: understanding fiscal capacity in imperial China and absolutist regimes. J Comp Econ (Forthcoming) https://www.sciencedirect.com/science/article/pii/S014759671830194X
Maden JB, Murtin F (2017) British economic growth since 1270: the role of education. J Econ Growth 22:229–272
Mayshar J, Moav O, Neeman Z (2017) Geography, transparency and institutions. Am Polit Sci Rev 111(3):622–636
McCloskey DN (2010) Bourgeois dignity: why economics can't explain the modern world. University of Chicago Press, Chicago
Michalopoulos S, Papaioannou E (2013) Pre-colonial ethnic institutions and contemporary African development. Econometrica 81(1):113–152
Michalopoulos S, Papaioannou E (2016) The long-run effects of the scramble for Africa. Am Econ Rev 106(7):1802–1848
Mitchener KJ, Weidenmier M (2005) Empire, public goods, and the Roosevelt corollary. J Econ Hist 65(3):658–692
Mitchener KJ, Weidenmier M (2008) Trade and empire. Econ J 118(533):1805–1834
Mokyr J, Nye JVC (2007) Distribution coalitions, the industrial revolution, and the origins of economic growth in Britain. South Econ J 74(1):50–70
Musgrave R (1959) Theory of public finance; a study in public economy. McGraw-Hill, New York
Naidu S, Yuchtman N (2013) Coercive contract enforcement: law and the labor market in nineteenth century industrial Britain. Am Econ Rev 103(1):107–144
North DC (1981) Structure and change in economic history. Norton, New York
North DC, Thomas RP (1971) The rise and fall of the manorial system: a theoretical model. J Econ Hist 31(4):777–803
North DC, Thomas RP (1973) The rise of the Western world. Cambridge University Press, Cambridge, UK
North DC, Weingast B (1989) Constitutions and commitment: the evolution of institutions governing public choice in seventeenth century England. J Econ Hist 49:803–832
North DC, Wallis JJ, Weingast BR (2009) Violence and social orders: a conceptual framework for interpreting recorded human history. Cambridge University Press, Cambridge, UK
Nunn N (2008) The long-term effects of Africa's slave trades. Q J Econ 123(1):139–176

Nunn N, Wantchekon L (2011) The slave trade and the origins of mistrust in Africa. Am Econ Rev 101(7):3221–3252
O'Brien P (1982) European economic development: the contribution of the periphery. Econ Hist Rev 35(1):1–18
O'Brien PK (2011) The nature and historical evolution of an exceptional fiscal state and its possible significance for the precocious commercialization and industrialization of the British economy from Cromwell to Nelson. Econ Hist Rev 64(2):408–446
O'Brien PK, de la Escosura LP (1998) The costs and benefits for Europeans from their empires overseas. Rev Hist Econ J Iber Lat Am Econ Hist 16(1):29–89
Ober J (2015) The rise and fall of classical Greece. Princeton University Press, Princeton
Ogilvie S (2007) 'Whatever is, is right'? Economic institutions in pre–industrial Europe (Tawney lecture 2006). Econ Hist Rev 60(4):649–684
Ogilvie S, Carus AW (2014) Institutions and economic growth in historical perspective. In: Aghion P, Durlauf SN (eds) Handbook of economic growth, vol. 2 of Handbook of economic growth, Elsevier, pp 403–513, chapter 8
Olson M (1965) The logic of collective action. Harvard University Press, Cambridge, MA
Olson M (1993) Dictatorship, democracy, and development. Am Polit Sci Rev 87:567–576
Oppenheim F (1922) The state. B.W. Huebsch, New York
Parker G (1976) The "military revolution," 1560–1660–a myth? J Mod Hist 48(2):195–214
Parker G (1988) The military revolution: military innovation and the rise of the West, 1500–1800. Cambridge University Press, Cambridge, UK
Persson T, Tabellini G (2000) Political economics: explaining economic policy. MIT Press, Cambridge, MA
Pincus S (2009) 1688 the first modern revolution. Yale University Press, New Haven/London
Pincus S, Robinson JA (2014) What really happened during the Glorious Revolution? In: Galiani S, Sened I (eds) Institutions, property rights, and economic growth: the legacy of Douglass North. Cambridge University Press, New York
Pirenne H (1925) Medieval cities. Doubleday Anchor Books, New York
Prior SJ (1878) The life of the right honourable Edmund Burke. G Bell, London
Puga D, Trefler D (2014) International trade and institutional change: medieval Venice's response to globalization. Q J Econ 129(2):753–821
Riker WH (1962) The theory of political coalitions. Yale University Press, New Haven
Root HL (1991) The redistributive role of government: economic regulation in old régime France and England. Comp Stud Soc Hist 33(02):338–369
Rosenberg N, Birdzell LE Jr (1986) How the west grew rich, the economic transformation of the industrial world. Basic Books, New York
Rosenthal J-L (1992) The fruits of revolution. Cambridge University Press, Cambridge, UK
Rosenthal J-L, Bin Wong R (2011) Before and beyond divergence. Harvard University Press, Cambridge, MA
Rubin J (2017) Rulers, religion, and riches: why the west got rich and the Middle East did not. Cambridge University Press, Cambridge, UK
Salter A, Young A (2018) Polycentric sovereignty: the medieval constitution, governance quality, and the wealth of nations. Soc Sci Q (forthcoming)
Sng T-H (2014) Size and dynastic decline: the principal-agent problem in late imperial China 1700–1850. Explor Econ Hist 54(0):107–127
Sokoloff KL, Engerman SL (2000) History lessons: institutions, factor endowments, and paths of development in the New World. J Econ Perspect 14(3):217–232
Stasavage D (2002) Credible commitment in early modern Europe: north and Weingast revisited. J Law Econ Org 18(1):155–186
Stasavage D (2003) Public debt and the birth of the democratic state. Cambridge University Press, Cambridge, UK
Stasavage D (2011) States of credit. Princeton University Press, Princeton

Stasavage D (2014) Was Weber right? The role of urban autonomy in Europe's rise. Am Polit Sci Rev 108:337–354
Stasavage D (2016) What we can learn from the early history of sovereign debt. Explor Econ Hist 59(Suppl C):1–16
Stigler GJ (1971) The theory of economic regulation. Bell J Econ Manag Sci 2(1):3–21
Strayer J (1965) Feudalism in western Europe. In: Coulborn R (ed) The idea of feudalism. Archon Books, Hamden, pp 15–26
Tawney RH (1926) Religion and the rise of capitalism. Verso, London
The Prize in Economics 1993 – Press Release (1993). http://www.nobelprize.org/nobelprizes/economic−sciences/laureates/1993/press.html
Tilly C (1975) Reflections on the history of European state-making. In: Tilly C (ed) The formation of nation states in Western Europe. Princeton University Press, Princeton, pp 3–84
Tilly C (1990) Coercion, capital, and European states, AD 990–1990. Blackwell, Oxford
Troesken W (2015) The pox of liberty: how the constitution left Americans rich, free, and prone to infection. University of Chicago Press, Chicago
Vidal-Robert J (2013) War and inquisition: repression in early modern Spain. Working Paper. Department of Economics, University of Warwick
Vidal-Robert J (2014) Long-run effects of the Spanish inquisition, CAGE Online Working Paper Series, Competitive Advantage in the Global Economy (CAGE) 192
Vries P (2015) State, economy, and the great divergence: Great Britain and China, 1680s–1850s. Bloomsbury, London
Waldinger F (2012) Peer effects in science: evidence from the dismissal of scientists in Nazi Germany. Rev Econ Stud 79(2):838–861
Weber M (1922/1968) Economy and society. Bedminster, New York
Weber M (1930) The Protestant ethic and the spirit of capitalism. Allen and Unwin, London
Wittfogel K (1957) Oriental despotism: a comparative study of total power. Yale University Press, New Haven
Wolitzky A, Acemoglu D (2011) The economics of labor coercion. Econometrica 79(2):555–601
Xue MM, Koyama M (2017) Autocratic rule and social capital: evidence from imperial China. Mimeo
Zanden V, Luiten J, Buringh E, Bosker M (2012) The rise and decline of European parliaments, 1188–1789. Econ Hist Rev 65(3):835–861
Ziblatt D (2017) Conservative parties and the birth of democracy. Cambridge University Press, Cambridge, UK

Merchant Empires

Claudia Rei

Contents

Abstract

The Age of Merchant Empires started with the implementation of the Cape Route in 1498 and ended in 1874 with the extinction of the English East India Company. Europeans engaged in the business pursued trade (therefore merchant) and maintained their overseas possessions by force (therefore empires), but population density and established state hierarchies in Asia prevented them from engaging in full-fledged colonialism from the outset. By settling trade outposts in port cities around the Indian Ocean and the Far East, Europeans acquired the necessary network for the supply of a continuous stream of spices, and other Asian goods, to be loaded on ships sailing to Europe. Maintaining an empire required a steady supply of Eastern products implying necessarily the availability of capital and the development of capable shipping technology. But the longevity

C. Rei (✉)
University of Warwick, Coventry, UK
e-mail: c.rei@warwick.ac.uk

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_58

of merchant empires also depended on a sophisticated administration of trade and personnel in Asia and the defense of trade interests. These multi-stranded enterprises were controlled by merchants and/or kings, to whom the prerogative of international trade belonged in early modern Europe. Organizational control had considerable implications on the way merchant empires were run as well as their long-term commercial success. The Asian territorial expansion that some merchant empires pursued reached colonial proportions even before the official start of colonialism in Asia when the administration of overseas territories was formally assumed by governments in Europe.

Keywords
Merchant empires · East India companies · Asian trade · Cape Route · Long-distance trade · Shipping · King · Merchants · Organization · Incentives

Introduction

Following the Great Maritime Discoveries of the fifteenth century, the Age of Merchant Empires marked the period when several European countries established direct trade connections with Africa, Asia, and the Americas. Based on all-sea routes, post-1500 trade broke the centuries old pattern of long-distance trade dominated by Venice and other Italian city states, who controlled Eastern trade coming through the Levant and arriving in the Eastern Mediterranean via land while relying on multiple intermediaries. Merchant companies, on the other hand, gathered investments from multiple agents, developed elaborate organizational structures headquartered in Europe and directly engaged with the distant locations producing exotic products, on which they based their trade. Not only were these the first joint-stock companies in history, they were also the first multinational structures carrying out activities on multiple continents. Regardless of the country they belonged to, these enterprises indelibly marked the histories of the regions where they operated and affected the power structure of the world.

It is often difficult to separate the Age of Merchant Empires from Colonialism, especially in South America where the arrival of European traders in the early sixteenth century occurred hand in hand with territorial encroachment. A combination of low population density, tropical conditions ideal for the production of cash crops, and the availability of precious metals precipitated the rapid conquest and colonization of that particular region. In Africa and especially in Asia, higher population densities and state hierarchies in place led European settlers to secure agreements with local sovereigns allowing for the establishment of coastal warehouses and fortresses to conduct trade on their shores. Such agreements were most often granted under coercion due to the military advantage of Europeans, who would (and non-rarely did) threaten to destroy cities, forge alliances with neighboring rulers, and eventually annihilate entire kingdoms. Reproachable as they were, these strategies resulted in the establishment of trade channels separate from formal territorial control, which was delayed to a later stage.

This chapter focuses on the merchant companies of early modern Europe that conducted trade across wide regions of the globe and not on the formal colonial rule that ensued. Started upon arrival of the Portuguese in India in 1498, England, the Dutch Republic, Denmark, France, and Sweden all established trading companies in the East, which endured until the dissolution of the last East India Company in 1874. Even though this chapter concentrates on merchant operations in Asia, the main focus relies on the European companies.

International Trade in Early Modern Europe

In the sixteenth century, international trade was a royal prerogative. Monarchs could lawfully exert their right of exploration of distant trade sources themselves (as it was the case with the Portuguese king after Vasco da Gama's arrival in India in 1498), or they could instead grant exploration rights to a subject or group of subjects in their realm. Whether owned by kings or merchants, these merchant companies were typically granted exclusive rights within the home country to conduct trade in specific areas for a given time period, both of which were clearly defined in the company's charter. In 1600 Elizabeth I issued the first charter of the English East India Company granting a group of London merchants the exclusive privilege within England, of conducting trade and traffic across sea or land (already known or yet to be discovered) in the area between the Cape of Good Hope and the Strait of Magellan for 15 years. Subsequent renovations of the charter extended the company's privileges for longer periods of time.

These companies operated therefore in very large portions of the world in the absence of a legal international framework to regulate or monitor their activities. They were subject to the set of norms in their charters, specifically designed to fit their novel forms of operation overseas. The charters allowed them to pursue, defend and protect their trade, and exert jurisdiction wherever they conducted business. The formidable power granted to these companies indelibly associates the trade they pursued with the military and political dominance they exerted outside their countries of origin (Findlay and O'Rourke 2007). To understand the various activities carried by these multinational ventures, this section focuses on three main areas of operation of these companies: trade, administration, and defense.

Trade

The prime *raison d'être* of these enterprises was trade in merchandise that was not produced in Europe but where it was in high demand and thus expensive. Asian products, especially spices, had reached Europe centuries before the emergence of merchant companies. Known to make food palatable, spices were also useful food preservatives, especially when appropriate food storage was not an option. Spices were gathered in the producing regions around the Indian Ocean by multiple Muslim traders, who then transported them in caravans through Central Asia and the Levant

until Constantinople or Cairo. From the Eastern Mediterranean, spices were brought by ship to Genoa or Venice and from there distributed to other markets in continental Europe. The many intermediaries of this process and the crossing of various, and not always peaceful, territories made spices in Europe limited and expensive.

Throughout the fifteenth century, Portugal's exploration voyages down the African coast resulted from investments in cartography, navigational instruments, and sailing ships fit for open-sea rather than coastal navigation. The technological prowess of Portugal and her propitious geography outside of the Mediterranean – and thus away from the influence of the Italian City states – facilitated the search for alternative routes to bring spices from Asia into Europe. The discovery of the Eastern passage around the southern tip of Africa in 1487 added to the contemporaneous geographic knowledge that the Atlantic and Indian oceans were linked and provided an early signal of the physical viability of an all-sea route to Asia.

Da Gama's first voyage to India in 1498 was a commercial failure, but subsequent voyages were not. In spite of the volatility of profits of early voyages, the commercial exploration of the Cape Route had substantial implications on European spice markets from the start. After 1503, real pepper prices declined in many European cities, as did the prices of other fine spices, suggesting serious disruptions to the Silk Route (O'Rourke and Williamson 2009). These traditional routes, however, would only be displaced a century later upon arrival of the Dutch in Asia (Steensgaard 1974). Caravan intermediaries gave way to perils in the high seas where storms and pirates could loom with more or less predictability. But the rising importance of the Cape Route in the sixteenth century went beyond the merchants' economic incentive of undercutting European spice prices. It depended heavily on the successful transportation of merchandise from Asia to Europe, implying therefore developments of the sailing ship, the workhorse of merchant empires (Unger 2011).

The sailing ship was the dominant technology for the time period of interest, as it evolved slowly from the year 1000 until the last quarter of the nineteenth century when steamers finally replaced it in long-distance routes (Graham 1956). Countries that were able to innovate and improve the sailing ship to serve their best purposes at any given time period achieved a prominent position in discovery and trade (Rei 2016).

Venice's early technological edge used a combination of oars and sails, allowing for the control of navigation in the Mediterranean roughly from the eighth to the fourteenth centuries. By the fifteenth century, Portugal's exploration voyages required small and easily maneuverable lateen sailed ships appropriate for the exploration of unknown coasts in unchartered waters. After 1498, however, the rising volume of trade increased capacity needs and replaced the earlier light and swift vessels by much larger square-sailed ships, such as galleons and carracks. In Lisbon, the shipbuilding industry grew rapidly and quickly specialized in the production of these very large vessels (Costa 1997). Ultimately, these heavily armed vessels proved inefficient as they were slower and more vulnerable to storms or attacks by foreign vessels. The foundation of the Dutch East India Company (VOC) in 1602 came after a century of Dutch experience in the shipping of bulk goods around Europe, from Portugal to the Baltic (Israel 1989). By the mid-sixteenth

century, the Dutch shipping expertise produced the *fluit*, a relatively small and light cargo vessel with modified sails and hull allowing for relatively larger cargo capacity and no added vulnerability (Eriksson 2014). Though better suited for safe routes in northern Europe or within Asia, *fluits* were not uncommon on the Cape Route and were also exported to most other merchant empires (Barbour 1930). The Dutch competitiveness in shipping stopped after 1650, halting the rapid growth of the Dutch Republic in the eighteenth century (van Zanden et al. 2009). By that time Britain had consolidated her power in the Indian Ocean and had a more powerful navy than any other empire. Long-distance trade leadership therefore cannot be dissociated from technological dominance in shipping.

Administration

The multinational nature of merchant empires required a sophisticated administration spanning across territories under different jurisdictions. Headquartered in Europe, companies faced the challenging task of managing trade operations through a network of overseas ports that could potentially extend from the Cape of Good Hope to Japan. Trade networks differed across empires. Portugal and the Dutch Republic centralized operations in a main port city (Goa and Batavia, respectively), from where they coordinated communications with Europe as well as with other port cities in Asia. England had a more flexible network of outposts, arguably more easily adaptable to changing market conditions (Erikson 2014). Denmark, France, and Sweden had much more limited areas of operation concentrated in specific subregions of Asia. Adding to the geographic complexity, the slow intercontinental communication depended not only on distance but also on the monsoon and the non-negligible risk of loss at sea. Depending on the European country of departure, the voyage from Europe to Asia took some 9 months for vessels departing in the spring, and the return trip from Asia was usually delayed until March of the next year to avoid the stormy season. Arrival in Europe by the end of the year made for an 18- to 24-month commercial cycle, imposing a lag in the implementation of decisions coming from Europe and implying a large degree of autonomy of workers in Asia, especially those in high-rank positions. These features affected the companies' administration of trade, of their own personnel, and also the administration of justice as the resolution of conflicts could not depend on European judicial structures.

Successful trade depended on the timely arrival of vessels and the ready supply of merchandise to be loaded and shipped. These demands required the procurement of goods from local producers (often selling exclusively to the companies in imposed monopsony schemes), to be weighted, valued, and stored in local warehouses until the arrival of company vessels. Coordination of these activities involved vessels arriving from and departing to Europe but also vessels engaged in intra-Asian routes. For instance, Europeans engaged in the very lucrative Japan and China trades, acting as intermediaries between the two Asian powers who had severed contact due to earlier conflicts, while widening the spectrum of Asian products arriving in Europe.

The bureaucratic administration to maintain operations in Asia involved locally hired labor, mostly for unskilled jobs (e.g., loading and unloading ships), but also labor hired in Europe to work in Asia whether in low-rank positions of trust (e.g., guarding warehouses), low-rank skilled positions (such as bookkeeper), or high-rank supervisory positions (such as director of operations in Japan). Hiring people in Europe to work in Asia constituted a classic moral hazard problem. Overseas jobs were attractive as they offered nonseasonal wages, unlike most jobs in Europe, and the potential to make a fortune (Marshall 1976). But these jobs were also notoriously risky: the trip to Asia was perilous, tropical port cities in Asia were stricken by disease, and military conflict could often emerge. From the companies' perspective, the choice of compensation should take into account the job's risks and uncertainties but also elicit honest behavior from agents directly dealing with highly valuable merchandise in distant locations.

Economic theory has long dealt with incentives to induce workers' effort in imperfect monitoring situations (Lazear 1995). In the context of merchant empires, economic historians in this line of work have studied wage structures, private trade, and career progression as mechanisms to deal with the inherent moral hazard problem. Rei (2013) uses a principal-agent model yielding different pay schemes in companies controlled by different parties, which correspond to different monitoring abilities. The model's implications are then illustrated with archival data on labor compensation of Portuguese and Dutch overseas workers. The latter were paid a larger fraction of total compensation in the form of wages when compared to their Portuguese counterparts, which is consistent with Dutch merchants controlling the VOC and, as such, implementing a monitoring structure providing better information on workers than merchants working for the king of Portugal. Hejeebu (2005) focuses on compensation beyond wages and argues that private trade in the English East India Company was complementary to the company's trade. Allowing workers to use the company's resources and enjoy its protection while conducting trade in their own account encouraged workers to fulfill the company's orders so to keep the opportunity of making a fortune of their own. The combination of private trade and a dismissal policy for workers who failed to meet the company's objectives in a given position ensured worker's effort and the pursuit of the company's goals. Rei (2014) uses data on Dutch overseas workers to illustrate the internal career and wage structures of the VOC. There were stable career paths, on-the-job training, fast tracks in promotions largely from within the company's ranks and sizeable returns to tenure. All these papers show that the personnel policies of merchant empires were designed to hire and retain workers, foreshadowing the practices of modern multinational companies.

Other than trade and personnel, merchant companies also administered justice in their geographic areas of action. As empire workers moved to overseas territories where the European court system did not reach, companies were granted judicial authority on behalf of the home countries courts, which constituted a delegation of an important state prerogative. Initially, the administration of the law was aimed at the resolution of conflicts within the community of European settlers, soldiers, privateers, and company officials. But the divide between these subjects and the

indigenous population was not always clear, especially in the Portuguese empire where the policy of intermarriage with locals and the Church's active proselytism extended jurisdictional claims to a growing Christian population (Benton 2002). Other empires may not have followed these same policies, but they faced similar challenges, especially with the expansion of territorial control, which encompassed larger indigenous populations, particularly in the English case.

Like other Europeans, the English encountered pre-existing local legal arrangements in Asia, so they implemented a parallel judicial system where local rulers could still conduct justice independently from the company's courts. The 1661 charter conferred judicial authority in civil and criminal matters directly on the Governor and Council, the only court of law in India until 1683 when the Courts of Admiralty were established (Fawcett 1934). The company's legal control expanded gradually with the company's territorial expansion. The concession of Bombay as a royal dowry of Catherine of Braganza upon her marriage to Charles II in 1661 transferred the city from Portuguese to English hands, but only in 1668 was it transferred to the East India Company, as it became too heavy expense for the English crown. Bombay became one of the main settlements of the company in India, together with Madras and Calcutta, which came under the company's control in 1698. As the company expanded its reach in the subcontinent, it developed a judicial hierarchy resembling that of England, with judges appointed to criminal and civil courts from the ranks of company employees. By 1726 such authority moved to the king marking a turning point in the judicial policy, mainly to avoid litigation against the company in England (Fawcett 1934). Continued attention to the company's affairs led to the Regulating Act of 1773, which was intended to bring profound reform of the company's administration in India.

Defense

The last major area of activity of merchant empires entailed military defense, due to the valuable nature of Asian goods. This scope of activity involved four main aspects that made it necessary. First, trade agreements with local Asian rulers were largely attained under coercion of the non-European party, whose lower military capacity made it impossible to refuse such agreements. Second, these valuable products, transported across the oceans, were coveted by several European countries engaged in long-distance trade and also by pirates in the high seas. Third, all merchant companies were granted monopolies within the home country but competed with each other in Asia for the best locations to establish settlements and acquire a steady supply of goods; very often such commercial activities led to bitter territorial disputes lasting several years, as in the case of the four installments of the Anglo-Dutch Wars. Lastly, as competition between England and the Dutch Republic intensified over the years, so did the urge of the corresponding companies to control larger portions of territory in India and Southeast Asia, some of which were managed as spice-growing plantations using locally enslaved labor, which required coercion power. Moreover, these plantations could themselves be targets of foreign interests and thus had to be defended.

From the early charters, all companies were given exclusive rights of trade and conquest in vast areas of the globe. This opened the door to diplomatic agreements but also to the protection and defense of their trade interests. The conflicting nature of merchant empire operations at multiple levels led, from the very beginning, to the investment in armed ships, resorting to travel in convoys to protect cargo, and the employment of labor exclusively dedicated to military defense. Vessels on the Cape Route carried not only spices but also soldiers who were initially bound to the protection of the companies' fortified settlements but were later involved in key battles with local kingdoms for vast portions of territory.

Portugal's military edge over local sovereigns was very visible upon arrival in Asia in the late fifteenth century, and it was only challenged by the Dutch who, nearly a century later, were able to displace the Portuguese from a large number of ports, especially in the spice-producing regions of Southeast Asia. The English, on the other hand, challenged and displaced the Portuguese from many outposts in the spice-producing region of Malabar in the West Indian coast. With the Portuguese out of the key areas of interest, the seventeenth century was marked by the stark commercial rivalry over trade and territories around the Indian Ocean, yielding to fierce confrontation of the English and the Dutch East India Companies. The Dutch were more successful in the 1600s, but British naval superiority and the decline of the Dutch Republic after the 1730s allowed the English East India Company to control larger portions of territory than any other empire. The decisive battles of Plassey (1757) and Buxar (1764) against Indian powers gave the English East India Company full control of Bengal, thus becoming the major political and military actor in the region.

The control of Bengal catapulted the company to a new level of influence: the original venture of London merchants now commanded a larger army, ruled over a greater number of subjects, and controlled a wider territory than the United Kingdom itself (Mcaulay 1877). Administering justice and having its own army were not unusual for early modern merchant companies, but the East India Company's power reached such unprecedented levels that it effectively became a state within the state. The company's growing authority in India and the consolidation of its commercial position could no longer be dissociated from its military strength. As such, the company came under thorough scrutiny, especially from parliament. Though in favor of maintaining the empire, prominent statesman Edmund Burke denounced the company as "a state in the disguise of a merchant" (Burke 1870: 23). Adam Smith, who mostly argued for the end of the empire and the development of a free-trade relationship with the independent colonies, characterized the East India Company as the "'strange absurdity' of a company state" (Stern 2011: 3). The ensuing imperial debates throughout the nineteenth century resulted in the gradual incorporation of the company into the British state and empire.

International trade in early modern Europe was a risky and expensive venture that went well beyond the pursuit of distantly produced exotic goods. Merchant empires did primarily pursue trade, which depended on the efficient administration of operations in the East and the defense of their interests against rival empires. The monopoly nature of merchant companies operating on multiple continents made

them necessarily large, at a time when high operational costs would advise a smaller optimal firm size (Anderson et al. 1982). Such high costs derived from the multivariate nature of challenges faced by merchant companies in a context of slow communications, constant development of competitive shipping technology, unpredictable dangers associated with weather *en route*, tropical climates prone to disease in Asian destinations, and complexities inherent to the administration of trade, personnel, and justice. These constraints led countries to the use of similar means to explore trade in distant locations. All countries pursued spices in Asia, all used the sailing ship on the Cape Route, and all mounted elaborate operations in search of profit. Yet, countries organized overseas operations very differently. The next section focuses on these differences and the economic incentives behind them.

Contrasting Empires

From the sixteenth to the nineteenth centuries, the expansion of European operations in Asia did not follow a uniform pattern despite sharing similar challenges and objectives. On the one hand, the variety of locations in Asia demanded adjustments to each company's settler strategy. For example, the Dutch settlement in Malacca – a renowned emporium of trade on the strait of the same name, linking the Bay of Bengal and the South China Sea – was necessarily very different from the Dutch settlement at the Cape of Good Hope, the main port of call of the VOC with the prime role of supplying victuals to calling vessels. On the other hand, European countries could have concentrated operations in Asia in a single city to which all other (minor) trade outposts would report and from where all communications to Europe departed and arrived (such was the case of the Portuguese and Dutch empires). Alternatively, they could have chosen a small number of main cities in their trade network to concentrate to serve as regional centers of operation (such was the case of England).

As many of the business decisions in merchant empires, the choice of how to geographically organize operations in Asia derived directly from the strategies devised by the owners of the enterprise located in Europe. If we understand merchant empires as partnerships between kings who had the prerogative of international trade and merchants who effectively conducted and managed trade, organizational control conforms to three distinct possibilities: king/government control (as in the case of Portugal), merchant control (observed in England, the Dutch Republic, and Sweden), and mixed control between king and merchants (as in Denmark and France). The theory of the firm shows that property rights of ownership matter as they determine the ultimate residual claimant of the firm, who controls inalienable business decisions that directly affect the firm's efficiency and success (Grossman and Hart 1986). In the context of merchant empires, the divergent long-term performances highlight the importance of company ownership.

Figure 1 shows the relative shipping of merchant empires to Asia as a percentage of total tonnage per decade in the sixteenth and seventeenth centuries. Alone in the long-distance trade market, Portugal held to her full share of the market until the

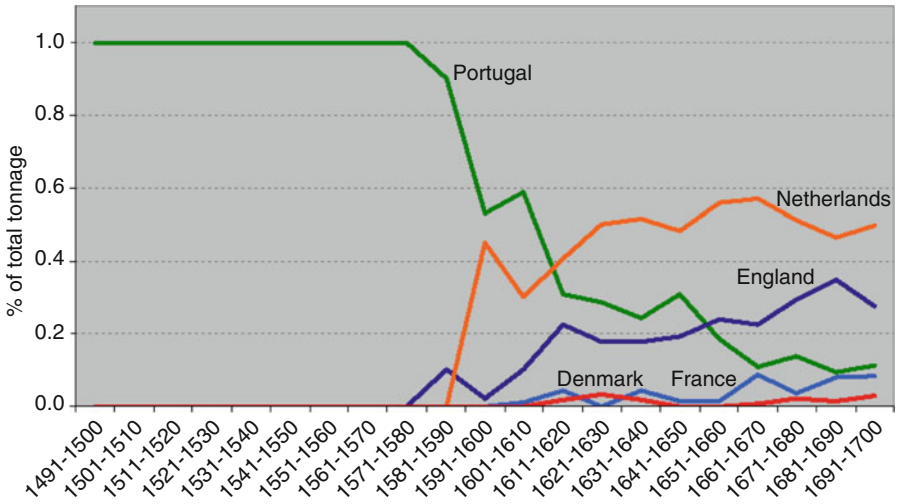


Fig. 1 European shipping to Asia (Source: Rei 2011: 117)

1590s when English and Dutch vessels began competing on the Cape Route shortly before the foundation of their respective trade companies in 1600 and 1602. The continued and marked decline of Portugal's market share throughout the seventeenth century suggests an inability to cope with competition when it arrived. Portugal, lost most of her trade posts in Asia to England and the Dutch Republic, and as such was unable to sustain trade volume at sixteenth century levels, since the decline is also visible in absolute terms. By the end of the 1700s, Portugal was as small a player in the long-distance trade market as were the Danish and French companies that never shipped much to Asia.

If firm ownership is indeed a factor affecting the long-term performance of merchant empires in Asia, it becomes imperative to understand why some kings chose to control the monopoly of trade while others chose instead to charter monopoly rights to private agents. Was Elizabeth I, the English monarch chartering the East India Company in 1600, more farsighted than her distant cousin Manuel I of Portugal in 1498? Were Portuguese merchants less entrepreneurial than English merchants, leaving the king of Portugal no alternative but to control the monopoly of trade himself? Why would the French king share the control of empire with merchants?

This Section is organized in three parts. The first, provides answers to all the above stated questions with a simple version of the model of organizational choice in Rei (2011). The model relies not on differences between monarchs, or merchants in different countries, but on differences between monarchs and merchants in any country. The second part, provides an interpretation of historical events at the emergence of merchant empires in light of the simplified model. The third part, provides evidence on the observed historical differences across merchant empires with different control structures.

A Model of Organizational Choice

Long-distance trade in a given country results from the cooperation between king and merchants. The king can either award the monopoly of trade to private merchants or keep it to himself. Merchants contribute with management skills in the form of effort either in a merchant-controlled firm or as royal employees in a king-controlled firm. King and merchants own cash, which they can either invest in long-distance trade or in other alternatives assumed to yield a lower return, which is consistent with historical evidence on the profits of the East India Companies (Chaudhuri 1965). As such, provided they have cash to spare, king and merchants will invest in long-distance trade.

In the context of uncertainty and incomplete contracts, the probability of success of long-distance trade depends on merchants' effort, non-contractible business decisions by the controlling party, and the realization of uncertainty after the contract is signed.

Preferences of king and merchants are similar with respect to profit that they both want to maximize, but differ substantially in two other aspects. First, merchants effort is costly. Adding a similar term to the king's utility function does not change the result, so this case is ignored for simplicity. Second, the king values being in charge of non-contractible decisions, which enter separately in the king's utility function but not in the merchants'. Adding a similar preference in the merchants' utility function does not change the result so long as merchants care less for business decisions than the king. Again, for the sake of simplicity, this case is ignored. This last characteristic fundamentally distinguishes king and merchants. Even though they both care for trade, which itself is affected by effort and business decisions, the king cares for being in charge of the enterprise, which implies not only the choice of non-contractible business decisions but also ruling over a larger number of people, controlling a vast amount of territory, and spreading his religion and culture, among other features.

The timing of the model develops in three sequential stages. First, merchants and king negotiate over the firm's control structure and the sharing rule of profits to be divided among them; both of these decisions are contractible. Second, after the contract is signed, uncertainty is realized, eliciting the choice of business decisions by the controlling party. Finally, merchants choose the level of effort based on the previously determined uncertainty and business decisions.

Figure 2 shows the graphic solution of the model. Utilities of king (v) and merchants (u) are measured in the yy- and xx-axes, respectively. When king and merchants cooperate, they get utility pairs represented by the higher curved frontiers upon variation of the sharing rule: KC for king control and MC for merchant control. The line closer to the origin represents the set of disagreement points from where king and merchants start the negotiation stage. In the absence of cooperation, these utility pairs represent the king and merchant's outside options given by their alternative choices in the use of cash (c), assumed to have a lower return (λ). This frontier also represents each party's relative bargaining power: moving toward the yy-axis has the king's outside option rising (and the merchants' decreasing) in the no

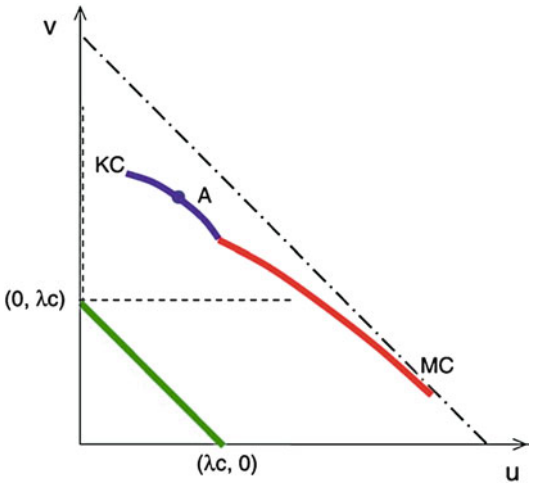


Fig. 2 Graphic solution of the model (Source: Rei 2011: 119)

cooperation case, so the king has a stronger bargaining position than the merchants. The reverse is true for the merchants. In the limit, when the king owns all the cash in the economy, the outside option is (0, λc), from where both parties negotiate, ending up in the KC frontier, say point A, where both parties are better off relative to the case of no cooperation. The reasoning is symmetric for merchants: the more bargaining power they have, the closer they are to (λc, 0), and the negotiation will yield a point in MC. A non-extreme disagreement point could lead to a mixed solution where the control of the enterprise alternates between king and merchants.

In this model, whether the solution lies on MC or KC depends on the relative bargaining power of negotiating parties, which results directly from the firm's financial structure. Thus, institutions emerge because they satisfy the incentives of those in charge, not because they are efficient or socially optimal. An example of a social utility function is the linear sum of king and merchants' utilities. The benevolent social planner would maximize the sum of utilities at the highest indifference curve of slope − 1, which touches the MC frontier closest to the xx-axis yielding very high utility for the merchants and very low utility for the king. Since there is no social planner, both parties maximize their own utility function according to their relative bargaining positions, so we may never attain the socially optimal solution. Either party is strictly better off when in control, that is, KC strictly dominates MC for the king, and the reverse is true for the merchants. Thus, when a party has the bargaining power to control the enterprise, it will never give up on such control.

The implications of the model are very clear: a cash-constrained monarch charters monopoly rights to merchants who control the enterprise, whereas a cash-flush monarch retains ownership of the venture and hires merchants as royal servants to manage the company. The control structure has stark implications on the model's choice variables: merchants work harder in a firm they control, and business decisions differ greatly whether they are made by the king or the merchants. The different choice variables in turn imply different probabilities of success under king

or merchant control, with the latter being larger. The model's implications suggest companies should differ greatly according to the control structure, which provides a vindication of the theory of the firm in the context of merchant empires, but also illustrates the historical differences empirically observed.

Historical Context of the Emergence of Merchant Empires

In light of the model's implications, this section focuses on the historical context of the emergence of merchant empires so as to infer the relative bargaining power of king and merchants in different countries from the surviving evidence on their relative financial endowments.

Vasco Da Gama's arrival in India in 1498 came after nearly a century of crown-sponsored naval expeditions down the African Coast and into the Atlantic in search of the Eastern passage to India. Some of these expeditions were military ventures, mostly in Northern Africa. Others reached uninhabited regions, such as the Atlantic islands of the Azores. Still others reached populated areas with potential for trade in gold, slaves, and Guinean pepper, mostly located in the Gulf of Guinea. This trade started in the late 1430s, giving the crown a comfortable fiscal position, which can be inferred from the frequency of *cortes* – meetings between the king and representatives of the clergy, the nobility, and the general estates with the main purpose of raising taxes. Before the start of the expansion in 1415, *cortes* occurred every one and a half years; from then until 1498, the average time between meetings doubled to 3 years, and after 1498 it tripled to 9.1 years (Rei 2011). This evidence suggests the king of Portugal had strong bargaining power to keep the control of the monopoly of trade in his hands after 1498. Estimates of the labor and capital costs of Portugal's exploration voyages show variable but very high returns on investment for the Portuguese king, further confirming the reinforcement of his strong financial status pre-1498 (Rei 2011). In light of the model, the Portuguese monarch had the financial leverage to keep the ownership of the monopoly of Asian trade, so this is a case of king control.

The chartering of the English East India Company in 1600 came at the end of Elizabeth I's long reign (1558–1603), which was marked by internal and external conflicts. The dispute of the throne by Mary Queen of Scots, the Nine Years' War with Ireland (1594–1603), and the Anglo-Spanish War (1585–1604) all contributed to this turbulent period of English history and added to the crown's financial distress. The pressing financial needs on account of war could not be met by government revenues: from a surplus in 1583–1585, the government balance turned into a deficit by 1597–1600 (Goldsmith 1987). The absence of formal capital markets before the 1688 Glorious Revolution led English monarchs to engage in confiscatory practices by the name of forced loans (North and Weingast 1989). This unpopular form of raising crown revenue hurt the monarch's credibility of repayment and added to the general dissatisfaction that, in time, resulted in the fall of the monarchy. The crown had plenty of pressing needs for the money raised by forced loans, but had the decision been to invest that capital in long-distance trade, it would not have been

enough to cover the costs of the East India Company's voyages in the first 15 years of the charter (Rei 2011). In the context of the model, England's precarious financial situation by 1600 rendered little or no bargaining power to the crown, who chose instead to allocate property rights of trade to the group of merchants who formally requested it, in exchange for customs duties. This is a clear case of merchant control.

From 1595, multiple companies of Dutch merchants sent ships to Asia on the Cape Route to take part in long-distance trade. Competition among these small companies increased spice prices in Asia but lowered prices in Europe as more quantity became available for sale. Encouraged by the Dutch government (not a king in this case), these companies reached an agreement to merge, and the United Dutch East India Company was born in 1602. The early companies had varying degrees of success, but their emergence in the late sixteenth century shows that merchants in the Dutch Republic had the necessary funds to engage in the expensive business of long distance trade, which, in the context of the model, would imply stronger bargaininng power in the negotiation for control. The degree of involvement of the government in the VOC was greater than that of the English crown in the English East India Company. Not only did the government initiate efforts to encourage the merger, it also provided convoy protection to the company's ships in times of war, for which it received an annual payment. The waiving of the first such payment can be seen as a contribution by the Dutch government to the assets of the VOC in the amount of 25,000 fl. corresponding only to 0.4% of the original capital stock of the company (Glamann 1981). As such, the Dutch government seems to have had some degree of bargaining power, since it was able to place the country in better competitive standing relative to other European rivals. But its bargaining power was probably not that great or we would have seen a larger initial investment in the Dutch company. This small initial contribution is suggestive of a cash-constrained Dutch government, still actively involved in the Eighty Years' War (1568–1648) in the first decades of VOC operations. The resulting control structure of the VOC is a clear case of merchant control, but the government had more of a say in the company than in the English case.

Denmark's East India Company, chartered in 1616, had a limited scope of operation in the Indian Ocean, locating mostly on the southern coast of India and in Ceylon. Like the English and the Dutch companies, it was a joint stock, but in many other aspects, it differed greatly from those other companies. The initiative to file the petition of monopoly rights for royal approval came from two Dutch citizens living in Copenhagen and not from Danish nationals, directly contrasting with England and the Dutch Republic. Additionally, Christian IV had a far more active role in the foundation of the company than his English counterpart or the Dutch government. The company faced capital shortages from the beginning, prompting the monarch to force contributions from private investors. The crown itself contributed 12.5% of the necessary capital, the nobility 15.5%, another 35% from citizens of Copenhagen and the university, 29.5% from other towns in Denmark, and 7.5% of the capital from abroad – 2.5% from Hamburg and 5% from the Dutch Republic (Feldbaek 1981). The king again played a decisive role in sponsoring the company through the Thirty Years' War (1618–1648), which was ruinous for Denmark. By

1630, the king had become the owner of half the company's shares, effectively controlling the enterprise. As such, Denmark offers a case of mixed control: the king was never fully able to carry out trade on his own and needed to rely on (or at times force) merchants to participate with part of the necessary capital. The reorganization of the company in 1670 proved similar in terms of bargaining power balance, with the king insisting on the financial support of Copenhagen merchants (Furber 1976). In the context of the model, Denmark's king and merchants would be located in an intermediate disagreement point, thus sharing or alternating control of the company.

The French East India Company was chartered in 1664 and designed by Jean-Baptiste Colbert, the long-time minister of Louis XIV, which denotes a stronger involvement of the king in the enterprise, as in the case of Denmark. This centrally planned company followed a series of unsuccessful companies since the early 1600s, dictating the late arrival of the French in India. The company was conceived to be similar to the VOC, but the lack of interest of French merchants in a company they could not control was evident early on (Manning 1996). Of the original 15 million livres in stock, the crown's subscription of three million forced subscriptions by the royal court, great financiers, and several French cities. Even so, subscriptions totaled only a little over eight million livres, with merchants becoming only a small minority of the shareholders (Furber 1976). From the start the company was plagued with capital problems that worsened with France's involvement in the War of Spanish Succession (1701–1714). As the company turned to the crown for a loan to which the king agreed so long as the directors and stockholders also contributed with given amounts. Barring the Portuguese in India, in no other empire was the involvement of the state more dominant. Still, the forcing of stockholders to engage in further investments shows the king alone was incapable of assuming sole control of the enterprise. Moreover, the incomplete subscription of the company's original capital stock is evidence of the merchants' bargaining power shown in their reluctance to participate in the state company. France, therefore, offers another case of mixed control in light of the model of organizational choice, where the balance of bargaining power probably favored the king a little more than in the case of Denmark.

In 1731 Frederick I chartered the Swedish East India Company after several merchant petitions for a company to engage in Eastern trade. Rather than competing directly with the bigger empires, Sweden specialized in Far Eastern trade with China, notably in tea and porcelain. The late chartering of this company owed to both internal and external factors. On the one hand, Sweden had a belligerent seventeenth century, draining capital to alternatives other than long-distance trade (Hermansson 2004). On the other hand, the suspension and subsequent failure of the short-lived Ostend Company, in 1727 and 1731, created a vacuum of investment opportunities for shareholders in that joint stock, favoring both the already existing Danish company and the newly created Swedish company (Koninckx 1993). The timeliness of the latter event suggests that merchant capital circulated in eighteenth-century Europe but also that kings willing to charter trade companies were relatively more abundant. The end of the Ostend Company resulted from pressure by the British and Dutch governments on Charles VI, the Holy Roman Emperor, to suspend the

company or they would not recognize his daughter Maria Theresa as his rightful successor (Furber 1976). Merchants petitioning the king of Sweden for the company in 1731 further requested the monarch to activate the necessary diplomatic efforts so that the company was recognized by other maritime powers in order to prevent a fate similar to that of the Ostend Company. These additional merchant demands and the late entry of Sweden into the long-distance market suggest the king had little influence in the company and thus no bargaining power beyond that of chartering the monopoly of trade. In the context of the model of organizational choice, Sweden is therefore a clear case of merchant control.

Different Firms

At the foundation of merchant empires, king and merchants' negotiating positions were central to determine the control structure of the enterprises. But the control structure itself implied fundamentally different firms. This section discusses just three areas where these differences were particularly visible.

First, merchant empires were faced with classic agency problems resulting from the nature of long-distance trade. Organizational control led to different incentive structures within the firm, which resulted not only in different pay schemes for personnel hired in Europe to work in Asia but also in different incentives at the upper levels of management. Irwin (1991) argues the Anglo-Dutch commercial rivalry of the early seventeenth century can be understood in light of the subtle institutional differences deriving from the English and Dutch India Companies' charters. At a time when both empires were establishing their foothold in Asia, the Dutch Republic was far more successful at eliminating competitors and gaining territorial control of key ports around the Indian Ocean. As a result, quantities and profits of the Dutch company were substantially greater than those of the English Company. This differential success, Irwin argues, is consistent with a strategic trade policy where the board of directors of the English company maximized profits to please shareholders, while in the Dutch company, they maximized a mix of profits and turnover. Since the English company was controlled by merchants alone, managers only served shareholder interests. But the Dutch company resulted from the government push for the merger of the small pre-companies, and thus, shareholder interests were eroded in favor of the government's. Irwin further calibrates the model with 1622 data evaluating the implications of alternative trade policies.

A second area where organizational differences were manifested was the shipping technology of merchant empires. Even though all empires used the sailing ship on the Cape Route, vessel size differed across empires controlled by different parties. The increasing size of Portuguese vessels on the Cape Route in the sixteenth century was a natural result of the rising trade volume as the all-sea route to Asia was opened. Portugal's ships, however, reached larger sizes than any other vessels on the Cape Route (Boxer 1948), fitting a king-controlled organization: large vessels carried more merchandise and were more fitting of the monarch's strategy of empire that valued glory and prestige, which the very large ships could project in distant

locations (Rei 2016). The problem with these very large vessels was the increased vulnerability in the event of storms or enemy attack, which resulted in higher loss rates. From the early seventeenth century, the Dutch specialized in the construction of smaller and more seaworthy vessels, which they exported to several other empires but never to Portugal, who kept invested in very large vessels well into the first half of the seventeenth century. Portugal's loss rates reflected the trend in ship size increasing from one loss in 10 ships between 1497 and 1550, to nearly one in every five a century later, or one in every seven after dropping losses due to enemy attack resulting from competition on the Cape Route (Rei 2016). The Dutch Republic on the other hand conducted far more voyages than Portugal, at a much lower loss rate of 3%, between 1602 and 1794 (Bruijn et al. 1987). Portugal's relatively higher loss rates could result from aspects beyond vessel size, such as worse shipping quality, or poorer shipping practices, such as travelling outside of convoys or overloading vessels. All these alternative explanations are associated with a poor organizational structure, but they are either unverifiable (shipping quality) or associated with ship size. For example, Portugal also used convoys, but larger vessels were slower and often sailed solo, thus becoming even more vulnerable (Solis 1955). Larger vessels also encouraged the bad practice of overloading, a well-known problem in the Portuguese empire mentioned repeatedly in secondary sources (Guinote et al. 1998). All these variations highlight the impacts of differential organizational control.

Finally, one of the most striking differences between Portugal's royally controlled empire and all other empires was the preoccupation with religion. From the early days, Portuguese ships on the Cape Route carried spices, soldiers, and priests. Trade and defense were common activities of all merchant empires, but only the Portuguese proved to be fervent converters of souls. A naïve approach could potentially see religious interests detracting from the purpose of trade and thus being at the origin of Portugal's decline and relative lack of commercial success in the long run. In merchant empires, the interest in religion (or the lack of thereof) can, however, also be interpreted as a manifestation of organizational control. In the context of the model, monarchs care for being in charge of the enterprise as they retain control over multiple decisions or factors that enter directly in their utility function. A prime example of these factors would be for the king to rule over a large number of people in distant locations, where he could spread his religion and way of life. These concerns were well beyond the scope of trade and touched the domain of colonialism. Kings cared for religion in all countries, but since they had less bargaining power than the king of Portugal in the negotiation for control, we do not see religion take a prime role in any of the merchant or mixed control empires.

The case of England is a very clear example of the monarchs' religious concerns, as the crown itself became the head of the Church in England after 1534 in the context of the English Reformation. Even though the East India Company chose not to interfere with religious beliefs and practices in India, as it would be counterproductive to the successful conduction of business, it could not avoid religious matters altogether. The company employed Christian servants in distant lands and thus allowed for the provision of religious services to them. Moreover, the increasing

scrutiny of parliament over the company's affairs after the Glorious Revolution slowly changed the balance of power between the government (parliament in this case) and merchants, which became clearly visible in the company's attitude toward religion. By 1813, the renewal of the charter required the company to provide for the "'religious and moral improvement' of its Indian subjects" (Carson 2012: 3). The incorporation of the company into the British state was merely 45 years away.

The Demise of the Companies and the Rise of Colonialism

Through their operations in Asia, East India Companies adapted to different market conditions and competitive scenarios. Their fate varied in terms of commercial success and duration of operations. This section focuses on the later life of merchant companies, which laid the ground for the ensuing stage of colonialism when countries took up efforts to establish an official presence in Asia beyond the original trade motives that brought their servants there.

Portugal's case is different from the other empires because the control of the king already implied colonial rule, even though Portugal's policy in Asia never entailed territorial encroachment beyond port cities. The demise of the empire was neither easy nor fast. The arrival of competition in the Indian Ocean saw Portugal's role in the Asian market decline substantially as most of her trade outposts in the Indian subcontinent and Southeast Asia either fell to rivals or were conceded in diplomatic agreements that recognized Portugal's restored independence from Spain in 1640. By the early eighteenth century, Portugal's minimal presence in Asia included three port cities on the west coast of India (Goa, Daman, and Diu) and also East Timor and Macau in the Far East. At this time, the Portuguese contribution to Eastern trade was negligible with Lisbon purely maintaining annual communication with the East and allowing returning ships to participate in the Brazil trade, where Portugal had redirected her colonial efforts. Despite not serving trade purposes anymore, the rusty structure of Portugal's empire in Asia endured long after the peaceful independence of Brazil in 1822. The slow death is consistent with the preferences of the party in control that encompassed more than trade objectives. The empire outlived Portugal's monarchy (which fell in 1910), the first republic (which collapsed in 1926), and was cherished by the fascist dictatorship that ensued. The start of the Colonial War in Africa in early 1961 precipitated events in Portuguese India, which fell in December 1961 with the Annexation of Goa, Daman, and Diu into the Republic of India, itself independent from the United Kingdom in 1947. The fall of Portugal's dictatorship in 1974 ended Portugal's Colonial War and started the clumsy decolonization process concluded in 1975. East Timor was abandoned that same year and subsequently invaded and occupied by Indonesia, her much larger neighbor. Macau remained a territory under Portuguese administration until its agreed transfer to China in 1999.

At the time of its demise in 1874, the English East India Company was a very different entity from the one chartered by Elizabeth I in 1600. The rebound from fierce competition by the Dutch in the seventeenth century and the subsequent

consolidation of territorial control resulting from the Napoleonic Wars saw the company strengthen its establishment in Asia in the eighteenth and nineteenth centuries. The battles of Plassey and Buxar marked the clear expansion of the company's territorial reach, strengthened by the favorable outcome of the fourth (and last) Anglo-Dutch War (1780–1784), which further consolidated the company's territory. The rising importance of nation states in the eighteenth century meant closer examination of the company's activities due to conflicting sovereignty of overseas territories and necessarily implied the decline of the company as a trade body, as it was incompatible with the state (Hejeebu 2016). Successive acts of parliament since 1773 curbed the powers of the company in a series of attempts to create the legal framework for the scope of its international operations. The 1813 Act was especially relevant, as it opened India to missionaries, for the first time raising concerns of religion in the context of a merchant company whose control was no longer purely merchant. The 1833 Act, on the other hand, removed all remaining trade monopolies from the company, officially relegating its commercial interests to second order, while renewing the company's political and administrative authority in India on behalf of the crown. The gradual transformation of the company into a branch of the British administration came to a conclusion with the 1858 Act that nationalized the company in the sequence of the Indian Rebellion of 1857. The company continued to manage the tea trade on behalf of the government until its stock was formally dissolved in 1874, making it the longest-lived merchant India Company. Formal colonial rule in British India started then.

The astounding success of the VOC in the seventeenth century was not replicated in the second century of operation of the company. The decline of the Japan and the China trades, especially after the loss of Formosa in 1662, led to a reorientation of the VOC toward Bengal, though the main strongholds of the Dutch were still in the Southeast Asian archipelago. In the long run, the geographic concentration of the VOC became a disadvantage as European consumption patterns began shifting away from spices and toward Indian textiles, raw silk, tea, and coffee, starting in the late seventeenth century (Chaudhuri 1978). The long decades of hostilities between the Dutch and English companies came to an end after the Glorious Revolution, which placed William of Orange, the Dutch head of state, on the English throne. This major shift in the balance of power between the Dutch and the English saw the Dutch drop their policy of ruthless elimination of all European competitors toward England (Chaudhuri and Israel 1991). Barred from fighting the English, who benefited from the break strengthening their positions in the Malabar coast, Bengal, and the Persian Gulf, the Dutch singlehandedly pursued fight with the French in southern India. The concentration of efforts in this area in the 1690s weakened the VOC's overall position in Asia. From the 1730s, the gradual exclusion of the VOC from various areas of commercial interest, and the symmetric rise of the English company, led to the steady decline of the VOC's revenue. Peace with the British came to an end when the Dutch began supporting the rebels, who had revolted against the British in the American Revolutionary War (1775–1783), quickly leading the way to the Fourth Anglo-Dutch War. The beleaguered Dutch Republic fell in 1795 in the context of the French Revolutionary Wars (1792–1802), bringing major disruption to the United Provinces that were only to regain independence in 1813. The VOC

was nationalized and its charter revoked in 1800. The territories of the VOC came under the administration of the Dutch government, starting a period of formal colonial rule in the Dutch East Indies.

Repeated financial problems of the French company led to multiple reorganizations throughout the company's history. Despite its smaller role on Asian trade, the French company showed territorial aspirations early. Successive attempts to establish a colony in Madagascar from the early days proved fruitless until the company finally gave it up in 1685, transferring it back to the royal domain. The decline of the Mughal Empire on the other hand led the French company to intervene in order to protect and expand its territorial interests in southeast India, implementing a policy of alliances with local rulers in the 1740s. These ambitions confronted directly with British interests in the region and confrontation between the two companies ensued in 1744. Though tempting, given the local conditions in India, the expansionary policy proved disastrous to the company's finances, bringing substantial reductions in the company's dividends (Boulle 1981). The situation was further aggravated by the Seven Years' War (1756–1763), which opposed Britain and France in Europe and rekindled the old conflict between the two East India companies in Asia. By 1769, the French company was no longer able to support its debts and was abolished by the crown, which took over the administration of French India, starting a period of formal colonial rule.

Though the Danish and the Swedish East India Companies had different control structures and different starting times, they were both dissolved in the early 1800s. Denmark had a more eventful history in Asia than its Scandinavian counterpart, due largely to the initial capital difficulties that were recurrent and resulted in the bankruptcy of the first two Danish companies in 1650 and 1728, respectively. The third Danish company also resulted from the cooperation of king and merchants in the initial capital stock. Upon dissolution of the French East India Company in 1769, Copenhagen merchants argued for the liberalization of the India trade, and, when renewing the 1772 charter, the state kept the monopoly of China trade but gave up the monopoly of India trade, allowing private merchants to compete directly with the company for a fee. At the same time, the company retained the administration of territories in India, giving rise to conflicts with private merchants (Feldbaek 1981). War with Britain in the context of the Napoleonic Wars (1803–1815) with the attack on Copenhagen in 1807 led to the end of operations of the company, while the remaining Danish settlements in India were sold to Britain in 1845. The Swedish company also succumbed at British hands, but differently. The company's charters were renewed in succession for periods of 20 years, but the 1784 Commutation Act enacted by the British parliament dropped the tea tax from 119% to 12.5%, effectively ending the smuggling of tea into Britain and putting the Swedish company's activity in jeopardy. Profits fell sharply, and by 1811 the Swedish company declared bankruptcy. Both the Danish and Swedish companies had small impact on Asian trade, in part because neither of them controlled much territory in Asia. The impossibility of victualing vessels en route, or the dependence on trade adversaries, hindered trade prospects of these smaller companies. Nevertheless, participation in Asian trade proved rather meaningful for the small Nordic countries, none of which engaged in subsequent colonial efforts.

Conclusion

The emergence of the Cape Route in 1498 opened Asian trade to countries with sufficiently developed shipping technology and enough capital to invest in long-distance trade, in the pursuit of notable arbitrage opportunities. Mounting expeditions involved producing ships, fitting and victualing them, as well as paying sailors, merchants and soldiers. Such expensive tasks led to innovative forms of pooling capital in the form of joint-stock companies, mostly participated by merchants but also by monarchs who held the royal prerogative of international trade in early modern Europe. The relative contributions of king and merchants led to varying control structures, with relevant implications on the organization's incentive structure. In Portugal, the king was wealthy enough to fund trade operations and thus retained control of the enterprise, effectively starting colonialism though without territorial expansion beyond port cities. In England, the Dutch Republic, and Sweden, merchants had stronger bargaining power than the monarch (or the government in the Dutch case) and controlled their respective merchant companies. In the case of Denmark and France, organizational control was shared between merchants and king, who retained considerable influence in the companies' decisions, especially in France.

Maintaining empires involved more than merchant initiative; it also required a powerful army and navy as well as the administration of operations in Asia coordinated from Europe, leading to a classic moral hazard scenario. Companies dealt with these administrative problems in ways resembling modern corporate practices, but they could not avoid the belligerent nature of operations in Asia, where the defense of trade interests often originated conflicts between competing parties. Additionally, events in Europe impacted operations in Asia, notably various wars but also the Glorious Revolution and the French Revolution. Finally, the waging of war against local powers also led companies to act well beyond the commercial objectives that led Europeans to Asia in the first place. From merchant entities, companies gradually transformed into political and military units, laying the ground for formal colonialism, which ensued after the demise of the companies, except in the case of the Danish and Swedish East India Companies.

The demise of the companies revealed a very different world than the one that had seen them emerge. Country monopolies were contested as more subjects wanted to take part in Eastern trade. The companies' sovereignty over distant territories and subjects raised legal and moral concerns that the companies were never designed to address. In the end, companies that were able to adapt to changing circumstances in Asia fared better in the long run.

References

Anderson GM, McCormick RE, Tollison RD (1982) The economic organization of the English East India Company. J Econ Behav Organ 4(2–3):221–238

Barbour V (1930) Dutch and English merchant shipping in the seventeenth century. Econ Hist Rev 2:261–290

Benton L (2002) Law and colonial cultures: legal regimes in world history; 1400–1900. Cambridge University Press, Cambridge
Boulle PH (1981) Chapter 6, French mercantilism, commercial companies and colonial profitability. In: Blussé L, Gaastra F (eds) Companies and trade. Leiden University Press, Leiden, pp 97–117
Boxer CR (1948) Fidalgos in the Far East 1550–1770: fact and fancy in the history of Macao. Martinus Nijhoof, The Hague
Bruijn JR, Gaastra FS, Schöffer I, with assistance of van Eyck van Heslinga ES (eds) (1987) Dutch-Asiatic shipping in the 17th and 18th centuries. Vol. I, introductory volume. Martinus Nijhoff, The Hague
Burke E (1870) The works of the right honorable Edmund burke. Vol. 7: speeches on the impeachment of Warren Hastings. Bell and Daldy, London
Carson P (2012) The East India company and religion, 1698–1858. Boydell & Brewer, Suffolk
Chaudhuri KN (1965) The English East India company: the study of an early joint-stock company, 1600–1640. Cass, London
Chaudhuri KN (1978) The trading world of Asia and the East India company – 1660–1760. Cambridge University Press, Cambridge [Eng.]/New York
Chaudhuri KN, Israel JI (1991) Chapter 13. The English and Dutch East India companies and the glorious revolution of 1688-9. In: Israel JI (ed) The Anglo-Dutch moment: essays on the glorious revolution and its world impact. Cambridge University Press, Cambridge, UK, pp 407–438
Costa LF (1997) Naus e galeões na ribeira de Lisboa: a construção naval no século XVI para a Rota do Cabo. Patrimonia, Cascais
Erikson E (2014) Between monopoly and free trade: the English East India company 1600–1757. Princeton University Press, Princeton and Oxford
Eriksson N (2014) Urbanism under sail – an archeology of fluit ships in early modern everyday life. Elanders, Stockholm
Fawcett C (1934) The first century of British justice in India. Oxford University Press, Oxford
Feldbaek O (1981) In: Companies and trade, Blussé L, Gaastra F (eds) Chapter 8 The organization and structure of the Danish East India, West India and Guinea companies in the 17th and 18th centuries. Leiden University Press, Leiden, pp 135–158
Findlay R, O'Rourke K (2007) Power and plenty: trade, war, and the world economy in the second millennium. Princeton University Press, Princeton
Furber H (1976) Rival empires of trade in the orient, 1600–1800. University of Minnesota Press, Minneapolis
Glamann K (1981) Dutch-Asiatic trade: 1620–1740. Den Haag, Nijhoff
Goldsmith RW (1987) Premodern financial systems: a historical comparative study. Cambridge. Cambridge University Press, New York
Graham GS (1956) The ascendancy of the sailing ship 1850-85. Econ Hist Rev 9(1):74–88
Grossman SJ, Hart O (1986) The costs and benefits of ownership: a theory of vertical and lateral integration. J Polit Econ 94(4):691–719
Guinote P, Frutuoso E, Lopes A (1998) Naufrágios e outras perdas da 'Carreira da Índia': séculos XVI e XVII. Grupo de Trabalho do Ministério da Educação para as Comemorações dos Descobrimentos Portugueses, Lisboa
Hejeebu S (2005) Contract enforcement in the English East India company. J Econ Hist 65(2):496–523
Hejeebu S (2016) Chapter 3. The colonial transition and the decline of the East India company, c. 1746–1784. In: Chaudhary L, Gupta B, Roy T, Swamy AV (eds) A new economic history of colonial India. Routledge, London/New York, pp 33–51
Hermansson R (2004) The great East India adventure: the story of the Swedish East India company. Breakwater Publishing, Göteborg
Irwin DA (1991) Mercantilism as strategic trade policy: the Anglo-Dutch rivalry for the East India trade. J Polit Econ 99(6):1296–1314

Israel JI (1989) Dutch primacy in world trade, 1585–1740. Clarendon Press, Oxford
Koninckx C (1993) Chapter 5. The Swedish East India company. In: Bruijn JR, Gaastra FS (eds) Ships, sailors and spices: East India companies and their shipping in the 16th, 17th, and 18th centuries. NEHA, Amsterdam, pp 121–138
Lazear EP (1995) Personnel economics. MIT Press, Cambridge/London
Mcaulay TB (1877) Government of India (10 July 1833). In: Speeches of Lord Macaulay, corrected by himself. Longman, Green, and Company, London
Manning C (1996) Fortunes a faire ñ the French in Asian trade, 1719–48. Ashgate Pub, BrookÖeld
Marshall PJ (1976) East Indian fortunes. Oxford University Press, Oxford
North D, Weingast B (1989) Constitutions and commitment: the evolution of institutional governing public choice in seventeenth-century England. J Econ Hist 49(4):803–832
O'Rourke KH, Williamson JG (2009) Did Vasco da Gama matter to European markets? Econ Hist Rev 62(3):655–684
Rei C (2011) The organization of eastern merchant empires. Explor Econ Hist 48(1):116–135
Rei C (2013) Incentives in merchant empires: Portuguese and Dutch labor compensation. Cliometrica 7(1):1–13
Rei C (2014) Careers and wages in the Dutch East India company. Cliometrica 8(1):27–48
Rei, Claudia (2016) Turning points in leadership: Shipping technology in merchant empires, Manuscript
Solis, Duarte Gomes (1955) Alegación en favor de la Compañia de la India Oriental comercios ultramarinos, que de nuevo se instituyó en el reyno de Portugal. Edição organizada e prefaciada por Moses Bensabat Amzalak. Editorial Império, Lisboa
Steensgaard N (1974) The asian trade revolution of the seventeenth century – the East India companies and the decline of the caravan trade. The University of Chicago Press, Chicago
Stern PJ (2011) The company-state: corporate sovereignty and the early modern foundations of the British empire in India. Oxford university Press, Oxford
Unger RW (2011) Shipping and economic growth 1350–1850. Brill, Leiden/Boston
Zanden V, Luiten J, van Tielhof L (2009) Roots of growth and productivity change in Dutch shipping industry, 1500-1800. Explor Econ Hist 46(4):389–403

Colonial America

Joshua L. Rosenbloom

Contents

Abstract

The first permanent British settlement in what became the United States was established in 1607, nearly 170 years prior to the American declaration of independence. This chapter examines the economic development of the British North American colonies that became the United States. As it describes,

J. L. Rosenbloom (✉)
Department of Economics, Iowa State University, Ames, IA, USA

NBER, Cambridge, MA, USA
e-mail: jlrosenb@iastate.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_60

abundant natural resources contributed to the remarkable growth in the size of the colonial economy and allowed the free white colonial population to enjoy a standard of living among the highest in the world at that time. While living standards were comparatively high, there was not much improvement over time. Scarce labor and capital, the corollary of abundant resources, also played an important role in shaping colonial institutions, encouraging reliance on indentured and enslaved labor and influencing the development of representative governmental institutions. The colonial period ended with the American declaration of independence in 1776. For most of the colonial era, the colonists had happily accepted their relationship to Britain. Changes in British policies following the end of the Seven Years War in 1763, however, created tensions between Britain and the colonies that ultimately led to the colonies to declare their independence.

Keywords
Economic growth · North America · Mercantilism · Slavery · Institutions · Colonics

Introduction

Reflecting the dominant themes in the cliometric literature, this chapter is concerned with the economic history of those British mainland North American colonies that became the United States in 1776. It is important at the outset to acknowledge the backward-looking nature of this selection criterion. During the seventeenth and eighteenth centuries, Britain established a number of other colonies in the Americas, including parts of coastal Canada and the West Indies. At the same time, other European nations were also engaged in colonization efforts in North America. The Spanish had established colonies in parts of the Southwest and Florida, France had colonized Quebec, and until the 1660s, the Netherlands controlled parts of what would become New York and New Jersey.

Focusing on European colonization also diverts attention from the experience of the indigenous peoples who had occupied North America for millennia prior to the arrival of European explorers. For this latter group, European colonization proved profoundly destructive. Exposure to European diseases, such as smallpox, decimated native populations. Natives were exposed to these diseases through contact with European fishing expeditions even before permanent European settlements were established. Thus, when the first permanent European settlements were established, they encountered indigenous communities that were already disrupted, faced less resistance than might otherwise have been the case, and were often able to occupy lands that had been cleared by native inhabitants.

Despite the merits of these different perspectives in informing historical understanding of the colonial era, the cliometric literature has mostly adopted a vantage point that casts the history of this phase of American economic development as background for the subsequent development of the United States, asking how

developments in the years leading up to American independence from Britain shaped the subsequent evolution of the US economy. This chapter will largely follow this approach.

Economic Performance and Living Standards

Income at the End of the Colonial Period

Quantitative data on which to base measurements of income in the colonial period are quite limited. Nonetheless, recent scholarship has shed additional light on incomes near the end of the colonial period, suggesting that by this time, free white residents of North America enjoyed living standards that compared favorably with Britain, which, according to Maddison's estimates (Bolt and van Zanden 2013), had the highest per capita income in the world at the time.

Allen et al. (2012) have gathered time series of unskilled wages and the cost of living for workers in three British North American colonies as well as in a number of other locations around the world. Converting nominal wages in each location into their equivalent in grams of silver and then deflating these by the cost of subsistence, they have computed comparative welfare ratios for each location. These comparisons are depicted in Fig. 1. By the time of the American Revolution, laborers in Philadelphia had the highest earnings of any location represented in their data, roughly 25% higher than laborers in London. Laborers in Boston and Maryland,

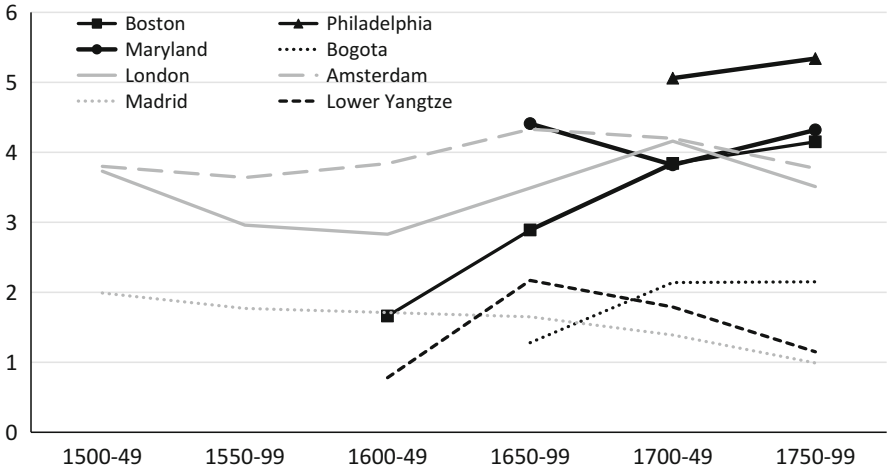


Fig. 1 Subsistence ratios (wages/cost of subsistence), by location, 1500–1799. (Source: Robert C. Allen, Tommy E. Murphy and Eric B. Schneider, "The Colonial Origins of the Divergence in the America: A Labour Market Approach," IGIER- Universita Bocconi, Working Paper no. 402 (July 2011), Table 4, p. 45)

which had lagged behind London, were by the 1770s quite close to London and well ahead of their counterparts in South America and China.

Peter Lindert and Jeffery Williamson (2016a) have undertaken the more ambitious task of constructing estimates of national income by combining social tables, describing the occupational and class structure of the population, with estimates of labor and property income for each group. Their approach allows them to construct aggregate per capita income estimates in 1774 for the colonies as well as examine the distribution of income within the colonies. Their estimates show that per capita incomes were higher in America than England by 1774 and much more equally distributed. According to their estimates, the Gini for free American colonists was 0.4, well below the average of 0.57 calculated for four Northwest European countries at the time. Even including the enslaved, their North America Gini rises only to 0.44. Adjusting for exchange rates and cost of living differences is challenging, as they acknowledge, but what seems clear is that while incomes at the top of the distribution remained higher in England, the colonies offered considerable economic opportunity for those further down the income distribution, a fact consistent with the rising tide of British immigration to North America in the decades before the Revolution.

Economic Growth

Estimating longer run trends in colonial incomes is more difficult. Demographic evidence attests to the robust extensive growth that characterized the colonial era. The first permanent British settlement in colonial America was established in 1607, at Jamestown, in what is now Virginia. A second settlement was established in Massachusetts in 1621. By the time the American colonies declared their independence in 1776, the population of colonial America had increased from a few hundred settlers to approximately 2.5 million. Nonetheless, this population remained confined primarily to a relatively narrow strip of land along the Atlantic seaboard stretching from present-day Georgia, in the South, to what is now Maine, in the North.

The growth of the European and African populations in colonial America was accompanied by a decline in the indigenous population. Estimates of the size of the precontact indigenous population vary considerably, but there is little question that smallpox and other European diseases hit the native population quite hard. Ubelaker (1988) estimates that the indigenous population east of the Mississippi River fell from about half a million in 1600 to 254,485 in 1700 and 177,630 in 1800.

Relying on a variety of censuses, tax rolls, and other documents, historical demographers have been able to work out reasonably detailed estimates of the growth of European-American and African-American populations. High rates of fertility, early marriage, and relatively small numbers never married combined to produce rapid rates of natural increase throughout the colonies. Voluntary migration and the importation of slaves further increased population growth rates. Together these factors contributed to population growth rates close to 2.8% per year, on

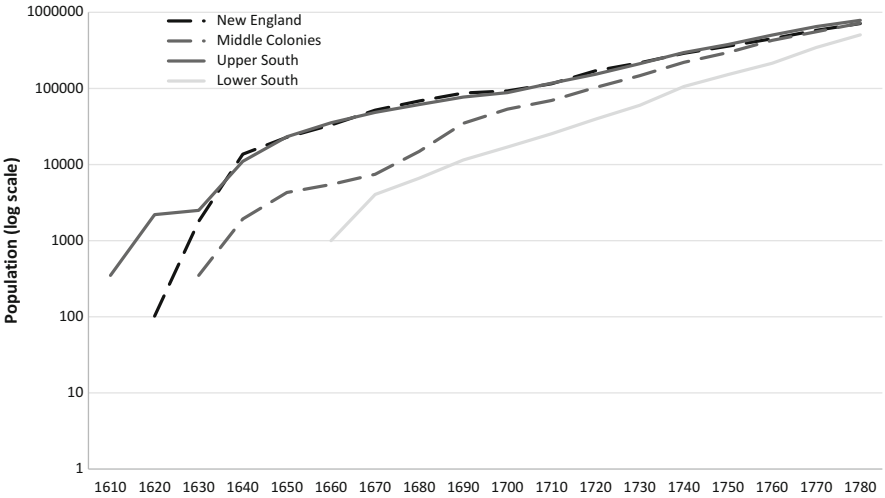


Fig. 2 Regional population growth, 1600–1780. (Source: Carter et al. (2006), series Eg1–59)

average, for most of the colonial period. This rate was sufficient to produce a doubling of population every generation. One contemporary observer, Thomas Malthus, characterized the rate of increase as "probably without parallel in history" and used it as support for his contention that in the absence of constraints, population would increase at a geometric rate (Galenson 1996, p. 169).

Figure 2 plots regional growth in population numbers on a semilog scale. Growth rates in the first few decades of settlement were quite rapid, reflecting the contributions of immigration to an initially small base, but then slowed as natural increase became the dominant source of growth. As the regions of earliest settlement, the Chesapeake and New England accounted for virtually all of the colonial population through the end of the seventeenth century. After 1680, however, the Middle Atlantic colonies (New York, New Jersey, and Pennsylvania) attracted growing numbers of immigrants and expanded rapidly. Settlement of the Lower South (the Carolinas and Georgia) did not begin in earnest until nearly 1700, but thereafter the region grew quite quickly, though the population of this region remained much smaller than the other colonial regions.

The early colonists experienced extreme hardships as they adjusted to a new land. Shortages of food, the challenges of adapting to conditions, and the disease environment all contributed to initial high rates of mortality (Perkins 1988, p. 6). As colonial settlements became more established, however, living conditions improved and mortality rates declined.

Quantifying early living standards has, however, proved difficult. Based on the robust growth of colonial population and the diversifying economy it supported, early accounts assumed that per capita incomes must have been rising in the colonial period. McCusker and Menard (1985), for example, in their influential assessment of the state of colonial economic history, suggested that per capita incomes in the

eighteenth century must have grown at least as fast as British per capita income and might have grown twice as fast – leading them to suggest that per capita income growth was in the range of 0.3 to 0.6% per year.

More recent scholarship has argued, however, on the basis of new data and more refined analytical techniques, that after overcoming the initial challenges of settlement, the pace of aggregate economic growth was quite small. In view of the limited quantitative data available for the colonial period, these estimates rely largely on backcasting income levels using indices for a few key indicators. Mancall and Weiss (1999), for example, applied the method of controlled conjectures to construct per capita GDP estimates for the period 1700–1800. They began with the identity that per capita GDP is equal to output per worker times the labor force participation rate. That is:

$$Q/P = (L/P) * (Q/L) \tag{1}$$

where Q is GDP, P is population, L is the labor force, L/P is labor force participation, and Q/L is output per worker.

Output per worker can then be decomposed into a weighted sum of per worker productivity in different economic sectors:

$$\begin{aligned} Q/P &= (L/P) * \left[(1 - S_a) * (Q/L)_n + S_a(Q/L)_a\right] \\ &= (L/P) * (Q/L)_a[(1 - S_a)k + S_a] \end{aligned} \tag{2}$$

where the subscripts a and n denote agriculture and nonagriculture, respectively, S_a is the share of the labor force employed in agriculture, Q/L is average output per worker, and k is the ratio of output per worker in nonagriculture to agriculture.

Beginning with known values of per capita GDP in 1800 and assuming that relative labor productivity, k, was constant at its 1800 level, it is possible to project backward per capita income on the basis of estimates of just three series: labor force participation, the sectoral distribution of labor, and labor productivity in agriculture. The first two series can be derived primarily from available demographic data and rest on a relatively sound quantitative base, but measuring labor productivity is more challenging. To derive estimates of agricultural labor productivity, Mancall and Weiss first computed total food production by aggregating estimates of consumption by different demographic groups (e.g., children, adult males, adult females, and slaves) and adding net exports. Dividing total food production by the aggregate labor force results in a measure of average labor productivity.

In the absence of firm evidence about trends in food consumption, their baseline case assumed constant levels of consumption of agricultural products over time, an assumption they justified based on the constancy of military rations over time. Combining all of the evidence, they calculated that per capita income (expressed in 1840 prices) increased only from $64 in 1700 to $68 in 1770 and then fell to $67 in 1800. This is a growth rate of just 0.08% per year for the shorter period and 0.04% per year for the entire century.

Mancall and Weiss explicitly acknowledged that they could not definitively measure agricultural production; nonetheless they noted that their estimates of per

capita GDP were constrained by the range of plausible values for domestic agricultural production. Assuming a more rapid rate of growth of agricultural productivity, for example, would result in higher rates of per capita GDP growth, but even assuming that agricultural productivity grew as fast as it did in the first half of the nineteenth century would result in a growth rate of per capita GDP of only about 0.2% per year, well below the range posited by McCusker and Menard. Yet, assuming this rate of growth, and accepting the levels of GDP in 1800, implies that the value of food consumed by free colonists in 1700 would have been lower than the value of the diet consumed by slaves in 1800. Mancall and Weiss argued that this implication seemed implausible and so concluded that likely rates of per capita GDP growth could not have been higher than 0.1% per year and were likely closer to zero.

In subsequent work, Mancall et al. (2004) and Rosenbloom and Weiss (2014) have constructed similar estimates for the colonies and states of the Lower South and the Middle Atlantic regions, respectively. Applying the method of controlled conjectures at a regional level allowed them to incorporate additional, region-specific evidence about agricultural productivity and exports and reinforced the finding that there was little if any growth in GDP per capita during the eighteenth century. Lindert and Williamson (2016b) have also attempted to backcast their estimates of colonial incomes. Their estimates rely in part on the regional estimates of Mancall, Rosenbloom, and Weiss, but the independent evidence they present is consistent with the view that economic growth was quite slow during the eighteenth century.

Wealth Accumulation

One of the richest sources of information about colonial living standards is provided by probate inventories. In one of the first studies to utilize these data, Jones (1980) drew a sample of 899 inventories from randomly selected counties in each region of the colonies in 1774. After adjusting the age distribution to reflect that of the population and reweighting observations to reflect the fact that wealthier decendants had a higher probability of entering probate, she was able to construct estimates of per capita wealth holding by region. These are summarized in Table 1.

The first column of Table 1 shows average total net worth per free capita by region. On this basis, there appears to be a wide gap in wealth accumulation between

Table 1 Probate wealth per capita, by region, 1774

	Per free capita			Per capita
	Net worth	Slaves and servants	Nonhuman	Nonhuman
New England	38.2	0.2	38.0	36.4
Middle colonies	45.8	1.7	44.1	40.2
South	92.7	31.1	61.6	36.4
All colonies	60.2	11.8	48.4	37.4

Source: Alice Hanson Jones (1980), pages 54, 58

colonial regions: residents of the southern colonies had accumulated nearly twice as much wealth as residents of the Middle Atlantic colonies and almost 2.5 times as much as residents of New England. This difference, however, reflects almost entirely the effects of slavery on the distribution of wealth. As the second and third columns show, most slave wealth was concentrated in the South, and regional differences narrow considerably if attention is confined to nonhuman wealth. When the definition of the population is broadened to include the enslaved as well as free residents, as shown in column 4, regional differences in nonhuman wealth per capita are nearly equalized. Thus, one can conclude that while slavery allowed the free residents of the southern colonies to amass a greater concentration of wealth, much of it in the form of property rights in labor, physical capital accumulation was remarkably similar regardless of region.

While Jones was able to provide a detailed cross-sectional snapshot at the end of the colonial period, several other studies have sought to use probate inventories to illuminate trends over time. The results of these studies largely support the picture of a relatively static standard of living. Main and Main (1988) analyzed the economic growth and development of southern New England using a sample of over 16,000 inventories from 1640 to 1774. Figure 3 plots the evolution of estate values and their major components in constant prices. There was, as they concluded, "no doubt that wealth in southern New England was growing in real terms, and the principal category in which that growth occurred was in land and buildings" (Main and Main 1988, p. 36). Indeed, the per capita value of many other categories of wealth was actually declining over time. Thus, while New Englanders cleared land and invested in additional improvements to this land, there was little growth in other markers of material well-being. Lindert and Williamson's (2016b) recent reanalysis

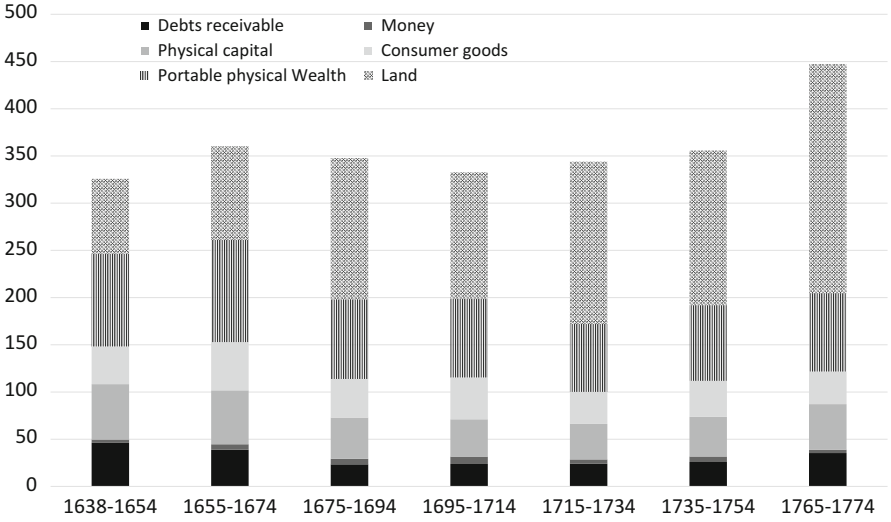


Fig. 3 Probate wealth in southern New England, 1640–1774 (constant sterling). (Source: Main and Main (1988), p. 36)

of these probate data further reinforces the impression of a relatively static economy. Using regression techniques to control for age, location, and occupation, they conclude that only farmers in the later-settled hinterland regions experienced significant gains in wealth over time as they continued to add to their improved land and accumulate additional livestock.

Ball and Walton (1976) made use of inventories from Chester County, Pennsylvania, to measure changes in agricultural productivity over the majority of the eighteenth century. To estimate productivity growth, they constructed indexes of outputs (grain and livestock) and inputs of land and capital from probate inventories and combined these with estimates of labor input from other sources. Setting productivity to 100 in 1714–1731, they found that it had increased to 108 by 1750–1770 but fell back to 105 in 1775–1790 (Table 6, p. 110).

The Colonial Economy

Economic historians have tended to view the colonial economy through one of two lenses. The first approach focuses primarily on the high ratio of land and natural resource to labor that European colonists encountered in North America. According to this "demographic" model, natural resource abundance raised labor productivity, especially in agriculture, contributing to the colonists' high standard of living and removing the demographic constraints that limited population growth in Europe (Smith 1980). Unchecked natural increase combined with migration, both voluntary and forced, to produce the economy's rapid extensive growth. The second view of the colonial economy focuses on the role played by key exports as drivers of economic growth. This "staple exports" thesis emphasizes European demand for tobacco, rice, indigo, and other exports as a major determinant of the size and structure of the colonial economy (McCusker and Menard 1985).

The truth, as is often the case, lies somewhere between these two views. There is no question that land abundance made entry into farming relatively easy, thus encouraging early marriage and high rates of marital fertility. At the same time, mortality was low because abundant food and forest products contributed to a better nourished, better housed, and healthier population, while low population density discouraged the transmission of diseases. Quantitatively, the production of food and fuel for domestic, and mostly local, consumption dominated the economy. Mancall and Weiss (1999), for example, estimated that colonial exports amounted to only about 10% of economic activity across the eighteenth century. Even in a highly export-dependent region, such as the Lower South, Mancall et al. (2008) estimated that foreign exports amounted to only 20–25% of GDP and were declining in importance over time.

Yet, exports were essential to colonial economic survival. The potential contribution of colonial exports to the larger empire was one of the primary motivations for colonization, and export earnings were crucial to the colonies' ability to pay for imports of manufactures and other goods that could not be produced domestically. Differences in export crops also resulted in the emergence of distinctive patterns of

economic organization across the colonies, as reflected in the distribution of wealth noted earlier. Moreover, volatility in international trade played a role in contributing to short-run fluctuations within the colonial economy, although it seems likely that these effects were concentrated in the commercially oriented port cities and did not greatly affect hinterland farmers.

Regional Differentiation in the Colonial Economy

In the Southern colonies, climatic conditions were conducive to the cultivation of crops that found lucrative markets in Europe. In Virginia, Maryland, and parts of coastal North Carolina, tobacco cultivation spread widely, stimulating demand for labor and encouraging immigration. Tobacco did not require major capital investments and was characterized by few economies of scale, leading to a society dominated by small holders working their land with family labor and possibly the assistance of a few servants. By the late 1600s, however, planters in the upper South had begun to import slaves, allowing some producers to expand the scale of production.

Further south, in the low country of South Carolina and Coastal Georgia, early colonists discovered that conditions were favorable for the cultivation of rice. In contrast to tobacco, rice cultivation relied on relatively large capital investments to control irrigation. As a result, rice was grown mainly on relatively large plantations, and colonists in coastal South Carolina and Georgia relied heavily on slave labor to provide an adequate work force. Interestingly, at its founding, Georgia was intended to be a free colony and slavery was prohibited. Settlement proceeded slowly, however, until its founders realized that the prohibition on slavery was discouraging commercial development and allowed the use of enslaved workers. Beyond the narrow coastal region, rice was not a viable crop, and the interior of both colonies was settled primarily by small holders and independent farmers.

The climate of the northern colonies more closely resembled that of northwest Europe, limiting export opportunities. Pennsylvania, New Jersey, and New York nonetheless supported the development of small farms raising livestock and growing wheat and other grains. Flour produced in the region found markets in Southern Europe and the West Indies. Conditions in New England were not as favorable for producing agricultural surpluses, but the region did develop markets supplying food to the West Indies, where intensive sugar cultivation squeezed out local food crops.

Reflecting the greater complexity of regional trading relationships, the northern colonies developed dense and relatively sophisticated merchant communities that helped to organize and finance regional and international trade and provide shipping services. By the late colonial period, Boston, New York, and Philadelphia had become bustling urban centers. The largest, Philadelphia had over 30,000 residents in 1775, while New York had 25,000, and Boston 16,000. In comparison, Charleston, the only significant urban center in the South, had a population of just 12,000.

Free and Unfree Labor in the Colonies

Regional differences in export crop production correlated closely with the use of enslaved labor. Although slavery was legal throughout the colonies, and there were slaves everywhere in British North America, by the 1770s over 90% of the slave population was concentrated in the colonies from Maryland to Georgia. It is tempting to explain the regional distribution of slavery in terms of the distribution of export staples, but the linkage between labor institutions and crops was mostly indirect. While investments in tidal irrigation needed to cultivate rice encouraged large-scale production, tobacco could be cultivated profitably on a small scale and was subject to no economies of scale (Wright 2006).

Because of land abundance, entry into agriculture was relatively easy, and few free men in the colonies were willing to work for wages rather than operating their own farm. As a result, planters who wished to expand the scale of production beyond what could be cultivate with family labor were required to turn to bound labor. As Evsey Domar (1970) has famously observed, it is not possible to simultaneously have free labor, free land, and large-scale production. The availability of unfree labor created the potential for planters to expand production, and the value of export crops provided the means to acquire unfree labor for this purpose.

During the initial tobacco boom in the Chesapeake, the high price of African slaves, which was determined by their productivity in West Indian sugar cultivation, discouraged planters from the use of enslaved labor. Instead, planters who wished to expand production relied largely on indentured servants from Europe. During the seventeenth century, indentured servitude emerged as the primary mechanism to finance the migration of labor to the Americas. Although the high returns to labor in North America made immigration an attractive prospect of English laborers, the high cost of trans-Atlantic passage (In excess of half a year's income) posed a significant obstacle (Grubb 1985). Indenture contracts provided a mechanism by which prospective immigrants could finance their passage.

Indenture contracts entered into with ship captains sailing to North America committed immigrants to work for a specified term (typically 3–4 years) in exchange for passage to the colonies. Once they arrived in America, their contract would be sold to a planter seeking to hire additional labor. At the end of their service, indentured laborers would receive a small payment from their employer and would be able to set out as independent farmers. Other migrants came as "redemptioners," committing to repay ships captains once they arrived in America by selling themselves into servitude to an American master to obtain the funds needed to pay for their passage (Grubb 1986).

Data on immigration in the colonial period are scattered and incomplete, but several scholars have ventured estimates that between half and three quarters of European immigrants arriving in the colonies came as indentured servants or redemptioners. Grubb (1985) found that at the end of the colonial period, close to three-quarters of English immigrants to Pennsylvania and nearly three-fifths of German immigrants arrived as servants. Galenson (1981) has examined the terms of several large samples of indenture contracts and concluded that the length of

service required to repay the cost of passage varied with characteristics likely to be associated with differences in individual productivity. He also concluded that over time, contract terms varied in response to larger supply and demand forces in a manner consistent with the operation of a highly efficient market.

In the 1680s, tobacco planters in the Chesapeake began to shift away from indentured servants toward reliance on enslaved Africans. Galenson (1984) attributes this transition to the falling price of African slaves and improvements in the market for labor in England that made indentured servants more expensive. At first, servants continued to be used to fill positions requiring more skill, but as the number of slaves increased and slave owners expanded their land holdings making it more difficult for new arrivals to gain a foothold as independent farmers, migration to the Chesapeake became less attractive to potential migrants. By the early 1700s, the bulk of indentured servants preferred to migrate to the Middle Atlantic region, and tobacco cultivation became inextricably linked to slave labor. In South Carolina, which was settled only beginning in the 1690s, planters relied from the outset on slave labor to establish commercial rice production. Consistent with the close connection between export-oriented agricultural production and the use of slaves, Mancall et al. (2008) document a strong positive correlation between slave imports to the Lower South by decade and changes in export values.

Figure 4 shows the rapid rise in the slave share of the population in both regions after 1680. Because the mainland colonies were relatively small importers in the larger Atlantic slave market, they faced what was effectively a perfectly elastic supply of slave labor at the world price, which was determined largely by supply and demand conditions in Africa and the West Indies. Consistent with this view, Mancall et al. (2001) show that between 1720 and 1800, long-run movements in

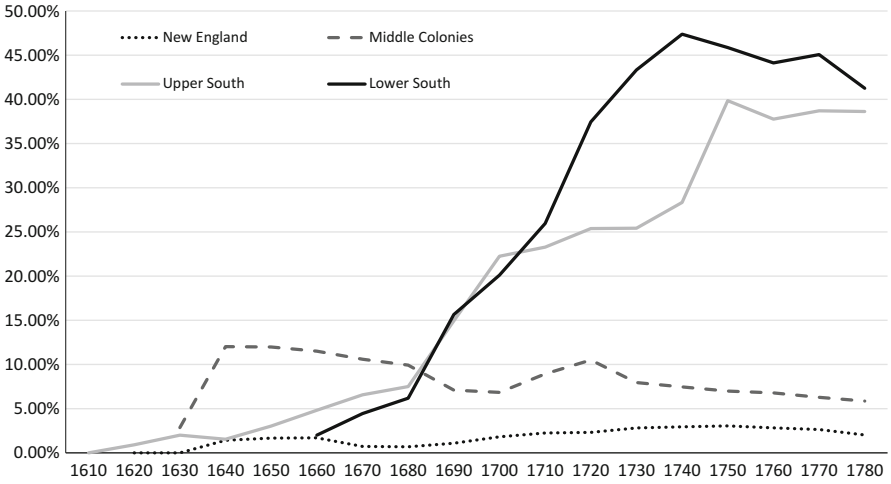


Fig. 4 Black population as percentage of total population, by region, 1610–1780. (Source: Carter et al. (2006), series Eg1–59)

South Carolina slave prices paralleled those in the West Indies and that in the short run, when demand for labor drove up local slave prices, the volume of imports increased and prices ultimately fell.

Institutions and Colonial Economic Development

Institutions and Economic Development

Influenced by North (1990), cliometricians and, more generally, economists have come to see institutions as one of the primary determinants of economic development. According to North, institutions are "the rules of the game" and their means of enforcement. In this definition, they encompass both formal structures of law and governance and informal structures, such as social norms and conventions. For the most part, economists have argued that institutions that provide relatively secure property rights and promote more democratic forms of governance encourage growth by creating conditions conducive to private investment and innovation. If institutions account for differences in economic performance, however, the question becomes why do some societies have institutions better adapted to promote growth than others?

Several recent and influential articles have sought to answer this question by arguing that variation in contemporary institutions, and hence economic development in the Americas, is explained by the interaction between European colonization and the initial conditions that European colonizers encountered in different parts of the world (Engerman and Sokoloff 2012; Acemoglu et al. 2002). The mechanisms identified differ somewhat, but both arguments emphasize a "reversal of fortune," through which conditions that were initially conducive to economic prosperity led to the development of institutions that discouraged subsequent growth, while conditions that were initially less favorable led to development of institutions more conductive to growth.

Engerman and Sokoloff, concentrating on Spanish and British colonization in the Americas, argued that differences in climate, geography, and natural endowments created opportunities for the extraction of precious metals (Latin America) or concentration in production of export staples (the West Indies) that caused them to develop highly unequal societies in which a small European elite exploited a large enslaved population. The resulting legal and political regimes produced high per capita incomes initially but proved poorly adapted to exploiting the opportunities for modern economic growth that emerged during the Industrial Revolution. In contrast, they argue that the absence of a substantial indigenous population in North America, and the fact that soil and climate resulted in limited opportunities to produce crops characterized by major economies of scale, led to the development of relatively egalitarian and representative societies in the mainland colonies. This more egalitarian social structure facilitated the transition in the nineteenth century to modern economic growth.

Acemoglu et al. (2002) adopt a broader perspective, seeking to explain global variations in modern economic performance in terms of the density of population and the corresponding complexity of the societies that European colonizers encountered. Where Europeans encountered (and took over) existing extractive institutions, as in Mexico, India, and Peru, they perpetuated extractive institutions founded on exploitation and inequality. In more sparsely settled and less prosperous areas of the world, such as North America, Australia, and New Zealand, they established institutions of private property and political equality that were better suited to promoting growth in the long run.

Institutions in Colonial America

For both Engerman and Sokoloff and Acemoglu, Johnson, and Robinson, colonial America is reduced in essence to a single data point. Such broad-brush theorizing helps to establish a broader context, but risks missing important local variations and the historical contingencies that underlie the development of American institutions. As we have seen, the economic and social structure of the British colonies that became the United States varied quite substantially, producing a highly unequal distribution of the slave population and wealth holding across the colonies. Yet, all followed similar paths in terms of political structure and long-run economic development.

An important commonality across colonial America was the shared influence of British institutions that the colonists brought with them (Jones 1996). In comparison to other European nations in the seventeenth century, Britain was characterized by a relatively weak monarch, who was obliged to negotiate collectively with quasi-independent local leaders in parliament (Elliot 1992). In comparison, in Spain, where the monarchy was relatively strong, individual notables negotiated directly with the King, resulting in a patchwork of agreements, all dependent on individual relationships (Irigoin and Graf 2008). When early colonists created institutions of local governance, they followed the template provided by the relationship between parliament and the King, establishing bodies to represent their collective interests vis-a-vis the empire, rather than seeking to negotiate individual arrangements.

Another manifestation of the English King's weakness was the limited resources available to devote to the project of American exploration and colonization. Instead of directly financing voyages of exploration, the King encouraged exploration by granting corporate charters or monopolies to private investors who undertook the expense and risk on their own account. Meanwhile, conditions in America encouraged the early development of relatively robust and egalitarian forms of self-government within the colonies.

The first colonial charter for a mainland colony was granted to the Virginia Company in 1606. It gave investors in the company the right to establish a colony and claim the protection of the King. Importantly, the charter established the precedent of private land ownership, specifying that land was to be held in free and common socage, a form of tenure that established secure property rights and made the land fully alienable and inheritable.

Organized as a joint-stock company with the aim of earning profits for its investors, the Virginia Company initially adopted a hierarchical model of control in which the early colonists were housed in communal barracks and farmed the company's land. This top-down organization was poorly adapted, however, to the conditions of land abundance that the early settlers encountered. As Morgan (1971) describes, the company found it nearly impossible to compel the colonists to work on the company's behalf. As a result, the early colonists devoted little effort to food production, and the early years of the colony were characterized by food shortages, epidemic disease, and hardship.

The problem of incentives was resolved when the company opted to grant each settler title to his own parcel of land and allowed them to farm on their own account. Tensions between settlers in America and investors in England and the long delays in communication across the Atlantic further obliged the Company to allow the colonists to establish a representative assembly to make the colony's laws. When the company's charter was revoked in 1624 and Virginia became a crown colony administered directly by the king, the settlers were successful in retaining self-government through the representative assembly, subject to the right of the royally appointed governor's veto.

The Virginia Colony was followed in 1629 by the granting of a charter to the Massachusetts Bay Company, which was also organized as a joint-stock company. Like the Pilgrims who settled at Plymouth in 1620, the Puritan founders of the Massachusetts Bay Company were religious dissenters, and in 1630 the leaders of the Company migrated to New England, taking the charter with them. With the investors in the company present in Massachusetts, tensions between settlers and investors in England were reduced, and the company evolved into a colonial government.

Several more efforts at colonization, including the grant of Maryland to Lord Baltimore, New York to the Duke of York, and Pennsylvania to William Penn, followed a model of proprietorial land grants. The charters granted these proprietors broad powers to govern their new territories and to distribute land within them. However, as in Virginia, land abundance and distance favored the development of local assemblies to represent the settlers and a shift toward local self-governance.

The development of colonial self-government was no doubt facilitated by the limited value of colonial exports. Because of this, the English government largely adopted a policy of benign neglect toward the colonies, exercising quite limited supervision. Political turmoil in England resulting from the English Civil War also diverted attention from the colonies during the middle years of the seventeenth century, allowing traditions of local governance checked only by the veto power of the royally appointed governor to become established.

In sum, while factor endowments and distance played a role in encouraging the colonies to develop mechanisms of self-government, political norms that promoted representation through an elected assembly that represented a unified set of colonial interests to the King were also important. By the late 1600s, when South Carolina was settled, these models were adopted even in an environment characterized from the outset by large-scale plantation agriculture and a heavy reliance on slave labor.

The Colonial Monetary System

At the outset, the colonies were primitive agricultural economies with limited international trading relationships. In the approximately 170 years from the first settlements to American independence, the American economy grew substantially in scale and developed more complex and sophisticated international trading and financial relations with England and other foreign countries. To support the growing scale and complexity of economic transactions, the financial system had to evolve. Among the notable innovations that the colonists introduced was the widespread use of paper money as a medium of exchange (Grubb 2016; Perkins 1988; West 1978).

For the colonists, specie (gold and silver coins) constituted the basis of the monetary system. The colonies did not produce gold or silver, however, and were prohibited from minting their own coins. Specie had to be acquired by selling exports, and the bulk of the specie in use in America came from exports to Spanish America. Exports to England also allowed the colonists to acquire pound sterling credits in the form of bills of exchange drawn on accounts in England. These foreign exchange earnings were used primarily to pay for colonial imports and little specie circulated within the colonies to facilitate domestic commerce.

In lieu of specie, the colonists relied heavily on barter for local exchange. In the Chesapeake transactions were often denominated in weights of tobacco. However, tobacco was not used as a medium of exchange. Rather merchants might advance credit to planters for the purchase of imported items, to be repaid at harvest with the specified quantity of tobacco. Elsewhere book credit accounts helped to facilitate transactions and reduce the need for currency. The colonists regularly complained about the shortage of specie, but as Perkins (1988) observed, the long-run history of prices does not suggest any tendency of prices to fall, as would be expected if the money supply was too small.

While English merchants provided commercial credit to finance international trade, there were no banks or credit institutions to which colonial legislatures could turn when confronted with extraordinary expenses. In 1690, Massachusetts issued £7000 in paper bills of credit to pay wages promised to soldiers who had participated in a military campaign against French Quebec and made the bills legal tender in the payment of colonial taxes. As the bills were returned to the treasury, they were to be retired from circulation. The experiment was a success and was soon followed by most other colonies, mostly in connection with large military expenditures. Details of these monetary emissions varied in terms of maturity, whether interest was paid and whether the bills were made legal tender or not (Grubb 2016). In most cases, however, bills were linked explicitly to future redemption through taxes to be collected or repayment of loans.

With only a few exceptions, the colonies' issuance of these notes did not give rise to inflationary pressures. There is by now a large literature that has analyzed the relationship between note issuance and prices and finds little evidence of any correlation between the series (Weiss 1970; Wicker 1985; Smith 1985; Grubb 2016). As Grubb (2016) has argued, this suggests that while the circulation of bills

of credit may have facilitated exchange by substituting for book credit or other forms of barter, they did not assume the role of currency.

Despite the colonies' history of limited currency depreciation, by the early 1750s, English merchants had begun to express concern that loans they had extended to the colonists might be repaid in depreciated colonial currencies rather than in pound sterling. In 1751, under pressure from these trading interests, Parliament intervened in colonial note issue, passing an act aimed at the New England colonies, restricting the maturity of new issues in Rhode Island, Connecticut, and Massachusetts to 2 years and prohibiting their designation as legal tender in private transactions. This was followed in 1764 by a second act applying to the remaining colonies and prohibiting legal-tender provisions both for private debts and for public obligations as well. A number of colonies, including New York and Pennsylvania, defied these prohibitions, however, and continued to issue new notes (Perkins 1988). Lobbying by colonial representatives was ultimately successful in winning a relaxation of these restrictions, allowing notes to be designated as legal tender for public debts.

The Colonies Within the British Empire

For most of the nearly 170 years between the founding of the Jamestown colony and the American declaration of independence, there was little tension between Britain's mainland North American colonies and the home country. The colonists identified as British subjects and thrived within the context of the expanding Atlantic economy that British policies supported. As Benjamin Franklin observed in 1775, "I never had heard in any Conversation from any Person drunk or sober, the least Expression of a Wish for a Separation, or Hint that such a Thing would be advantageous to America" (Quoted in Taylor 2016, p. 4). From a few small, isolated outposts on the edge of a largely unknown continent, they had grown by the 1750s into a well-established set of communities with a population close to 40% the size of the home country, enjoying relatively high living standards and supporting a complex set of international trading relationships. The context within which this growth occurred is usually termed "mercantilism."

Mercantilism

The key tenet of mercantilism was that national strength was enhanced through an inflow of specie that was achieved by maintaining a positive balance of trade with other nations. Colonies could contribute to these goals by producing specie directly (as was true of Spanish colonies in the Americas), by producing crops that were valued in export markets, or by acting as a source of goods that would otherwise have to be imported from outside the empire. At the same time, colonies could contribute as markets for manufactured goods produced at home.

The Navigation Acts, first passed by Parliament in 1651, were an attempt to achieve these ends by establishing the legal parameters for colonial trading relations.

Their key provisions were (1) vessels registered in foreign countries were excluded from carrying goods between ports within the empire; (2) goods manufactured on the European continent could not be directly imported into the colonies but had instead to pass through England; (3) certain valuable colonial exports, "enumerated goods," could be exported only to ports in Great Britain; and (4) bounties were authorized for highly valued colonial products. Among the enumerated goods were furs, ship masts, rice, indigo, and tobacco. Importantly, ships registered in the colonies were allowed to carry trade within the empire. Although subsequent acts modified bounties and adjusted the list of enumerated goods, the basic outlines remained consistent until American independence (Perkins 1988).

The requirement that major colonial exports pass through England on their way to continental markets and that manufactures be imported from England was the equivalent of imposing a tax on this trade. The resulting price wedge reduced the volume of trade and shifted some of the producer and consumer surplus to the providers of shipping and merchant services. A number of cliometric studies have attempted to estimate the magnitude of these effects to determine whether they played a role in encouraging the movement for independence (Harper 1939; Thomas 1965; Ransom 1968; McClelland 1969). The major difference in these studies arises from different approaches to formulating a counterfactual estimate of how large trade would have been in the absence of the Navigation Acts. In general, the estimates suggest that the cost to the colonists was relatively modest, in the range of 1–3% of annual income. Moreover, this figure needs to be set against the benefits of membership in the empire, which included the protection the British Navy afforded colonial merchants and military protection from hostile natives and other European powers.

Regional Variation Within the Colonies

Despite the common context created by membership in the British Empire, differences in climate and soil created varied economic opportunities in the American colonies that translated into significantly different patterns of economic development across the colonies. With a relatively short growing season and rocky, infertile soil, the New England colonies were poorly endowed to produce agricultural exports. During the initial years of settlement, the colonists participated in the fur trade, acquiring beaver pelts in trade with the indigenous population, until local supplies of beaver were decimated by over hunting. Northern forests also yielded tall trees that could be used as ships masts. Potash, used in making soap and gunpowder and a by-product of clearing fields by burning trees, also found an overseas market. But the value of these exports fell far short of the region's imports of textiles, hardware, and other manufactured goods. To make up this difference, New England merchants developed markets in the West Indies, supplying fish, grain, and livestock to support the intensive cultivation of sugar in these colonies. Boston merchants also earned considerable sums carrying goods within the empire (Walton and Shepherd 1979).

The Middle Atlantic colonies of New York, Pennsylvania, and New Jersey, settled later in the 1600s, emerged as a major source of food exports in the eighteenth century. Possessing better soil, a more favorable climate and excellent natural harbors in New York and Philadelphia, they became major exporters of bread, flour, wheat, and salted beef and pork for both the West Indies and southern Europe. As in New England, earnings from shipping services became a major source of export earnings in the Middle Atlantic region (McCusker and Menard 1985; Perkins 1988).

Although British policy officially discouraged colonial manufacturing enterprises, New England and the Middle Atlantic colonies also developed ship building and iron refining industries that benefitted from access to plentiful supplies of timber. Deforestation in England was a significant constraint on iron manufactures, and imports of pig and bar iron from the colonies were preferred to increasing dependence on European sources. By the late 1700s, the colonies accounted for 15% of world output of iron, ranking third behind Russia and Sweden (Perkins 1988).

The southern colonies conformed much more closely with mercantilist expectations than did the New England and Middle Atlantic colonies. The Chesapeake colonies of Virginia and Maryland, along with parts of North Carolina, became major producers of tobacco, which was in high demand, especially on the European continent. Exports of tobacco were the single largest source of export earnings for the mainland colonies. In the late colonial era, these colonies also began to export wheat and flour in response to rising world prices for food (McCusker and Menard 1985; Egnal 1998). Further South, in South Carolina and Georgia, rice provided a major source of export earnings. The region was also well adapted to the production of indigo, and production of this valuable dyestuff increased in the 1740s when Britain began to offer significant bounties for its production.

Table 2 provides a snapshot of colonial exports by region near the end of the colonial period. Tobacco, grain, rice, fish, and livestock made up close to 75% of the value of colonial exports, with the first two accounting for nearly half of total exports. By virtue of its role in the tobacco trade, the Upper South produced over 40% of all colonial exports, followed by the Lower South, and the Middle Atlantic colonies. New England produced the smallest value of exports and did so with a much more diverse trade than the other regions.

Economics, Politics, and Revolution

After nearly a century and a half of relatively harmonious development, relations between Britain and its North American colonies shifted radically after the conclusion of the Seven Years' War (often referred to in the American context as the French and Indian War) in 1763. Over the next 13 years, the colonists discovered a common identity in opposition to the British Empire, and in 1776 they declared independence. One important consequence of the Seven Years' War had been to largely eliminate France and Spain as rivals for control of North America and as potential allies for the indigenous peoples. As a result of the peace settlement reached in 1763, France had

Table 2 Annual average value of commodity exports from American Colonies, 1768–1772 (pound sterling)

Commodity	New England	Middle Atlantic	Upper South	Lower South	Total
Tobacco			756,128		756,128
Grains, grain products	19,902	379,380	199,485	13,152	611,919
Rice				305,533	305,533
Fish	152,555				152,555
Livestock, beef, pork	89,953	20,033		12,930	122,916
Wood products	65,271	29,348	22,484		117,103
Indigo				111,864	111,864
Whale products	62,103			25,764	87,867
Other	8,552	21,887	39,595	13,904	83,938
Iron		27,669	29,191		56,860
Deerskins				37,093	37,093
Flaxseed		35,956			35,956
Potash	22,399	12,272			34,671
Naval stores				31,709	31,709
Rum	18,766				18,766
Total	439,501	526,545	1,046,883	551,949	2,564,878

Source: McCusker and Menard (1985, pp. 108, 130, 172, 199)

ceded Quebec and most of its North American territories to Britain. Spain, which had also entered the conflict, gave up Florida (which included territories along the Gulf Coast as far west as the Mississippi River) but acquired New Orleans from the French. Now, essentially all of North America east of the Mississippi river was under British dominion.

In the absence of significant European competitors, British attitudes toward colonial territorial expansion shifted dramatically. So long as Britain was in competition with France for control of the North American interior, colonial efforts to secure new land were consistent with imperial goals. But, after 1763, this competition was eliminated, and Britain sought to slow expansion and avoid provoking renewed conflict with the indigenous population. In the colonies, the rapid pace of population growth created an insatiable demand for new land on which to farm; thus obtaining title to western lands became an important source of wealth for leading colonists (Egnal 1980; Egnal and Ernst 1972). British efforts to restrain colonial expansion thus placed them in direct conflict with influential colonists who hoped to profit through development of western lands.

At the same time that British leaders sought to restrain colonial expansion, they were struggling with the challenges of paying the costs of the recently concluded conflict. Defending their North American possessions had been a major expense, and the colonists were lightly taxed relative to citizens at home. It seemed natural that colonists should contribute to paying wartime debts through higher taxes. To the

colonists, however, these measures helped create a perception that Parliament was favoring British interests at colonial expense and provoked a growing resistance movement centered among the mercantile and trading elite in New England and Middle Atlantic port cities (Taylor 2016).

Beginning in 1764 with the Sugar Act, Parliament sought to increase revenues from its colonial possessions. The Sugar Act actually lowered duties on colonial sugar imports from the French West Indies from the prohibitive levels that had previously prevailed, but British officials expected that lower duties would reduce smuggling and increase tax revenues. The Act also provided for increased enforcement, cracking down on a large and illicit trade between New England and sugar producers in the French West Indies. And because colonial courts could not be relied upon to enforce these laws, jurisdiction was moved to British Admiralty courts.

The Sugar Act was followed in 1765 by passage of the Stamp Act, which imposed a tax on a broad range of legal documents and newspapers. To colonists, this effort to tax domestic commerce appeared to be a significant departure from previous acts that had focused on external trade. And it evoked a strong negative reaction from influential members of the colonial elite. Colonists threatened or intimidated the appointed tax collectors and forced many of them to resign from their positions or to agree to not enforce the tax. They also sought to organize a collective response, calling on colonial governments to send representatives to a Stamp Act Congress in New York. As the first such meeting of representatives from the different colonies, the Congress was an important step in the emergence of a sense of national identity. The boycott of British goods the Congress organized led British merchants to join the colonists in urging Parliament to rescind the tax. In 1766, Parliament backed down and repealed the tax.

Britain's need for revenue remained, however, and in 1767 the Townshend Act imposed a new set of duties on the colonists. Believing that taxes on external trade would not meet with resistance, the Townshend Act imposed duties on imports of glass, paint lead, paper, and tea. The tax was coupled, however, with plans to use part of the revenue it generated to pay the salaries of colonial governors, who until then were supported by locally collected taxes, a change that would have made them less dependent on colonial legislatures. The colonists objected to this interference, however, and newly emboldened by the success of their resistance to the Stamp Act, they launched another boycott. As the volume of business lost by British merchants mounted, so did pressure to repeal the Townshend Act taxes. In 1770 Parliament dropped all of the taxes except that on tea.

The final provocation came in 1773 in the form of the Tea Act. Seeking to aid the East India Company, which found itself in significant financial difficulties, Parliament granted the company a monopoly on the sale of tea in the American market. Taxes on tea were actually adjusted, so that the price colonists would pay fell. Yet, local merchants, angered by their exclusion from a lucrative trade and seeing Parliament favoring a British company at their expense, organized protests. In Boston on December 16, 1773, a group of colonists boarded several vessels carrying East India tea and dumped it in Boston Harbor. In response to the so-called Boston Tea Party, Britain ordered the complete closure of Boston Harbor and dispatched a

large military force to enforce this action. Responding to these actions, the colonists convened the first Continental Congress in Philadelphia in 1774, and by early 1775, the tensions had devolved into an armed conflict precipitated by the British march on Lexington and Concord.

After the Revolution: American Independence

Looking backward it is easy to view the success of the American Revolution as inevitable. It was, however, anything but that. Residents of the colonies were deeply divided about independence: loyalists, who favored remaining part of the empire, were probably almost as numerous as revolutionaries, and a large part of the rural population remained uncommitted to either side. The Revolution thus entailed significant internal conflicts. Moreover, the revolutionaries confronted practical challenges in taking on the much larger and better-supported British Army (Taylor 2016).

The armed conflict dragged on for 7 years, ending only in 1782. The colonists faced a more numerous, better equipped and better trained military force; but the British had to contend with the logistical challenges created by distance and slow communication. For the most part, British forces were content to occupy a few major ports and exert pressure by attempting to blockade trade at sea. Unable to defeat the British, the colonists were ultimately successful in outlasting them as the costs of the war mounted.

For colonists outside contested areas, the impact of the conflict was relatively light. For those closer to the fighting, the effects were mixed. On the one hand, increased demand for food and supplies caused by British occupation may have raised incomes. On the other hand, some colonists saw crops seized or destroyed, and in the South, British forces confiscated slaves and encouraged others to defect.

Most accounts suggest that independence was a negative shock to the colonial economy but have differed in their assessment of the magnitude of the effect. After the disruption of the war, the newly independent country was largely excluded from the trading networks it had participated in prior to independence, and under the Articles of Confederation, Congress lacked authority to establish tariffs and consequently lacked the bargaining leverage necessary to negotiate access to European markets. Some of these difficulties were removed with the ratification of the Constitution in 1787, and the outbreak of hostilities between the French and British after 1793 created new trading opportunities for American merchants.

Bjork (1964) has argued that the effects of the disruption in international trade were short-lived and relatively mild, as resources were diverted toward westward expansion and import-competing production. Lindert and Williamson (2013), on the other hand, offer the most pessimistic estimates of the decline in incomes after the Revolution. Weiss (2017), however, offers reasons to be skeptical of their estimates. Expressed in prices of 1840, Lindert and Williamson (2013) estimate that per capita income fell from $74 in 1774 to $59 in 1800, a drop of 20%. Since incomes are generally thought to have begun to recover after the early 1790s, this suggests that

the trough in incomes must have been even larger. Mancall and Weiss (1999) argued that between 1770 and 1800, incomes were relatively flat but inferred that there must have been some decline in incomes between 1770 and 1790 that was erased by the subsequent recovery of the 1790s. Rosenbloom and Weiss (2014) estimate that in the Middle Atlantic, per capita income fell from $78.70 to $65.50 in 1791, before recovering to $78 by 1800 (all quantities in constant 1840 prices).

Conclusion

Between 1607 and 1776, the 13 British North American colonies that became the United States were transformed from small isolated outposts of European settlement to a thriving economy with a population almost 40% the size of Britain. After overcoming the initial hardships of establishing the colonies, per capita incomes grew slowly (if at all). But the remarkable feature of this history is that the colonies were able to sustain rapid extensive growth for nearly two centuries without a reduction in living standards.

Writ large, the story is fundamentally one of resource abundance relative to labor and capital. When North American land was combined with European agricultural technologies and European institutions of private property, resource abundance created high returns to mobile resources. The result was a flow of trans-Atlantic migration and investment. This response required a variety of institutional innovations. To facilitate the movement of European laborers, indentured servitude was developed as a mechanism to finance profitable migration. Yet the resulting labor response was not sufficient to meet the labor needs of the colonies, and colonists resorted as well to the involuntary transportation of thousands of enslaved Africans to perform less desirable work in less attractive locations. Among both European- and African-American populations, rates of natural increase were relatively high. Abundant food and fuel contributed to a relatively healthy population and allowed earlier marriage and high rates of marital fertility. As a result, populations doubled roughly every 20–25 years.

Of course, European and African settlers did not arrive in an empty land. Their success coincided with the displacement of aboriginal peoples. Early European settlers encountered natives whose populations had already been disrupted by European diseases spread by early contacts with fishermen and explorers, thus facilitating settlement. The trade in furs (beaver in the North and deerskins in the South) was an important component of European economies in the early years of settlement, and both trade and conflict with native populations remained an important element throughout the colonial period.

In much the same way that institutions adapted to promote migration, the colonists adapted mechanisms of self-governance to their new conditions. Abundant land made it difficult for the early Virginia Company to maintain top-down discipline in its settlement. Only when land was distributed to the colonists and they were allowed to retain the profits from their effort did production begin to grow. Dealing with distant masters, separated by a month-long trans-Atlantic voyage, the colonists

asserted their rights to establish local governments empowered to make collective choices for most local decisions. Local assemblies developed strong traditions that provided a foundation for local rule when the colonists declared independence.

Nonetheless, for most of the colonial period, there was little thought of seeking independence. The colonies thrived within the imperial trading relationships that the British created. Under the Navigation Acts, American merchants enjoyed the same access to this trade as British merchants, and the protection of the British Navy on the seas, and the British military on land, provided important benefits to the colonists. After 1763, however, the dynamics of colonial relations shifted. Having largely eliminated the competition of French and Spanish colonies for control of North America, Britain wished to restrain the spread of American population and avoid provoking conflicts with native populations. The colonists saw opportunities to expand. At the same time, new taxes and new policies led colonists to question whether the British government was committed to their interests. The result was a rapidly growing movement of resistance that culminated in a formal declaration of independence in 1776.

References

Acemoglu D, Johnson S, Robinson JA (2002) Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q J Econ 117(4):1231–1294

Allen R, Murphy T, Schneider E (2012) The colonial origins of the divergence in the Americas: a labor market approach. J Econ Hist 72(4):863–894

Ball D, Walton GM (1976) Agricultural productivity change in eighteenth century Pennsylvania. J Econ Hist 36:102–117

Bjork GC (1964) The weaning of the American economy: independence, market changes, and economic development. J Econ Hist 24(4):541–560

Bolt J, van Zanden JL (2013) The Maddison-project http://www.ggdc.net/maddison/maddison-project/home.htm, 2013 version. Accessed 19 Oct 2017

Domar E (1970) The causes of slavery and serfdom: a hypothesis. J Econ Hist 30:18–32

Egnal M (1980) The origins of the revolution in Virginia: a reinterpretation. William Mary Q 37:401–428

Egnal M (1998) New World economies: the growth of the thirteen colonies and early Canada. Oxford University Press, New York

Egnal M, Ernst J (1972) An economic interpretation of the American revolution. William Mary Q 29(1):3–32

Elliot JH (1992) A Europe of composite monarchies. Past Present 137:48–81

Engerman SL, Sokoloff KL (2012) Factor endowments and institutions. In: Engerman SL, Sokoloff KL, Haber S (eds) Economic development in the Americas since 1500: endowments and institutions. NBER series on long term factors in economic development. Cambridge University Press, Cambridge, UK

Galenson DW (1981) White servitude in colonial America: an economic analysis. Cambridge University Press, Cambridge

Galenson DW (1984) The rise and fall of indentured servitude in the Americas: an economic analysis. J Econ Hist 44(1):1–26

Galenson DW (1996) The settlement and growth of the colonies: population, labor, and economic development. In: Engerman SL, Gallman RE (eds) Cambridge economic history of the United States, Volume I: the colonial era. Cambridge University Press, Cambridge, UK

Grubb F (1985) The incidence of servitude in trans-Atlantic migration, 1771–1804. Explor Econ Hist 22(3):316–339
Grubb F (1986) Redemptioner immigration to Pennsylvania: evidence on contract choice and profitability. J Econ Hist 46(2):407–418
Grubb F (2016) Colonial paper money and the quantity theory of money: an extension. NBER working paper 22192. NBER, Cambridge, MA
Harper LA (1939) The effect of the navigation Acs on the thirteen colonies. In: Morriss RB (ed) The era of the American evolution: studies inscribed to Evarts Boutell Greene. Columbia University Press, New York
Irigoin A, Graf R (2008) Bargaining for absolutism: a Spanish path to nation-state and empire building. Hisp Am Hist Rev 88:173–209
Jones AH (1980) Wealth of a nation to be: the American colonies on the eve of the revolution. Columbia University Press, New York
Jones EL (1996) The European background. In: Engerman SL, Gallman RE (eds) Cambridge economic history of the United States, Volume I: the colonial era. Cambridge University Press, Cambridge, UK
Lindert PH, Williamson JG (2013) American incomes before and after the revolution. J Econ Hist 73(3):725–765
Lindert PH, Williamson JG (2016a) Unequal gains: American growth and inequality since 1700. Princeton University Press, Princeton
Lindert PH, Williamson JG (2016b) American colonial incomes, 1650–1774. Econ Hist Rev 69(1):54–77
Main GL, Main JT (1988) Economic growth and the standard of living in southern New England, 1640–1774. J Econ Hist 48(3):27–46
Mancall PC, Weiss T (1999) Was economic growth likely in colonial British North America? J Econ Hist 59(1):17–40
Mancall PC, Rosenbloom JL, Weiss T (2001) Slave prices and the South Carolina economy, 1722 to 1800. J Econ Hist 61(3):616–639
Mancall PC, Rosenbloom JL, Weiss T (2004) Conjectural estimate of economic growth in the Lower South, 1720 to 1800. In: Guinnane TW, Sundstrom W, Whatley W (eds) History matters: economic growth, technology and population, essays in Honor of Paul A. David. Stanford University Press, Stanford
Mancall PC, Rosenbloom JL, Weiss T (2008) Exports and the economy of the Lower South region, 1720–1770. Res Econ Hist 25:1–68
McClelland PD (1969) The cost to America of British imperial policy. Am Econ Rev Pap Proc 59:382–385
McCusker J, Menard R (1985) The economy of British America, 1607–1789. University of North Carolina Press, Chapel Hill
Morgan ES (1971) The labor problem at Jamestown, 1607–18. Am Hist Rev 76(3):595–611
North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, New York
Perkins EJ (1988) The economy of colonial America, 2nd edn. Columbia University Press, New York
Ransom RL (1968) British policy and colonial growth: some implications of the burden from the navigation acts. J Econ Hist 28(3):427–435
Rosenbloom JL, Weiss T (2014) Economic growth in the mid Atlantic region: conjectural estimates for 1720 to 1800. Explor Econ Hist 51(1):41–59
Smith DS (1980) A Malthusian-Frontier interpretation of United States demographic history before c. 1815. In: Borah W et al (eds) Urbanization in the Americas: the background in comparative perspective. National Museum of Man, Ottowa
Smith BD (1985) American colonial monetary regimes: the failure of the quantity theory and some evidence in favor of an alternate view. Can J Econ 18(3):531–565
Taylor A (2016) American revolutions: a continental history, 1750–1804. W. W. Norton, New York

Thomas RP (1965) A quantitative approach to the study of the effects of British imperial policy upon colonial welfare: some preliminary findings. J Econ Hist 25:615–638
Ubelaker DH (1988) North American Indian population size, A.D. 1500 to 1985. Am J Phys Anthropol 77:289–294
Walton GM, Shepherd JF (1979) The economic rise of early America. Cambridge University Press, New York
Weiss R (1970) The issue of paper money in the American colonies, 1720–1774. J Econ Hist 30:770–785
Weiss T (2017) Review of Unequal gains: American growth and inequality since 1700, by Peter H. Lindert and Jeffry G. Williamson. J Econ Hist 77(3):952–954
West RC (1978) Money in the colonial American economy. Econ Inq 16:1–15
Wicker E (1985) Colonial monetary standards contrasted: evidence from the seven years' war. J Econ Hist 45:869–884
Wright G (2006) Slavery and American economic development. Louisiana State University Press, Baton Rouge

Property Rights to Frontier Land and Minerals: US Exceptionalism

Gary D. Libecap

Contents

Abstract

Property rights are the most fundamental institution in any society. They determine who has decision-making authority over assets and who bears the costs and benefits of those decisions. They assign ownership, wealth, political influence, and social standing. They make markets possible, define timelines, and provide incentives for investment, innovation, and trade. They mitigate the losses of open access and provide the basis for long-term economic growth. Economists and

Very helpful comments and direction were provided by Daron Acemoglu, Susan Carter, Robert Ellickson, Eric Edwards, Stanley Engerman, Richard Epstein, Peter Lindert, Deirdre McCloskey, Larry Neal, Claire Priest, Richard Sutch, Tom Weiss, and Gavin Wright. Excellent research assistance was provided by Chester Lindley

G. D. Libecap (✉)
National Bureau of Economic Research, University of California, Santa Barbara, CA, USA

Hoover Institution, Stanford University, Stanford, CA, USA
e-mail: glibecap@bren.ucsb.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_46

economic historians have long recognized the importance of secure property rights for economic outcomes. Other political economy, philosophy, and historical and legal literatures emphasize different, but critical, attributes based on *how* property rights are allocated and to whom. The linkages among the social, political, and economic effects are examined here with respect to US and Latin American frontier land and minerals. Property rights were sharply different across the two frontiers with apparent long-term consequences for economic growth, innovation, wealth distribution, private investment in public goods, as well as social and political stability. The distinct assignment of property rights to land and minerals is likely a basis for long-term US exceptionalism in economic performance, individualism, mobility, and optimism. The mechanisms through which property rights to land in a frontier society affect outcomes in a contemporary, highly urban one are complex. Because property rights to land were broadly distributed, Americans could participate in capital markets using land as collateral. This ability shaped opinions regarding markets, capitalism, and individual opportunity. In the twenty-first century, these critical attributes may be eroding, inviting more analysis from economists and economic historians.

Keywords
Property rights · Frontiers · Incentives

If a man owns a little property, that property is him.. . .it is part of him. . ..in some ways he's bigger because he owns it. John Steinbeck, *The Grapes of Wrath* (1939, 1976, 48)

Introduction

The above quotation points to attributes of property rights that are more individualistic, social, and political than are those emphasized in most economic discussions. Nevertheless, there are direct linkages across these characteristics, and emphasizing the relationships is the focus of this chapter. Attention is directed to frontiers in North and South America where property rights were distributed to land and minerals in very dissimilar ways, seemingly leading to starkly different consequences – economic, political, and social. These varying frontier experiences appear to have had long-term consequences. Frontiers provide a natural experiment because, by definition, they are areas where new rights to resources are emerging. The discussion is based on the existing literature and does not provide tests of hypotheses regarding property rights and various outcomes. Causality between property rights structures and observed variation in immigration patterns, middle-class development, innovation, social and political cohesion, individualism, reduced reliance upon the state, private investment in public goods, and long-term economic growth seems, however, to be supported by available research.

Property rights are the most fundamental institution in any economy and society. The economics and economic history literatures explore the role of secure property rights in mitigating the losses of the common pool, in promoting markets and

exchange, and in encouraging investment. These facilitate long-term economic growth and welfare. Older political economy, philosophy, and history literatures, as well as growing legal scholarship, point to widespread ownership of land and related resources in molding social, economic, and political relationships and the role of individuals relative to the state. In those literatures, the existence of secure private property rights leads to more independent, self-reliant, and individualistic citizens. They innovate, are politically conservative, and invest in local public goods. They rely on rents they discover and generate. The state is less important than the market, and the economy in turn is less centralized, more atomistic, market-based, and supportive of entrepreneurship.

For the USA, the frontier is defined according to the US Census as an unsettled region where population density was less than two persons/mile^2 (U.S. Census Bureau 2012). In Latin America, frontier regions were new to Europeans, but often had denser native occupation, and there were multiple frontiers. Nevertheless, for land and minerals, the definition of the frontier is used here for European settlers in the region. The settlement of temperate North America by European migrants, molded in part by the English common law of property and contract, along with the actions of colonial and US courts and legislatures, granted access to and ultimate ownership of unimaginable riches to common individuals in a piecemeal fashion. In Latin America, land and mineral ownership was retained by the sovereign with use rights to large tracts of land granted to political and economic elites. Mineral rents largely were reserved by the crown. In the postcolonial period, hierarchical, centralized control continued, although individuals could obtain title to their land. Minerals remained owned by the state. These differences in the property rights to land and minerals allocated to individuals by the state between North and South America appear to have had long-lasting, important economic, political, and social effects that are collected from the relevant literatures and described below. The outcomes underscore the perhaps underappreciated, far-reaching, and enduring impact of the nature of property rights to fundamental resources that exist in any society. They go far in offering an explanation for contemporary differences observed across nations and across times in development, equity, opportunity, entrepreneurship, and social and political stability.

The Economic Institutions of Property Rights

Property rights critically shape economic behavior by fixing incentives for resource use, investment, exchange, and inheritance. They set time frames and define the decision-makers for such actions. They determine flows of associated benefits and costs and designate who will receive or bear them. Externalities arise from incomplete property rights, when decision-makers do not internalize the full benefits and costs of their actions. Depending on the size of the imprecision in definition of property rights, incentives are distorted, leading to losses in potential resource value and welfare. Property rights can be informal (implicitly recognized) or formal (legally documented) and can range from state ownership to group rights to private

property rights. In all cases to be stable, they must be socially sanctioned and often are enshrined in an established law.

Property rights can exist with almost any imaginable array of attributes. They may (a) be held by a single party or be divided with one party having use rights and another actual ownership; (b) be permanent or short-term; (c) have comprehensive or restricted authority over use, exchange, investment, and assignment to heirs; (d) completely direct costs and benefits to rights holders or split costs and benefits among multiple entities, including users, sovereign rulers, politicians, and bureaucrats; (e) be well-defined or imprecise; and (f) be secure or insecure. This assortment of possible attributes leads to a similar multiplicity of economic outcomes for the same resources and decision-makers.

Economists and economic historians have long recognized the critical role of property rights in determining economic performance. North and Thomas (1973) emphasize the emergence of different freehold rights between England and Holland and France and Spain as a key source in differential patterns of economic growth. When reasonably well-defined and durable, private property rights critically contribute to economic growth (Davis and North 1971; North 1981, 1990; Acemoglu et al. 2001, 2005; Mehlum et al. 2006; Rodrik 2008; Dixit 2009; Besley and Ghatak 2010; North et al. 2009; Acemoglu and Robinson 2012; Alston et al. 2018). They facilitate greater investment when returns are uncertain or delayed (Besley 1995; Jacoby et al. 2002; Galiani and Schargrodsky 2010; Hornbeck 2010). They allow for the development of markets (Greif et al. 1994; Barzel 1997; Dixit 2009; Edwards and Ogilvie 2012). And finally, they reduce rent dissipation associated with common-pool resources (Gordon 1954; Scott 1955; Cheung 1970; Johnson and Libecap 1982; Wiggins and Libecap 1985; Grafton et al. 2000; Wilen 2005; Costello et al. 2008).

The principal long-term economic benefits of property, as an institution, arise from private property rights (Merrill and Smith 2010). Although group management of resource access and use has been effective in overcoming the losses of the common pool (Ostrom 1990), the conditions for successful collective action may be quite limited (Cox et al. 2010). Moreover, communal rights may not facilitate markets and asset trade outside the group nor risky, disruptive innovation within it. Finally, neither theory nor empirical evidence indicate success with state-owned rights or socialism in promoting long-term economic growth or in mitigating open-access losses (Barro 1991; Grafton et al. 2000; Costello et al. 2008).

The dominant performance of private property rights, relative to others, lies in the more complete alignment of private and social benefits and costs, lower transaction costs in decisions (initial assignment of rights may entail high transaction costs, Libecap 2008), market development, and incentives for value or rent creation (Ellickson 1993; Allen 2011). With state ownership or control, value may be lost, even if externalities are addressed. While private property rights assign ownership to individuals, group or state property rights *separate* ownership from actual decision-making, potentially generating distortions with possibly high welfare losses. There can be externalities associated with private ownership, and in that case, one question is why the property rights are incomplete in the first place, as compared to traditional

calls for state regulation or taxation (Pigou 1920; Coase 1960; Dahlman 1979). Externalities are far less likely for resources, such as land and minerals, where property rights can be defined and enforced at relatively low transaction costs. For other resources, such as air, water, or fish stocks, property rights are more difficult to define and enforce, and government ownership or regulation of access and use is more prevalent (Libecap 2008). For some resources, such as forestland and rangeland in the USA, political demand by advocacy groups, beginning in the late nineteenth century, led to their retention by the state, rather than assignment to private parties (Libecap 2007). The underlying arguments for retention by the state were not fundamentally due to an inability to define property rights effectively (Libecap 1981, 2007). The key problem for state regulation or ownership in either of these cases is that neither politicians nor bureaucrats are full residual claimants to the benefits and costs of their actions in the way that private owners are, so that incentives and outcomes differ, often creating other, and perhaps costlier, externalities (Libecap 2016).

Social and Political Institutions of Property Rights: Pre-frontier

Historically in Europe, ownership of land and all natural resources lays with the Creator, as represented on earth by the sovereign. The landed nobility had use rights that were held at the pleasure of the sovereign, and much of English and European history involved conflicts between the crown and the nobility over the extent and nature of those rights, their security, and taxation. Those who actually worked the land, thousands of serfs, peasants, and tenant farmers, had little or no authority over it and captured few of the benefits, with most rents taxed away by their feudal masters and passed on in part to the crown. Those who used the land were bound to it. This very centralized setting was static so as not to upset key relationships, and there was little incentive among tillers of the soil and grazers to innovate, and they had little political role in the society.

Political economists and philosophers during the European Enlightenment, including Adam Smith, John Locke, Jeremy Bentham, J.J. Rousseau, John Stuart Mill, David Ricardo, Edward Wakefield, and Robert Torrens (Winch 1965; Ellickson 1993; Linklater 2013; Priest 2019), debated the role of individuals in society, their potential for advancement, their relationship with the state, and the critical impact of widespread private ownership of land for advancing individual and resource potentials. The implications of land ownership as a threat for an authoritarian state also were clearly understood by Marx, Lenin, Stalin, and Mao Zedong. Land as a fundamental resource was key, and its ownership had contagious effects on the entire economic, social, and political orders in the short and long run.

The colonization of western hemisphere frontiers during the sixteenth and nineteenth centuries by England, Holland, France, Spain, and Portugal was molded by very different views of land and minerals distribution and ownership. In the Spanish, Portuguese, and French colonies, the process was controlled centrally by the crown. There was little emphasis on large-scale emigration with land being granted in large

tracts to political elites. Ownership remained with the crown, and those who received land grants held them at its pleasure. For English North America, the nature and distribution of property rights were in sharp contrast. Individuals, not the crown, were the ultimate owners of land, and for the most part, it was allocated in small plots. Vast numbers of immigrants were attracted by the opportunity to secure land, and their ability to own it had profound consequences in the development of English colonies and, subsequently, the USA.

The Magna Carta of 1215; the Glorious Revolution of 1688; land enclosures that broke up communal holdings, particularly in the eighteenth and nineteenth centuries; and the English agricultural and industrial revolutions ended the ultimate ownership of land by the crown as the representative of God on earth. Comparatively static, feudal, and communal obligations were broken. More materialistic economic and democratic political objectives became ascendant. The legal basis for alienable, private property in land became part of the common law. Ordinary individuals could own land and enjoy the benefits of using, investing, and trading it. William Blackstone commented in 1766 on the implications: "There is nothing which so generally strikes the imagination, and engages the affections of mankind, as the right of property; or that sole and despotic dominion which one man claims and exercises over the external things of the world, in total exclusion of the right of any other individual in the universe . . .." (quoted in Ellickson 1993, 1317). English colonization and migration to North America were driven by these ideals (Ely 2008). Those who migrated to and occupied frontier land eventually held it in fee simple as independent owners and not as a dependent peasantry (Story 1858).

Property Rights to Land on the US Frontier

Claire Priest (2019) summarizes much of the early literature and key elements of colonial and early federal land law. She argues that the colonists brought with them English laws, customs, and legal institutions and then modified them through the statutory enactments of local representative assemblies and rulings of common law courts. Gradually, colonial and early US property law became quite distinct from that in England, fundamentally transforming the economic, political, and social structure of the country (Priest 2019). Property rights in land became a liquid source of wealth, to be bought and sold and used to obtain credit. Because land was the most basic resource, its widespread ownership became the catalyst for colonial economic and political development. The ownership of property made individuals as special stakeholders in the society and dispersed political and economic power from elites in a manner that had not occurred in England. The easy circulation of land in the market facilitated extensive property ownership, undermining privileged inheritance and inalienability. Dynamic, open land markets became an essential ingredient for the credit system and its ability to support the growth of a middle class as well as to spur investment and innovation throughout the economy (Priest 2019).

There were many commentaries on the benefits of land ownership and exchange. Benjamin Franklin saw it as the way for ordinary persons to improve their position in

life and that of their children (Franklin 1751, quoted in McCulloch 1845). Thomas Jefferson saw a nation of numerous, small freeholders not only as good economics but good politics. The seemingly endless abundance of frontier land provided the perfect opportunity to create a society composed of small, independent, freeholding farmers that could support a republican form of government. Such citizens with an attachment to the land and to the country had virtue and a common interest in political stability and social cooperation. He notably stated that: "The earth is given as a common stock for man to labor and live on... The small landholders are the most precious part of a state" (quoted in Katz 1976, 480). Alexis de Tocqueville observed that being freeholders changed the way in which Americans thought of themselves and the political structure: "Why, in a quintessentially democratic country like America, does one hear no complaints about property in general such as those that often resound through Europe? Needless to say, it is because there are no proletarians in America. Since everyone has property of his own to defend, everyone recognizes property rights as a matter of principle" (de Tocqueville 1835, quoted in Goldhammer 2004, 273).

Later in the nineteenth century, as the frontier neared its end, the US Public Lands Commission endorsed the small-farm, homestead principle: "The maxim that He who tills the soil should own the soil is accepted as a fundamental principle of political economy... Small holdings distributed severally among the tillers of the soil is believed to be a fundamental condition for the prosperity and happiness of an agricultural population" (US Public Lands Commission 1880, xxii). Frederick Jackson Turner in 1893 in his famous thesis about the role of the frontier in US political and social development went further, claiming that America ultimately was shaped by small-farm frontier settlement as the underpinning for democracy, an independent citizenry, and generalized economic well-being (Turner 1893). This is the notion of US exceptionalism and its dependence on frontier resource ownership emphasized here.

Colonial Property Rights to Land

The English Crown granted colonial charters that conveyed land and lawmaking authority. Some colonies began as trading companies, such as the early Virginia, Plymouth, and Massachusetts Bay Companies, and the charters were to the owners of the corporation. Other colonies were proprietorships based on grants from the crown to an individual or a group of individuals, such as those to William Penn and Lord Baltimore, creating Pennsylvania and Maryland. A third type was the royal colony directly ruled by the crown, including New Hampshire, New York, New Jersey, Virginia, North Carolina, South Carolina, and Georgia. In all cases, a governor or proprietor served as the top official, and assemblies of elected representatives were authorized to enact legislation (Priest 2019).

Land sales were seen by the crown, proprietors, shareholders, and others as a major source of revenue, and accordingly, there was a need to attract immigrants to North America who would create small farms and cultivate the land. Large land

grants, in general, were not consistent with the policy. Especially in the Middle Atlantic and Southern colonies, a headright of 50 acres or more was given to those who would cover the transport costs of any immigrant. In repayment, the immigrant served an indenture period, 5 to 7 years, whereby they served their sponsor. Indenture contracts were supplemented later in the colonial period by redemptioners, who borrowed for their travel costs and were released from their indenture commitments upon paying off the loan (Ford 1910, 416; Grubb 1986; Abramitsky and Braggion 2006). The headright policy also encouraged the importation of slaves, but indentured servants had a future as freeholders with voting rights within the colonies. The headright/indentured servant system may have accounted for 50–66% of white male immigration to the American colonies between 1630 and 1776, or 300,000–400,000 people. Slave importation, primarily in the eighteenth century, was just over 255,000, largely to Bermuda, Barbados, and the southern US colonies (Priest 2019).

High-level colonial administrators were recruited with the promise of landed estates, and they expected to profit from rising land values and sales, stimulated by rapid new settlement by small holders. To quiet titles, attract settlers, and support land exchange, colonial administrators promised surveys of small parcels. Although most property boundaries followed natural terrain (metes and bounds), more systematic, rectangular survey of parcels was implemented in flat areas. Rectangular survey facilitated subdivision and sale (Ford 1910, 329–356; Libecap et al. 2011).

The extensive availability of fertile land to small holders, who could secure and cultivate freeholds, not only invited vast immigration but generated an egalitarian society with high levels of real per capita income. By 1751 the British North American colonies may have had 1 million inhabitants, compared to 52,000 or so in New France and a generally small number of immigrants to the Spanish and Portuguese colonies of South America (Linklater 2013). Lindert and Williamson (2013, 2014a, b) report that in 1774 the American colonies had the most equal distribution of income in the western world and per capita purchasing of income exceeded that in Great Britain. The process of distributing land on the frontier took time to sort out during the colonial period. Much of the political instability observed at the time, aside from opposition to English rule, came from frontier settlers against colonial proprietors and other administrators over issues of taxation of local production, protection against natives, and overall access to land (Bacon's Rebellion, 1676; Culpeper's Rebellion, 1677; Cary's Rebellion, 1711; Shay's Rebellion, 1786; and the Whiskey Rebellion, 1791–1794). During the postcolonial period, there was seemingly far less political volatility on the frontier. As described below, policies to grant land to migrants at low cost and to support their property rights and land markets became routine. The political coalition of small holders and would-be small holders, who also were land speculators and frontier territorial politicians who wanted their regions to have dense settlement so as to qualify for statehood, was a formidable one in Congress. At the same time, the federal government had no strong reason to hold on to frontier land. Until the rise of the conservation movement, there was no advocacy for maintaining federal ownership. For several reasons, land

policies promoted the distribution of frontier land to private claimants as quickly and at as low a cost as possible. As discussed below, political motives for land transfers were very different in Latin America.

Federal Property Rights Policies for Land

Following the end of the French and Indian War in 1763 and the Revolutionary War in 1783, frontier migrants began moving in large numbers beyond the Appalachian Mountains to the comparatively flat lands in the Ohio River Valley and elsewhere. Land in the vast new territory was distributed through direct sales from the federal government and from the issuance of military warrants, redeemable for small parcels, to compensate Revolutionary War soldiers (Ford 1910). Military warrants were bought and sold and were a major means of securing access to land by their ultimate holders. The Federal Land Ordinance of May 20, 1785, called for the survey of all lands ceded by the states to the federal government and all additional lands acquired through purchase from native tribes. It created the Public Land Survey System (PLSS) of grids of square townships of 6 miles square and 23,040 acres each, aligned along latitude and longitude. Each township was subdivided into 36 sections of land that could be further subdivided into half and quarter sections for purchase and sale. The survey made land a commodity with clearly defined parcel boundaries that were easily addressable (Libecap and Lueck 2011; Libecap et al. 2011). Libecap and Lueck (2011) estimate that the rectangular survey raised land values by 23% relative to the baseline alternative of metes and bounds, which invited boundary disputes and created irregular parcel sizes and shapes that hindered market exchange. The General Land Office was created in 1812 to administer and extend the survey across the continent and to distribute additional federal lands under land laws enacted by Congress (Table 1 below). The rectangular survey reduced gaps between properties and promoted dense, rapid settlement of the frontier. The federal government offered land at fixed prices, often $1.25/acre to raise revenue. This revenue objective ultimately was undermined by the inability to police squatting and occupancy and growing political demand for free land.

Figure 1 shows the movement of the frontier across the continent based on population density. In 100 years, the frontier went from the Atlantic seaboard to its announced end by the Census Bureau in 1890. As noted above, the Bureau defined the frontier as an area where population density was two persons/mile^2 or less, and in the figure the frontier is represented in the lightest yellow states. Land densities rose in established areas as the frontier progressed westward through time. Bazzi et al. (2017) provide a similar description of the frontier.

Table 1 lists the major federal land laws enacted by Congress that distributed property rights to land and minerals on the frontier. The demand for free small freeholds was incorporated into policy, beginning with the Preemption Act of 1830 and its many amendments (Kanazawa 1996) to accommodate and legally recognize squatter claims and on through the Homestead Act of 1862 and its adjustments. The Homestead Act effectively was ended by Congress in 1934. Under all laws, property

Table 1 Federal land distribution laws

Law	Date	Stated goal and brief impacts
Land Ordinance of 1785	May 20, 1785	Established the Public Land Survey System
Land Ordinance of 1787 (Northwest Ordinance)	July 13, 1787	Determined that the land south of Canada, north of Ohio, west of Pennsylvania, and east of the Mississippi River would be distributed by Congress and that Congress would institute governments and laws in this territory
Land Act of 1796	May 18, 1796	Made the rectangular system of 6 square mile townships permanent and determined the size of sections to be sold. Set minimum land prices
Preemption Act	May 29, 1830	Allowed settlers to occupy and purchase federal lands up to 160 acres at $1.25 an acre
Preemption Act	September 4, 1841	Permanently recognized preemption or squatter claims of land. Donaldson (1884) estimates around 175,000,000 acres were secured by individuals under the Preemption Acts
Graduation Act	August 3, 1854	Reduced the minimum prices of unsold federal government from $1.00/acre to $0.125/acre
Homestead Act	May 20, 1862	160 acres of federal land was made available to individual actual settlers after 5-year continuous residency
Coal Lands Act	July 1, 1864	Distributed coal lands at $20/acre and allowed individuals and associations to claim 160 acres and 320 acres, respectively
Timber Culture Act	March 3, 1877	Applied to 11 western semiarid states and territories to augment Homestead Act claims. Settlers paid $0.25/acre at filing and $1.00/acre when proving compliance
Desert Land Act	March 3, 1873	Authorized an additional 160 acres to Homestead claims if 40 acres of trees were grown
Timber and Stone Act	June 3, 1878	Authorized sale of land at $2.50/acre for land valuable for timber or stone in far western states and territories
Mining Lode Act	July 26, 1866	First major mining law, allowed individuals to claim ownership of ore veins
Mining Act	May 10, 1872	Second major mining law, added placer or shallow ore bodies; required a $100 investment in development to obtain title; procedure for obtaining title outlined
Oil Placer Act	1897	Recognized oil deposits as claimable as a placer ore deposit under the Mining Act of 1872
Stock- Raising Homestead Act	December 29, 1916	Authorized 640-acre homesteads to raise livestock

Sources: Material drawn from Donaldson (1884), Hibbard (1924), Robbins (1942), Gates (1968), and Lacy (1995)

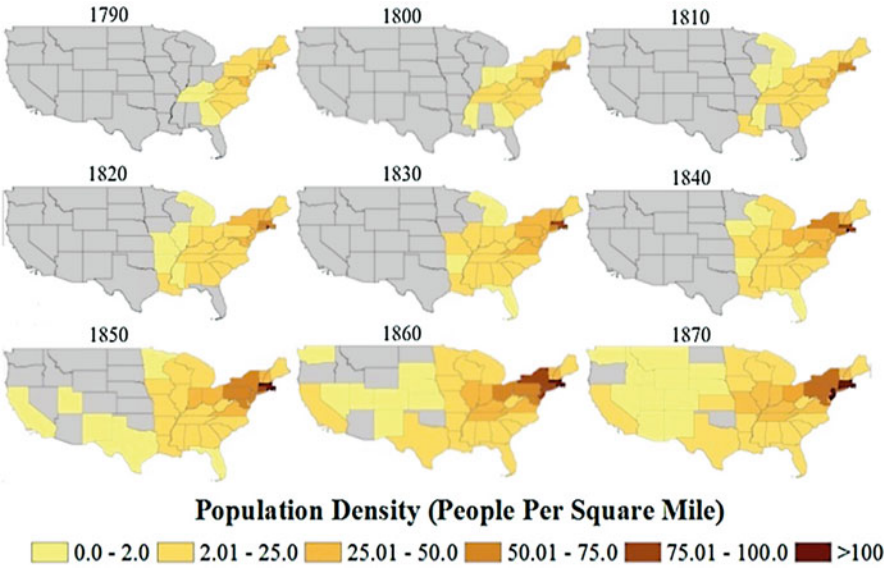


Fig. 1 The progression of the frontier. Notes: State population densities from 1790 to 1870 are from the US Census. Gray states are those with missing data. US Census Bureau (2012, September 6), retrieved from https://www.census.gov/dataviz/visualizations/001/. To construct the figure, population densities for each individual state were taken from the decennial censuses for the years 1790 to 1870. ArcGIS was used to create maps with contemporary state boundaries and their respective population densities for each census

rights to agricultural land were given out piecemeal in plots of 40 to 160 acres (later, up to 640 acres) with the requirement of occupancy and beneficial use (Hibbard 1924; Robbins 1942; Gates 1968). Through these land allocation laws, immense amounts were placed under private ownership. Under the Homestead Act, for example, some 2,758,818 original entries were made between 1863 and 1920 for 437,932,183 acres, an area larger than Alaska (Gates 1968).

There is an extensive literature on the wealth generated from migration to and along the frontier and associated capital gains from land sales. There were benefits from arriving early in a frontier region and obtaining, improving, and selling land. Relatively poorer migrants often benefitted disproportionately, and the rents generated led a more egalitarian wealth distribution than what was found in non-frontier areas. Major works include Lebergott (1985), Oberly (1986) for the period following the War of 1812, von Ende and Weiss (1993) for 1800–1860, Swierenga (1966) for 1840–1869, Kearl et al. (1980) for 1850–1870, Steckel (1989) for 1850–1860, Galenson and Pope (1989) for 1850–1870, Ferrie (1993) for 1850–1860, Stewart (2009) for 1860–1880, and Gregson (1996) for the latter part of the nineteenth century. Most of the studies examine the East and West North Central regions, where production conditions supported small-farm distribution. Easterlin (1960) reveals gradual convergence in per capita income and population patterns between 1840 and 1950 across the frontiers that are shown in Fig. 1.

The Private Provision of Public Goods by Land Owners

The hypothesis that small-farm distribution of property rights to land resulted in civic virtue among freeholders and political participation as suggested by Jefferson, de Tocqueville, and Fredrick Jackson Turner has not been tested. There naturally are measurement challenges, and within the USA there are no clear baselines for comparisons. The literature suggests such differences existed between US and Latin American frontiers as examined below. There are no comparative studies of political activity (voting participation rates, per capita involvement in political office, politician turnover rates) that might have varied with land allocation across and along the US frontier. There is, however, suggestive evidence for public goods investment in education in the USA that supports the relationship. Go and Lindert (2010) find that school enrollment and the number of teachers per capita were greater in the rural northern USA in the mid-nineteenth century, where small farms predominated, than in the South where there was a more heterogeneous mix of constituencies and farm sizes. They also find that student enrollments in the rural north were greater than in most of Europe by 1850. Northern schools relied more than elsewhere on local public money and governance. Go and Lindert (2010) credit the autonomy of local governments and the voting power of citizens in rural communities for this education achievement.

Goldin (1998, 2001) and Goldin and Katz (2010) explore the rise of high school or secondary education in the USA at the turn of the twentieth century that far outpaced that in Europe. Although the high school movement began in New England, it spread rapidly to the central Midwest and western USA that were dominantly rural, agricultural, and in the Midwest at least, characterized by small-farm ownership. The pedagogical focus was practical, pragmatic in emphasis, and egalitarian, aimed at understanding and using new technologies. Educated farmers and their children were more apt to adopt the new technologies and cropping changes underway at the time, and they needed high school education to be able to take advantage of the opportunities. Investment in general education, as compared to apprenticeships and firm (or farm)-specific training, was useful in a highly mobile society where labor would migrate to new locations and employment. School governance was decentralized with 130,000 independent school districts in America by the 1920s, locally supported and administered (Goldin 2001, 279). The demand for and local support of practical education in rural areas are consistent with the argument that small freeholders had incentives to innovate because they captured the rents from doing so and because they had strong ties to their communities. They were motivated to contribute financially and administratively to their schools. The school districts were small and homogenous with respect to citizen incomes and educational objectives. These conditions promoted local collective action to invest in schooling.

Related evidence for the impact of small-holder ownership on the frontier on democracy and individual initiative as claimed by Turner (1893) is provided by Bazzi et al. (2017). They map US frontiers and find that those who lived along them shared characteristics of rugged individualism, self-reliance, and opposition to government intervention and redistribution programs.

Property Rights to Minerals and Oil and Gas Deposits

Not only were frontier lands generally distributed in small parcels, but subsurface mineral deposits and oil and gas formations were secured initially by small holders. In most countries, the subsurface estate has been owned by the crown and later the state. In the USA, the Land Ordinance of 1785 included mineral lands that were to be sold to private bidders as surface lands, with reservation of 1/3 of the properties to the government. As it turns out, however, minerals were not prominent in the central and eastern USA, except for copper deposits in Michigan. Between 1776 and 1848, most mineral lands went to private ownership as agricultural lands, and the government did not claim the minerals below them (Lacy 1995). As the frontier moved into the far West, however, things changed. Rich gold and silver ore was found, beginning in California and then throughout the western region, and those individuals who discovered ore deposits claimed ownership as first possession even though minerals as separate resources from the surface land were not covered in the land laws until 1866. Eventually well over 600 mining camps were established throughout the West with bylaws regarding the requirements for defining, maintaining, and trading individual mining claims and arbitrating disputes over them. Figure 2 shows the range of mining camps in California in the late nineteenth century and the US West. Each camp had local mineral rights bylaws.

Libecap (1978) describes the institutional evolution of private property rights to ore on the Comstock Lode of Nevada from those defined by early mining camp rules to the actions of the territorial and state government and, ultimately, the federal government in the Mining Laws of 1866 and 1872. Mineral rights could be traded, and as surface ore was depleted, requiring more capital-intensive deep-vein mining, surface mining claims were sold and consolidated into new mining companies with their shares listed on the San Francisco Stock Exchange and other capital markets.

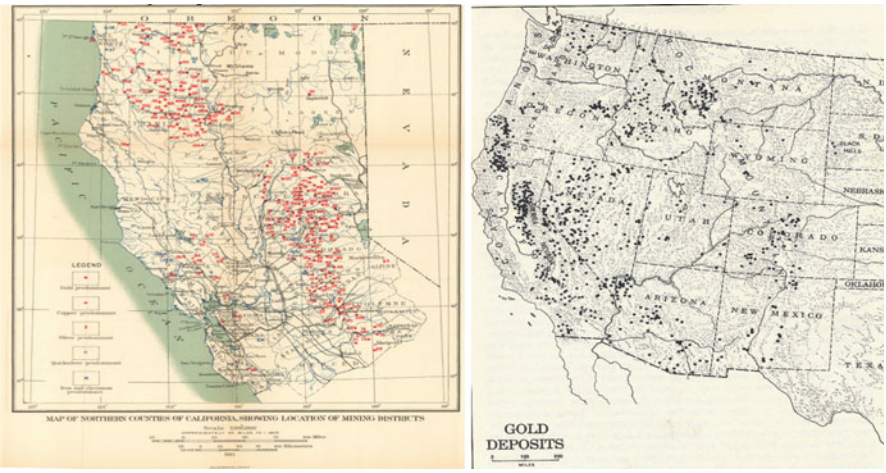


Fig. 2 Frontier mining camps in California and the far West. (Source: California, Hill (1912, 78) and gold deposits, Paul (1963, 4))

There is an extensive literature on mining camp rules, including that by Umbeck (1977), Clay and Wright (2005), and others noted by Libecap (2007). Drawing from public use samples from the 1850 to 1852 US Censuses for California, Clay and Jones (2008) analyze who went to the mining frontier in California and how they fared in their efforts to become rich. Most came from New York, Illinois, Michigan, and Wisconsin as well as Missouri, areas that once had been on the agricultural frontier. Clay and Jones find that this risky endeavor did not always pay. On average, migration may have lowered real earnings for many prospectors, relative to what they might have earned in their home locations. It was the promise of riches that drove migration and the ability to acquire individual ownership of ore in the quest for rents that drove migrants west. Overall, private property rights to minerals encouraged exploration, discovery, and production. As the mining industry developed, American mining engineering schools and technologies became world leaders. Human capital and physical technology investments in the USA led the country to produce beyond what its resource endowments would have otherwise suggested.

Ownership of major oil and gas deposits also went to private individuals. Oil discoveries in Pennsylvania in 1859 and in Oklahoma, Texas, Kansas, Illinois, and California in the late nineteenth century and into the twentieth century largely took place on private lands, and as noted above, surface owners held title to the minerals beneath their properties. As with minerals, private ownership and the potential to gain rents encouraged exploration and production. Those who specialized in search were called wildcatters, and they obtained leases from surface owners for exploration and later, if discoveries were made, for production. By the early twentieth century, prolific new oil fields were developed, particularly in the central and western USA. Fabled oil fields include Spindletop, Yates, Hendrick, East Texas in Texas, Oklahoma City and Seminole in Oklahoma, and Long Beach and Kern County in California. Given relatively low costs of locating, drilling, and producing in new fields, entry was easy, and production soared (Libecap and Wiggins 1984). The resulting output drove local economies; created local, self-reliant middle and upper classes, characteristic of Oklahoma and Texas today; and made the USA a major world producer.

There is nothing comparable in virtually all other countries, where minerals and oil and gas deposits are owned by the state. Surface property owners may or may not share in the rents generated from new discoveries, and they often resist exploration on their properties. Incentives for exploration and production are quite different. This explains why, for example, the current rapid adoption of nontraditional production techniques (fracking) is based on US innovation and occurs dominantly on private properties in the country, where surface property owners can anticipate a share in the returns.

Although private ownership of minerals and oil and gas encouraged exploration and discovery, competitive output may have been nonoptimal in an aggregate sense (Libecap and Wiggins 1984; Clay and Wright 2005). Figure 3 shows the many small mineral claims that often overlapped with unclear property rights, at least early in a mining camp. The rush to locate and prove or extract ore could have motivated excessive production. The figure also shows town lot drilling that was found where

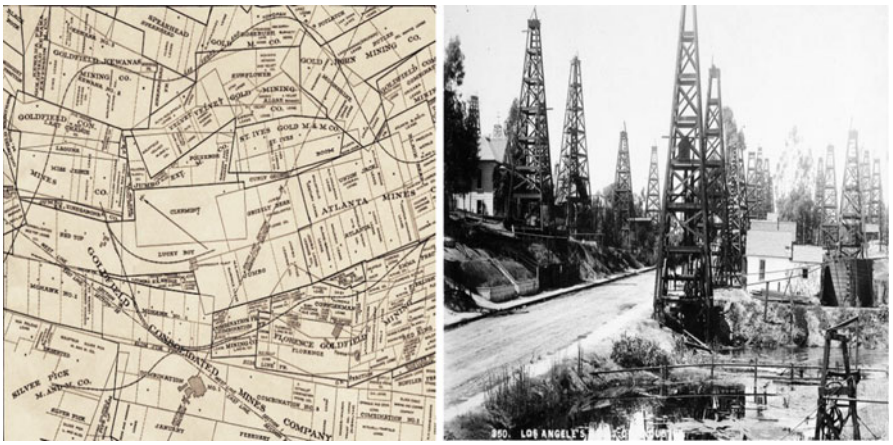


Fig. 3 Fragmented minerals and oil and gas ownership, Goldfield, Nevada, 1905, and Long Beach, California, 1920. (Source: https://www.mininghistoryassociation.org/Meetings/Tonopah/Goldfield%20Claims.jpg; http://texasalmanac.com/topics/business/history-oil-discoveries-texas, https://www.kcet.org/shows/lost-la/when-oil-derricks-ruled-the-la-landscape (2011))

surface ownership was extremely fragmented. With subterranean oil and gas deposits migratory, adjacent producers had an incentive to competitively drill and drain, generating classic common-pool resource losses.

It is difficult, however, to make strong welfare conclusions in either case. The trade-offs with more centralized, large minerals or oil and gas ownership, involving potentially less discovery and innovation, may offset the losses of rapid output. All in all, competitive output was mitigated by lease consolidation and large-scale hard-rock mining later in the life of most mineral regions and with the unitization of oil and gas fields that designated a single firm to develop a hydrocarbon formation (Wiggins and Libecap 1985).

Property Rights on Latin American Frontiers

Besides England of course, the western hemisphere was colonized by France (largely in Canada), Portugal (Brazil), and Spain (Central and South America) with limited colonization by Holland, Russia, and Sweden. Property rights to land and minerals were starkly different in Latin America, providing an opportunity to examine the consequences. The English language literature on frontier resource ownership is smaller, and only generalized conclusions are presented here. Nevertheless, the diverse patterns relative to the USA are quite clear (Alston et al. 2012). The differences in economic outcomes across the US and Latin American frontiers are in part attributable to the factors underscored by economists and economic historians on the clarity, enforcement, and durability of the property rights to land and minerals provided in the two regions. In Latin America, property

rights were ambiguously defined between the crown/state and private individuals, creating uncertainty as to decision-making, net rent capture, and durability. Short-term, extractive practices often were the result. Other differences in outcomes are attributable to the factors underscored in the political science, historical, legal, and other economics and economic history literatures. Property rights to land and minerals were distributed in a centralized way in large blocks to privileged parties, not in a decentralized manner to common people as in the USA. The political and social power structures that emerged in the two settings were, accordingly, very distinct, leading to lasting effects in political stability, social interactions, individual mobility and optimism, and economic growth.

Hennessy (1978) provides a summary of experiences on multiple frontiers in Latin America. Unlike the relatively orderly progression across North America shown in Fig. 1 above, Latin American frontiers varied within and across countries. Latin America was formed of generally extractive colonies (Acemoglu et al. 2001). The process of European migration and settlement was more centralized and limited, very different from English colonization in North America (Engerman and Sokoloff 1997; Acemoglu et al. 2001; Linklater 2013). Ownership of land and minerals was retained by the crown in the colonial period, and mineral ownership continued to be held by the state in postcolonial periods. In the colonies, use rights to land were granted by the Spanish and Portuguese crowns in large tracts to political elites who paid tithes or quit rent payments (Hennessy 1978; García-Jimeno and Robinson 2011). Large-scale immigration of the kind that occurred in the USA was both relatively unattractive because of limited access to desirable land and officially discouraged due to concerns in home countries of population shortages. The experience of common people on the land in Spain and Portugal was very different from their English counterparts who already had familiarity with broad private land ownership prior to emigration. Some 243,000 immigrants may have arrived in Latin America in the first 100 years of colonization along with perhaps 7 million more between 1820 and 1920, as compared to 34 million to the USA (Hennessy 1978).

Hennessy argues that long-term economic underperformance and political conflicts in Latin America have their origins in the initial land distribution (Hennessy 1978). It created enduring political and social elites and established internal country conditions that have influenced subsequent immigration, urbanization, and industrialization. Land grants – latifundia, encomienda, *sesmarias*, estancias, haciendas – not homesteads, were the typical rural institution. The estates often were near-feudal organizations with natives and immigrant farm laborers bound to the land and the patriarchal structure (Hennessy 1978). Others worked land in or near the grants as sharecroppers and tenants, with payments or crop shares to the large landholders (Leff 1997; Chowning 1997). Mandatory labor conscriptions were assessed in native communities for working the mines of Bolivia and Peru via the Mita (Arad 2013). There was little active smallholder participation in land or resource markets in the way that occurred in North America. Ownership and wealth were highly concentrated as were the political structures (Frank 2001; Arad 2013). Relative to the US frontier, a much smaller agricultural land or minerals-based middle class developed.

Privileged locals invested in agricultural export industries and may have neglected or resisted other development opportunities that could undermine their positions.

Engerman and Sokoloff (1997, 2012) and Acemoglu et al. (2001, 2002, 2005) point to factor endowments, climate and disease environments, and densities of native populations, along with general institutional differences emanating from the source countries. They argue that these factors explain the differential performance of North and South American colonial and postcolonial economies and for their unattractiveness for immigration. The underlying nature of the property rights granted to land and minerals, their distribution among the population, and associated incentives for innovation, market development, political participation, and social mobility also likely played critical roles. The US South had similar factor and environmental characteristics to much of Latin America, but small freehold farms did emerge alongside plantations (Engerman and Sokoloff 1997). Economic, social, and political outcomes there were more similar to the northern USA than to Latin America. Moreover, in temperate Argentina, Uruguay, Chile, and southeast Brazil, the best lands were often preempted by large grant holders, and small farmers had difficulties in obtaining title to their lands. Land conflicts due to incomplete property rights also occurred elsewhere in Latin America where small holdings otherwise would have been economically viable (Sanchez et al. 2010). The relatively fewer immigrants to these regions more easily became tenants or were employed as agricultural laborers or range riders, or gauchos, than small freeholders (Engerman and Sokoloff 1997; Garciá-Jimeno and Robinson 2011). In this regard, the Latin American frontier seems to have had an enduring impact, as F.J. Turner argued for the USA, but with very different consequences, because property rights were assigned to it in very different ways.

Conclusion

Property rights are the most fundamental institution in any society, an assessment recognized by early political economists and philosophers, as well as by later historians and legal scholars. Economists and economic historians have understood the direct economic importance of property rights, but the broader implications of the nature and distribution of the rights granted to land and minerals have not been central in recent research. Yet, these seem critical in explaining long-term differences in economic, political, and social performance. Property rights determine who has decision-making authority over assets and who bears the costs and benefits of those decisions. They make markets possible, define timelines, and provide incentives for investment, innovation, and trade. They mitigate the losses of open access and provide the basis for long-term economic growth. They assign ownership, wealth, political influence, and social standing. For these reasons, how property rights are defined and allocated to land and minerals determines who the players are and who has a lasting stake in the society and economy. The contrasting experiences of the North American and South American frontiers illustrate these arguments, and there likely are durable, path-dependent effects in the process of economic growth. The property rights distribution to land seems to have affected how citizens participated

in capital and other markets and how they assessed their ability to advance economically, socially, and politically through markets as compared to state intervention. Whether and how the effects of broad ownership of land are maintained in more urban, developed economies requires research attention.

References

Abramitsky R, Braggion F (2006) Migration and human capital: self-selection of indentured servants to the Americas. J Econ Hist 66(4):882–905

Acemoglu D, Robinson J (2012) Why nations fail. Crown, New York

Acemoglu D, Johnson S, Robinson J (2001) The colonial origins of comparative development: an empirical investigation. Am Econ Rev 91(5):1369–1401

Acemoglu D, Johnson S, Robinson J (2002) Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q J Econ 117(4):1231–1294

Acemoglu D, Johnson S, Robinson J (2005) The rise of Europe: Atlantic trade, institutional change, and economic growth. Am Econ Rev 95(3):546–579

Allen DW (2011) The institutional revolution: measurement and the economic emergence of the modern world. University of Chicago Press, Chicago

Alston LJ, Harris E, Mueller B (2012) The development of property rights on frontiers: endowments, norms, and politics. J Econ Hist 72(3):741–770

Alston E, Alston LJ, Mueller B, Nonnenmacher T (Forthcoming, 2018) Institutional and organizational analysis: concepts and applications. Cambridge University Press, New York

Arad LA (2013) Persistent inequality? Trade, factor endowments, and inequality in republican Latin America. J Econ Hist 73(1):38–78

Barro RJ (1991) Economic growth in a cross section of countries. Q J Econ 106(2):407–443

Barzel Y (1997) Economic analysis of property rights. Cambridge University Press, New York

Bazzi S, Fiszbein M, Gebresilasse M (2017) Frontier culture: the roots and persistence of "rugged individualism" in the United States. NBER working paper no. 23997. Centre for Economic Policy Research, London

Besley TJ (1995) Property rights and investment incentives: theory and evidence from Ghana. J Polit Econ 103(5):903–937

Besley TJ, Ghatak M (2010) Property rights and economic development. In: Rodrik D, Rosenzweig M (eds) Handbook of development economics, vol 5. Elsevier, New York, pp 4525–4595

Cheung SN (1970) The structure of a contract and the theory of a non-exclusive resource. J Law Econ 13(1):49–70

Chowning M (1997) Reassessing the prospects for profit in nineteenth-century Mexican agriculture from a regional perspective: Michoacán, 1810–60. In: Haber S (ed) How Latin America fell behind. Stanford University Press, Palo Alto, pp 179–215

Clay K, Jones R (2008) Migrating to riches? Evidence from the California gold rush. J Econ Hist 68(4):997–1027

Clay K, Wright G (2005) Order with law? Property rights and the California gold rush. Explor Econ Hist 42(2):155–183

Coase R (1960) The problem of social cost. J Law Econ 3:1–44

Costello C, Gaines SD, Lynham J (2008) Can catch shares prevent fisheries collapse? Science 321:1678–1681

Cox M, Arnold G, Villamayor Tomás S (2010) A review of design principles for community-based natural resource management. Ecol Soc 15(4):38. On Line: http://www.ecologyandsociety.org/vol15/iss4/art38/

Dahlman C (1979) The problem of externality. J Law Econ 22:141–162

Davis LE, North DC (1971) Institutional change and American economic growth. Cambridge University Press, New York

de Tocqueville A (1835) Democracy in America (trans: Goldhammer A (2004)). Penguin Putnam, New York
Dixit A (2009) Governance institutions and economic activity. Am Econ Rev 99(1):3–24
Donaldson T (1884) The public domain: its history, with statistics. Government Publishing Office, Washington, DC
Easterlin RA (1960) Interregional differences in per capita income, population, and total income, 1840–1950. In: Trends in the American economy in the nineteenth century, the conference on research in income and wealth, NBER. Princeton University Press, Princeton, pp 73–140
Edwards J, Ogilvie S (2012) What lessons for economic development can we draw from the champagne fairs? Explor Econ Hist 49(2):131–148
Ellickson RC (1993) Property in land. Yale Law J 102:1315–1400
Ely JW Jr (2008) The guardian of every other right: a constitutional history of property rights, 3rd edn. Oxford University Press, New York
Engerman SL, Sokoloff KL (1997) Factor endowments, institutions, and differential paths of growth among new world economies. In: Haber S (ed) How Latin America fell behind. Stanford University Press, Palo Alto, pp 260–291
Engerman SL, Sokoloff KL (2012) Economic development in the Americas since 1500. Cambridge University Press, New York
Ferrie JP (1993) 'We are Yankees now': the economic mobility of two thousand antebellum immigrants to the United States. J Econ Hist 53(2):388–391
Ford AC (1910) Colonial precedents of our national land system as it existed in 1800. No 352, History series, vol 2, no 2. Bulletin of the University of Wisconsin, Madison, pp 321–477
Frank ZL (2001) Exports and inequality: evidence from the Brazilian frontier, 1870–1937. J Econ Hist 61(1):37–58
Franklin B (1751/1845) Observations concerning the increase of mankind, peopling of countries, etc. In: McCulloch J (ed) The literature of political economy: a classified catalogue of selected publications. . .with historical, critical, and biographical notices. Longman, Brown, Green, and Longmans, Boston
Galenson DW, Pope CL (1989) Economic and geographic mobility on the farming frontier: evidence from Appanoose County, Iowa, 1850–1870. J Econ Hist 49(3):635–655
Galiani S, Schargrodsky E (2010) Property rights for the poor: effects of land titling. J Public Econ 94(9–10):700–729
García-Jimeno C, Robinson JA (2011) The myth of the frontier. In: Costa DL, Lamoreaux NR (eds) Understanding long-run economic growth: geography, institutions, and the knowledge economy. University of Chicago Press, Chicago, pp 49–89
Gates P (1968) History of public land law development. Public Land Law Review Commission, Washington, DC
Go S, Lindert PN (2010) The uneven rise of American public schools to 1850. J Econ Hist 70(1):1–26
Goldhammer A (2004) Alexis de Tocqueville, democracy in America, volume 1, translation. Penguin Putnam, New York
Goldin C (1998) America's Graduation from High School: the evolution and spread of secondary schooling in the twentieth century. J Econ Hist 58(2):345–374
Goldin C (2001) The human capital century and American leadership: virtues of the past. J Econ Hist 61(2):263–292
Goldin C, Katz LF (2010) The race between education and technology. Harvard University Press, Cambridge
Gordon HS (1954) The economic theory of a common-property resource: the fishery. J Polit Econ 62(2):124–142
Grafton RQ, Squires D, Fox KJ (2000) Private property and economic efficiency: a study of a common-pool resource. J Law Econ 43(2):679–714
Gregson ME (1996) Wealth accumulation and distribution in the midwest in the late nineteenth century. Explor Econ Hist 33(4):524–538

Greif A, Milgrom P, Weingast BR (1994) Coordination, commitment, and enforcement: the case of the merchant guild. J Polit Econ 102(4):745–776
Grubb F (1986) Redemptioner immigration to Pennsylvania: evidence on contract choice and profitability. J Econ Hist 46(2):407–418
Hennessy A (1978) The frontier in Latin American history. University of New Mexico Press, Albuquerque
Hibbard BH (1924) A history of the public land policies. Macmillan, New York
Hill JM (1912) The mining districts of the western United States. Bulletin 507. U.S. Department of the Interior U.S. Geological Survey. Government Printing Office, Washington, DC
Hornbeck R (2010) Barbed wire: property rights and agricultural development. Q J Econ 125(2): 767–810
https://www.kcet.org/; https://www.kcet.org/shows/lost-la/when-oil-derricks-ruled-the-la-landscape, 2011
Jacoby HG, Li G, Rozelle S (2002) Hazards of expropriation: tenure insecurity and investment in rural China. Am Econ Rev 92(5):1420–1447
Johnson RN, Libecap GD (1982) Contracting problems and regulation: the case of the fishery. Am Econ Rev 72(5):1005–1022
Kanazawa MT (1996) Possession is nine points of the law: the political economy of early public land disposal. Explor Econ Hist 33(2):227–249
Katz SN (1976) Thomas Jefferson and the right to property in revolutionary America. J Law Econ 19(3):467–488
Kearl JR, Pope CL, Wimmer LT (1980) Household wealth in a settlement economy: Utah, 1850–1870. J Econ Hist 40(3):477–496
Lacy JC (1995) Going with the current: the genesis of the mineral laws of the United States. 41st Rocky Mountain Mineral Law Institute. Mathew Bender/Lexus Nexus, San Francisco, pp 10-1–10-55
Lebergott S (1985) The demand for land: the United States, 1820–1860. J Econ Hist 45(2):181–212
Leff NH (1997) Economic development in Brazil, 1822–1913. In: Haber S (ed) How Latin America fell behind. Stanford University Press, Palo Alto, pp 34–64
Libecap GD (1978) Economic variables and the development of the law: the case of western mineral right. J Econ Hist 38(2):338–362
Libecap GD (1981) Locking up the range: federal land use controls and grazing. Ballinger Publishing, Cambridge
Libecap GD (2007) The assignment of property rights on the western frontier: lessons for contemporary environmental and resource policy. J Econ Hist 67(2):257–291
Libecap GD (2008) Open-access losses and delay in the assignment of property rights. Ariz Law Rev 50(2):379–408
Libecap GD, Lueck D (2011) The demarcation of land and the role of coordinating property institutions. J Polit Econ 119(3):426–467
Libecap GD (2016) Coasean bargaining to address environmental externalities. In: Bertrand E, Menard C (eds) The Elgar companion to Ronald Coase. Edward Elgar, Northampton, pp 97–109
Libecap GD, Wiggins SN (1984) Contractual responses to the common pool: prorationing of crude oil production. Am Econ Rev 74(1):87–98
Libecap GD, Lueck D, O'Grady T (2011) Large-scale institutional changes: land demarcation in the British Empire. J Law Econ 54(4):295–327
Lindert PH, Williamson JG (2013) American incomes before and after the revolution. J Econ Hist 73(3):725–765
Lindert PH, Williamson JG (2014a) Unequal growth: American incomes since 1650. Princeton University Press, Princeton
Lindert PH, Williamson JG (2014b) American colonial incomes, 1650–1774, NBER working paper 19861. National Bureau of Economic Research, Stanford
Linklater A (2013) Owning the Earth: the transforming history of land ownership. Bloomsbury, New York

McCulloch J (1845) The literature of political economy: a classified catalogue of selected publications. . .with historical, critical, and biographical notices. Longman, Brown, Green, and Longmans, Boston
Mehlum H, Moene K, Torvik R (2006) Institutions and the resource curse. Econ J 116:1–20
Merrill TW, Smith HE (2010) Property. Oxford University Press, New York
Mining History Association, https://www.mininghistoryassociation.org/, https://www.mininghistoryassociation.org/Meetings/Tonopah/Goldfield%20Claims.jpg
North DC (1981) Structure and change in economic history. Norton, New York
North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
North DC, Thomas R (1973) The rise of the western world: a new economic history. Cambridge University Press, New York
North DC, Wallis JJ, Weingast BR (2009) Violence and social orders: a conceptual framework for interpreting recorded human history. Cambridge University Press, New York
Oberly JW (1986) Westward who? Estimates of native white interstate migration after the war of 1812. J Econ Hist 46(2):431–440
Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, New York
Paul RW (1963) Mining Frontiers of the far west 1848–1880. University of New Mexico Press, Albuquerque
Pigou AC (1920) The economics of welfare. Macmillan, London
Priest C (Forthcoming, 2019) Credit nation. Princeton University Press, Princeton
Robbins RM (1942) Our landed heritage. The public domain 1776–1936. Princeton University Press, Princeton
Rodrik D (2008) Second-best institutions. Am Econ Rev 98(2):100–104
Sanchez F, Lopez-Uribe M, Fazio A (2010) Land conflicts, property rights, and the rise of the export economy in Colombia, 1850–1925. J Econ Hist 70(2):378–399
Scott A (1955) The fishery: the objectives of sole ownership. J Polit Econ 63(2):116–124
Steckel R (1989) Household migration and rural settlement in the United States, 1850–1860. Explor Econ Hist 26(2):190–218
Steinbeck J (1939/1976) The grapes of wrath. Penguin Books, New York
Stewart JI (2009) Economic opportunity or hardship? The causes of geographic mobility on the agricultural frontier, 1860–1880. J Econ Hist 69(1):238–268
Story J (1858) Commentaries on the constitution of the United States, vol 2. Little Brown, Boston
Swierenga RP (1966) Land speculator "profits" reconsidered: Central Iowa as a test case. J Econ Hist 26(1):1–28
Texas Almanac, https://texasalmanac.com/; http://texasalmanac.com/topics/business/history-oil-discoveries-texas
Turner FJ (1893) The significance of the frontier in American history. Report of the American Historical Association, pp 199–227. http://www.archive.org/stream/1893annualreport00ameruoft/1893annualreport00ameruoft_djvu.txt
U.S. Census Bureau (2012, September 6). Population Statistics. https://www.census.gov/dataviz/visualizations/001/
U.S. Public Lands Commission (1880) Report of the Public Lands Commission. 46th congress, 2nd session, house executive document 46. Government Printing Office, Washington, DC
Umbeck J (1977) A theory of contract choice and the California gold rush. J Law Econ 20(2):421–437
von Ende E, Weiss T (1993) Consumption of farm output and economic growth in the old northwest, 1800–1860. J Econ Hist 53(2):308–318
Wiggins SN, Libecap GD (1985) Oil field unitization: contractual failure in the presence of imperfect information. Am Econ Rev 75(3):368–385
Wilen JE (2005) Property rights and the texture of rents in fisheries. In: Leal D (ed) Evolving property rights in marine fisheries. Rowman and Littlefield, Lanham, pp 49–67
Winch D (1965) Classical political economy and colonies. Harvard University Press, Cambridge

Major Water Infrastructure and Institutions in the Development of the American West

Canals, Dams, and Hydropower

Zeynep K. Hansen and Scott E. Lowe

Contents

Abstract

In the history of western expansion in the United States, arguably no natural resource has impacted the economy of the American west more than water. As a consumptive natural resource, water is necessary for urban growth and development, industrial mining, and for irrigated agriculture. However, water resources also provide non-consumptive, in-stream benefits, by allowing for transportation, energy production, and recreation. This chapter addresses the roles that water resources played in enabling western expansion in the United States, first into the trans-Appalachian west, and later into the more arid western territories. We address the institutions that arose in tandem with the development of water resources, and the complexities that competing demands have introduced to the management of these (often) constrained water resources.

Keywords

Irrigated agriculture · Dams · Mining · Western expansion · Water resources

Z. K. Hansen (✉) · S. E. Lowe
Department of Econimics, Boise State University, Boise, ID, USA
e-mail: zeynephansen@boisestate.edu; scottlowe@boisestate.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_70

Introduction

Fresh water is a raw, consumptive natural resource, necessary for urban growth and development. However, fresh water is also an input to production that can be used to increase the productivity of mining and industrial processes and enable irrigated agriculture. In addition to its value as an input to production, water and water resources may offer transportation, energy, recreation, and environmental benefits, benefits that may be recognized in tandem with the consumptive or industrial uses and benefits that may in fact dominate the consumptive uses. In the history of western expansion in the United States, arguably no natural resource has impacted the economy of the American west more than water. However, for all of the benefits that they enable, water resources are not without costs and complexities. When scarce, water resources are virtually priceless but when abundant can be costly to restrain and even more costly when they generate damages via flooding. Unlike many of the in situ or sedentary natural resources, riparian water resources in the natural environment are transient and transboundary, crossing sociopolitical boundaries and meandering over time. Water resources are also fleeting – uncertain from day to day, month to month, and year to year, particularly in an age of heightened climate uncertainty and variability and even more so in the arid parts of the world in which water is *the* constraining input. Water resources are often of varying quality and can be rendered worthless due to degradation and pollution. Water is bulky and difficult to move but can be lost due to evaporation and groundwater recharge.

As much of the precipitation in the western United States falls in the high mountains and is stored in snowpack at a great distance from anthropogenic demand, humankind has been bequeathed with a natural, environmental buffer and storage service. However, the meting of water resources in the arid western United States rarely aligns with demand, so in order to overcome some of these complexities and uncertainties, humankind has developed great water works to store, pump, and move water and to put it toward beneficial uses. In tandem, water institutions and governance have evolved to regulate and distribute the vast water resources. Dams, canals, and aqueducts crisscross the American west, storing, elevating, and transporting water resources. Although effective and able to remove uncertainty from an otherwise exceedingly complex resource, major water infrastructure projects are costly and have often involved decades-long negotiations and politicking. This chapter addresses the role of major water infrastructure in the economic development of the American west.

Focusing first on western migration from the eastern seaboard to the trans-Appalachian west, we address the growth of the US canal system and the ways in which these canals enabled industrial growth, opened up new markets, and changed the composition of the Midwestern United States. Working in tandem with a growing rail industry, canals created a great commercial center in the Midwest. Contributions by Bogart (1913), Turner (1920), Rae (1944), North (1956), Cranmer (1960), and Ransom (1964), all addressed in this chapter, illuminate this epoch in the great American migration. This newfound access was not without cost – the

urbanization and industrialization that resulted had a deleterious impact on the quality of life, noted by Haines et al. (2003) and Chanda et al. (2008).

Following this great first migration into the trans-Appalachian west, we look to the second great migration westward, into the more arid territories. Perhaps based on prior experiences and the great wealth that was created in the Midwest, pioneers and pilgrims found new wealth in the arid west in the form of mining and agriculture. Unlike the east, where water supplies were plentiful, water in the new arid west was a constraining resource: water supplies were snowpack-determined, and often were not available for many months of the year. As such, major water works and infrastructure provided a man-made source of storage, allowed pioneers to reclaim the desert, lengthened the growing season, and enabled the planting of crops that were previously thought impossible to grow. Hydraulic mining for precious metals required the diversion of rivers and the development of high-pressure water extraction techniques. Pisani (1984) and Reisner (1986) provide insights into the mining and agricultural booms in western American history.

The growth of water infrastructure and storage in the arid western United States necessitated a new era of water governance. Federal legislation, including the Homestead and Reclamation Acts, facilitated the migration west but also introduced a number of issues and complications. Coman (1911), Libecap (1981, 2011), Libecap and Hansen (2002), Hansen and Libecap (2004a, b), McCool (1994), and Rhode (1995), among others, detail these issues and complications.

Hansen and Libecap (2004a) discuss US land policy and show that the Homestead Act of 1862 worked well initially in the northern plains while sufficient precipitation (precipitation without prolonged drought periods) allowed farmers to utilize familiar farming practices on small farms. Although there were warnings (e.g., Powell's call for a minimum of 2560-acre pastoral regions in the arid lands) and proposed bills to change federal policy, these were not considered, and strong sentiment for maintaining the emphasis on small farms remained. The prolonged drought during 1917–1921 was especially difficult for the homestead farmers of the northern plains, and as a result, many small farms failed. The prevalence of 160-acre and 320-acre farms in the region (and as a result a decline in average farm size) was due largely to the policy model and the restrictions imposed by the Homestead Act. Moreover, the prescribed dry farming techniques that called for intense cultivation of small farms (Libecap and Hansen 2002) produced further disastrous results for farmers in arid regions. Hansen and Libecap (2004b) address these externalities and analyze the origins of the great drought of the 1930s, linking the severe drought and wind erosion to the prevalence of small farms created by the Homestead Act (Hansen and Libecap 2004b). Farm size played a major role in contributing to severe wind erosion as well as inhibiting corrective action as most farmers were not operating at productively efficient sizes to adopt wind erosion controls, such as putting a part of their cropland to fallow.

Prior to the development of major water infrastructure in the arid western United States, conflicts between water users were particularly common, and given the limits of in-stream supplies, demand often outpaced supply (Coman 1911). Although the presence of large dams in the region, partially due to the Reclamation Act, enabled

the spatial and temporal transfer of water resources across seasons making agricultural production viable, there were many negative environmental impacts, including the major declines in wild fish stock due to very low water levels in many streams and rivers (Hansen et al. 2014).

As cities grew, urban demands increased, creating the need to move water from distant sources. Technological developments allowed for hydroelectric power. Vast amounts of hydroelectric power fueled an industrial renaissance and electrified the farm, thus allowing agriculture to thrive in areas where even canals and aqueducts could not reach. Erie and Joassart-Marcelli (2000), Libecap (2009), Kitchens and Fishback (2015), Kitchens (2014), Kline and Moretti (2014), Reisner (1986), and Pisani (1984) all reinforce the importance of major water infrastructure in the electrification of the farm and home.

Westward Expansion

Scholars note that many canal openings, such as the Erie Canal in 1825, were viewed as "epoch-making events" – allowing for the growth of many of the larger northwestern cities and their development into centers of global commerce (Turner 1920). With the Appalachian Mountains geographically dividing the Midwestern lands from markets on the eastern seaboard, canals such as the Erie "provided dramatic proof that a canal to the West was both technically feasible and economically profitable" (Ransom 1964, pp. 365). The canals served as the "highway for a new migration" and provided thoroughfares that moved people into the new west and in turn returned the surpluses of the American west – namely, agricultural commodities and raw natural resources – to the urban centers from which they fled (Turner 1920; Rae 1944). The opening of canals in Ohio had a profound (and predictable) impact on commodity prices. The availability of new markets in the east resulted in a two- or threefold increase in prices paid for western agricultural commodities such as timber, wool, pork, wheat, and corn. Nonagricultural mineral commodities such as coal, salt, pig iron, and stone found easy access to eastern market. Similarly, the access to commodities sold by eastern markets resulted in a decline in the cost of finished goods such as pins, coffee, and sugar, as purchased by Ohio residents (Bogart 1913). Some canals, such as the Wabash and Erie Canal in Indiana, stimulated so much migration that the populations of the counties that it served had quintupled in less than 10 years (Rae 1944). Political historians have identified the new American west, at this time, as holding the balance of power and setting the course for future national progress, particularly with regard to the public domain, tariff systems, banking and currency systems, and interstate commerce rules (Turner 1920).

The growing canal system, together with the related growth in railroad access, "gave birth to the cities of Chicago, Milwaukee, St. Paul, and Minneapolis, as well as to a multitude of lesser cities" (Turner 1920, pp. 137). The Sault Ste. Marie Canal in Canada, at that time representing a small fraction of the total distance of water-based and canal traffic for all of the Midwestern commerce, saw more tonnage than

the Suez Canal (Turner 1920). This Great Lakes traffic resulted in the development of several freshwater port cities, such as Chicago, Detroit, Cleveland, and Buffalo, with Chicago recognized by historians as the "metropolis of the Mississippi Valley" (Rae 1944). The growth resulted in a subsequent industrial and commercial revolution in the Great Lakes region after 1866 – "the tonnage doubled; wooden ships gave way to steel; sailing vessels yielded to steam; and huge docks, derricks and elevators, triumphs of mechanical skill, were constructed" (Turner 1920, pp. 150). Similarly, the economic and policy practices such as the fixation of rates by government officials – practices that would eventually be identified with railway management – had their genesis in early canal development and management (Bogart 1913).

Historically, there has been some debate regarding the overarching economic impact of canal investment. Ransom (1964) notes that many economists are quick to point out the indirect and induced effects of canal investments but ignore the fact that the returns from some canals were insufficient to justify the investments that were made in their construction. Cranmer (1960) identifies three distinct cycles in canal investment between 1815 and 1860: a first wave (1815–1831) noted by Cranmer as "canal mania," in which approximately $50 M was spent on both public and private canals that were viewed to be some of the most successful canals constructed in the United States; a second cycle (1832–1844) in which over $70 M was spent on largely public canals; and a third cycle (1845–1860) in which approximately $56 M was spent, ending precipitously at the outset of the Civil War. The first phase of development was intended to serve as main arteries of transportation, whereas the last wave was intended to improve navigation to and from the Great Lakes (Rae 1944).

Ransom (1964) notes that by 1860, approximately 4,250 miles of canal had been constructed in the United States, with a total capital investment of over $190 M, reflected by a number of "spectacular successes... and spectacular failures"(1964, pp. 366). Using benefits estimated based on canal traffic and revenues generated, Ransom identifies approximately 25 canal segments (located in New York, Ohio, Pennsylvania, Indiana, Illinois, Maryland, and Virginia,) representing $102 M in state investment, and divides them into "probably successful" and "probably not successful" in terms of their benefits being greater than the costs of construction. Of the $102 M in state investment, he notes that only $16.3 M of the total was spent on canals that he deemed to be "probably successful." Toward the end of the canal construction boom, he argues that it is undeniable that some canals were worthwhile investments that contributed considerable impetus to growth and increased the efficiency of the larger US economy. However, he is quick to note that in other cases, canal investments were unnecessary – there were too many canals that were competing with one another, perhaps replicating existing low-cost transportation opportunities such as rail, as opposed to opening new avenues of trade between regions. In many parts of the Midwest, freezing winters would shut down canal commerce and transit for many months of the year, making rail the only option for moving goods.

In support of the growing canal system, and to assist in their construction, the federal government provided aid in the form of land grants (Rae 1944). Unlike the development of the rail system, which received a substantial land grant allocation totaling over 130 M acres, the land grants earmarked toward canal development

totaled only about 4.5 M acres (Rae 1944). As the precursors to the rail land grants, canal land granting adopted an alternating section practice, so as to stimulate development along the length of the canal – a practice which was later successfully applied to rail-focused land grants. Given the shorter length and capital intensiveness of the canal system, relative to the rail system, the land grant was viewed as only part of the financial package necessary to fund the canals and was of "slight assistance" (Rae 1944). At the time, there was very little political opposition to the land grants, particularly given the "prudent proprietor" doctrine, which posited that "since completion of the canals would greatly increase the price of the adjoining public land, Congress would simply be making a sound investment if it gave part of its estate away for a purpose that would enhance the value of the remainder" (Rae 1944, pp. 169). In reality, the intent of the land grant was not to fully fund the canal, but rather to "give some aid and encouragement to enterprises that were considered to be of national importance" (Rae 1944, pp. 177).

Canals provided a direct connection between markets in the east and the raw material in the west, stimulated migration, and set the foundation for further western expansion. However, the presence of the canals also opened up existing nearby lands that were inaccessible and/or wilderness (Bogart 1913). In 1845, before the completion of the Miami and Erie Canal, "not a single bushel of grain nor a single barrel of flour or pork was exported," but by 1846, "over 125,000 barrels of flour and almost 2M bushels of grain were sent through the canal to the northern market" (Bogart 1913, pp. 58–59). The first-growth forests of the region were harvested for lumber and sold in eastern markets. Whereas previously wooded areas were cleared by burning, the presence of canal boats made clearing the land and selling the lumber in eastern markets more profitable. In general, however, the development of a canal or rail system did not open up new land, but rather made existing land available where it had been previously inaccessible for commercial production (North 1956).

The development of major water infrastructure energized the US economy, expanded agricultural outputs, increased urbanization and industrialization, and ultimately impacted the standard of living of all Americans (Haines et al. 2003; Chanda et al. 2008). The aggregate of the impacts resulted in a complex, often counterintuitive, outcome. Throughout the nineteenth century, as the major water infrastructure of the nation was being developed, the human health of the nation was at first heterogeneous but eventually became more homogeneous, converging toward a steady state. Access to clean water and improved sanitation eventually increased personal hygiene and thus the quality of life; access to more varied food sources, at a lower cost, of a higher quality and the ability to refrigerate foodstuffs increased the quality of life. However, urbanization and industrialization had negative impacts on the quality of life as well. Improved transportation served as a vector for disease transmission, and as cities grew, with their inhabitants often working long hours inside factories, people "lived closer together and were exposed to a larger set of diseases" (Chanda et al. 2008, pp. 23). Haines et al. (2003, pp. 409) find that "being born in a county with water transport connections in 1840 also is consistent with both a contagion and a commercialization view. The result that farmers were taller and that laborers were shorter is also supportive of these rural-urban effects."

The Development of the Urban West: Mining to Agriculture

By the middle of the nineteenth century, although the mining boom in California was in full swing, agricultural development was still in its infancy, with many agricultural commodities still being shipped in from Europe and Asia (Pisani 1984). This would not last for long, as the skills that the miners acquired in building the vast mining waterworks that were used to extract precious metals from the mountain were eventually applied to building agricultural and municipal water infrastructure. Pisani (1984) notes that by 1867, the mining industry had constructed over 300 individual ditch systems across a total of 6,000 miles. Although many of these systems still remain, it wasn't the infrastructure itself that would drive agricultural development in California, but rather the skills that were developed – the ability to design and construct irrigation works and water systems and the adaptation of mining tool fabrication processes to agricultural and industrial machinery.

Although the mining boom in California was fleeting, the growth of its agricultural industry would persist, starting with an agricultural boom in the 1860s and 1870s that mirrored the mining boom of the 1840s and 1850s. Pisani (1984) notes that by 1870 the number of miners in California had fallen from a peak of 83,000, 10 years earlier, to a mere 36,000. Conversely, over this same time span, the number of farmers more than doubled from 20,000 to 48,000, and the acreage of wheat farmland more than quadrupled. The growth in demand for agricultural products spurred a demand for the delivery of water resources. By the end of the 1860s, there was only a single canal, aqueduct, or dam in California; over the next decade, nearly two dozen more would be constructed (Pisani 1984). By the turn of the twenty-first century, California would be carpeted with over 1,200 major reservoirs, and almost every major river in the state would be dammed, some more than 14 times (Reisner 1986).

Migration west after the Civil War saw the number of farms, and the total population of California increase by nearly 50% throughout the 1870s. Whereas, in the early years of the agricultural boom, many farmers were planting forage crops and cereals, by the turn of the twentieth century, the plantings had changed – vegetables, grapes, citrus, and orchards were the norm, which in turn demanded additional water resources and infrastructure. By the late 1870s, only 200,000 acres of farmland in California were irrigated, but a mere 10 years later, the irrigated acreage had increased by 500% (Pisani 1984).

The Development of Agriculture and Water Infrastructure in the Arid West

Water rights laws and major water infrastructure in the arid western United States have long gone hand-in-hand and are reflective of the close relationship between the land and water. This land-water relationship has been an essential attribute of westward expansion since at least the middle of the nineteenth century, when most of the land in the western states and territories was still in the public domain and homesteading was a major policy emphasis (Hibbard 1924). Unfortunately, the

Homestead Act of 1862, which was well designed for the humid lands located east of the 100th meridian, did not fit well with the arid conditions of the west (Xu et al. 2014). The western agricultural landscape boasts some of the most arid and rugged terrain in the continental United States. Agricultural lands in the arid western United States exhibit tremendous heterogeneity in topography, climate, soil types, and water resources, and for much of the arid west, melting snowpack comprises a majority of the agricultural water supply that is available during the growing season (Brosnan 1918; Hansen et al. 2011, 2014). Unlike farming in the east, which could rely on rain-fed, seasonal irrigation, agriculture in the western United States is entirely reliant on man-made irrigation and therefore is much more subject to the influence of the water infrastructure and water rights institutions where it is located.

Thomas Jefferson defined the dominant focus of US land policy as small, family farms when he claimed: "The earth is given as a common stock for man to labor and live on...(T)he small landholders are the most precious part of the state" (Hibbard 1924). US land policy, which began with the Land Ordinances of 1785 and 1787 that called for the orderly, systematic distribution of federal property to private claimants, continued with the Homestead Acts and between 1862 and 1935 became a period of major federal land dispersal (Gates 1979; Robbins 1942). The Homestead Act of 1862 allowed any family head to claim federal land between 40 and 160 acres and to receive title after satisfying the continuous residence and improvement requirement. The Homestead Act was modified in 1909 to increase the claims of land to 320 acres and again in 1912 to change residency requirement to 3 years. Moreover, the required fees and commissions under the law were adjusted. Following the Homestead Act, the Timber Culture Act, which granted 160 acres if settlers planted 40 acres of trees, was repealed in 1891 (Gates 1979). Hansen and Libecap (2004a) argue that the Homestead Act, the most commonly used law to claim federal lands, worked well initially, east of the 98th meridian, as the "familiar conditions" in soil quality and sufficient precipitation allowed farmers to transplant existing farming practices and knowledge. However, conditions were quite different in the Great Plains (Hansen and Libecap 2004a). Although John Wesley Powell made a warning in his "Report on the Lands of the Arid Region of the United States" to Congress in 1879 that arid lands necessitated new agricultural methods and called for a minimum of 2,560-acre homesteads for "pastoral regions" and proposed bills to change federal policy, these were not considered, and no significant modifications to land policy were implemented. The strong sentiment for maintaining the emphasis on small farms remained as illustrated by the statement of Representative George W. Julian of Indiana:

> If our institutions are to be preserved, we must insist upon the policy of small farms, thrifty villages, compact settlements, free schools, and equality of political rights, instead of large estates, slovenly agriculture, wide-scattered settlements, popular ignorance and a pampered aristocracy lording it over the people. This is the overshadowing question of American politics. Worster (2002)

Libecap and Hansen (2002) explain that at the time, there wasn't any scientific or experiential knowledge to support Powell's claim. Their view is that the science

was inconclusive to support the politically controversial modifications in federal land policy. Thus, they analyze the weather information problem in the Great Plains and show that the absence of knowledge about the region's climate combined with the positive experience of homestead farms during wet periods, which corresponded with major settlement in the northern plains, led to optimism. The rise of folk theories to explain the weather "rain follows the plow," which held that precipitation would increase with cultivation, and the pseudoscientific prescriptions for farming practices "dry farming doctrine," which explained how the use of tillage would solve or alleviate the problems with drought, were readily accepted (Libecap and Hansen 2002).

A more accurate understanding of the region's climate and the implications for small farms only emerged when the 1917–1921 drought was particularly hard on farmers. Prior to the 1917–1921 drought, optimism among farmers and western politicians was common, and Powell's suggested distributions, which were 16 times the size of existing allocations, were considered extreme, and it was commonly believed that they would drastically reduce the number of farmers that could settle in the region. Politicians, however, wanted to increase the number of farmers (and hence the population) in their jurisdictions in order to encourage economic development and therefore supported the existing land policies. As Representative Thomas Patterson of Colorado explained: ". . .our agricultural lands. . . are limited, and the number of our population following agricultural pursuits must also be limited. But to have that number as great as possible, to swell it to its maximum," the 160-acre homestead must not be exceeded (Patterson 1879 as quoted in Hansen and Libecap 2004a). In addition, a greater population expedited the territorial progress to statehood and increased the chance of having more voting members in the US House of Representatives. Thus, western politicians, including the territorial delegates, strongly opposed any major changes in the size of plots, and the Homestead Act remained generally unmodified except for a few minor changes, including the 1909 Enlarged Homestead Act that increased the claims of land to 320 acres. These 320-acre plots were still far smaller than those suggested by Powell, and subsequent events would reveal that they also were too small for long-term survival on the Great Plains.

A more complete analysis of the origins of the Dust Bowl of the 1930s, one of the most severe environmental disasters in North America, links this severe drought and wind erosion to the prevalence of small farms created by the Homestead Act (Hansen and Libecap 2004b). This analysis emphasizes the important role that farm size played in contributing to severe wind erosion as well as inhibiting corrective action. Most farmers were not operating at productively efficient sizes to adopt wind erosion control, such as putting a part of cropland to fallow. Larger farms had greater shares of fallow, and erosion was more severe in those Great Plains counties where cultivation shares were greater (Hansen and Libecap 2004b). Soil conservation districts, which could be as large as 600,000 acres, helped coordinate erosion control as farmers within districts entered into regulatory contracts. These regulations, combined with the gradual consolidation of farms, led to changes in cultivation practices that better protected the soil. Thus, when droughts affected

the Great Plains in the 1950 and 1970s, the region was less vulnerable to wind erosion. Nace and Pluhowski (1965) report that while the 1950s drought was at least as severe as that of the 1930s, the wind erosion in the 1930s happened to be much more extensive and damaging.

The homesteading led to a decline in average farm size in the Great Plains as the settlers claimed and subdivided the land that had been previously occupied by very large ranches. Under the federal land laws, ranchers could not obtain title to these large plots, and many were broken up as the homesteaders arrived (Libecap 1981; Dennen 1976; Fletcher 1960). For example, in Fergus County, Montana, in 1904, prior to major homestead migration to the northern plains, there were 472 farms or ranches with an average size of 1,300 acres. By 1916, however, the number of farms had grown by over eightfold to 3,843, and average farm size had fallen to 322 acres, a decline of 75% (Libecap and Hansen 2002).

Indeed, homestead settlement led to the proliferation of small farms throughout the Great Plains. Over one million original homestead entries were filed for 202,298,425 acres in western Kansas, Nebraska, the Dakotas, eastern Colorado, and Montana between 1880 and 1925 (U.S. Department of Interior, General Land Office, Annual Reports of the Commissioner). Most farms were 160 acres, although the average homestead sizes larger than that were due to the 1909, 320-acre Homestead law. The prevalence of 160-acre and 320-acre farms was due largely to the model and restrictions of the Homestead Act. Indeed, the dry farming techniques prescribed by the USDA extension service and experiment stations from the region's land grant colleges prior to 1917 called for intense cultivation of small farms of 160 or 320 acres (Libecap and Hansen 2002).

As mining opportunities and homesteading opportunities drew settlers west, those same miners and the earlier Mormon settlers from Utah (Coman 1911; Xu et al. 2014) heavily influenced the agricultural landscape in much of the arid western United States. In the arid northwestern United States for much of the nineteenth century, the only convenient water sources were located in close proximity to the riparian corridors; therefore early settlements rarely veered far from the major waterways. With the development of the Oregon Short Line Railroad in the 1880s, and major water infrastructure developments in the early twentieth century, settlements began to extend further away from the riparian corridors (Brosnan 1918).

A number of federal laws including the Carey Act (1894) and the Reclamation Act (1902) promoted the major water infrastructure projects in the arid western United States. The Carey Act encouraged private investment in water infrastructure and allowed for the private capture of profits from water sales (Xu et al. 2014). Unlike the Carey Act, which focused on private investments in major water infrastructure, the Reclamation Act provided federally subsidized funding for major water infrastructure projects in the arid western states and required the establishment of local water supply organizations to govern the use and distribution of water resources. These major water infrastructure projects profoundly transformed the agricultural landscape of the western states (Coman 1911; Hansen et al. 2011).

All told, the Reclamation Act authorized the development of hundreds of major water infrastructure projects, which were spread across the seventeen western states. With tens of billions in federal investment, Reclamation Act projects provide water resources to industrial, agricultural, and domestic users and hydroelectric resources for millions of those same users (Pisani 2002; Hansen et al. 2011). One of the primary goals of reclamation was to transform water resources into commodities that could be bought and sold (Pisani 2002). While the primary motivation for the Reclamation Act was to provide agricultural water, the water infrastructure provided a number of secondary benefits, including flood, tailings and debris control, fire protection, recreation, and navigation, among others (Hansen et al. 2011). Many of the projects that the Reclamation Act funded were financed through a Reclamation Fund, which was in turn financed through a cost-sharing agreement between the federal government and the local water supply organizations and through the sale of public lands (Pisani 2002). Immediately after the passing of the Reclamation Act, the federal government set aside some 40 million acres of public lands from use – including 1.5 million acres in Colorado, 2.7 million acres in California, 3.7 million acres in Idaho, 4.4 million acres in Nevada, and 8.5 million acres in Montana (Pisani 2002). The land that was "locked up" by the federal government through the Reclamation Act included many of the best reservoir sites and much of the publicly owned acreage that adjoined streams (Pisani 2002).

Over time, Hansen et al. (2011) find that the investments made by the federal government in the arid western states produced dividends. Those arid western counties with major water infrastructure were better able to deal with climatic variability, having more predictable agricultural production and fewer crop failures. The presence of major water infrastructure was particularly valuable during the regular drought or flooding events, when farms with access to major water infrastructure planted more acreage, of a higher value, with a greater harvest and fewer losses, than farms without access to a stable supply of water (Hansen et al. 2011, 2014).

Prior to the development of major water infrastructure in the arid western United States, conflicts between water users were common, and given the limits of in-stream supplies, demand often outpaced supply (Coman 1911). The riparian-based water laws that worked in the humid east, where water supplies were more plentiful and rarely were a binding constraint, did not work in the arid west. As settlements increased, population levels expanded, and large-scale agricultural practices replaced subsistence farming, a different rule of water law was needed. At the state level, in much of the arid west, appropriative-based water laws and rights were developed in order to establish and enforce the general rules of water use and to provide a mechanism for water distribution and ownership (Xu et al. 2014). These state-level rules of water law, many of which were being established at the same time that the territories were transitioning into statehood, and at a time when much of the nonirrigated land was still in the public domain, were profoundly influenced by earlier legislation. For example, the Desert Land Act (1877) set the stage for the wording of future appropriation legislation by requiring the identification of beneficial use and the documentation of the first date of use or prior appropriation

(Xu et al. 2014). However, because of the nature of the water right, which often includes a place of use and a point of diversion from the riparian corridor, these water rights introduce barriers to trade that can limit the transferability of water and therefore may result in economic inefficiencies (Xu et al. 2014).

Water is a scarce resource in most regions, which is especially true in the western United States. The economic development of the west mostly followed the possession of and the ability to use water. Competition and conflict over water was common both for American Indians and non-Indian settlers (McCool 1994). Although the case of *Winters v. United States* (also known as the Reserved Rights Doctrine) of 1908 established that the federal government implicitly reserved the water rights of reservation lands, western states adopted water codes and policies that allocated water rights on the basis of priority of beneficial use (McCool 1994). Prior appropriation – first in time, first in right principle, which was also used to settle disputes in many natural resources, such as grazing rights and land use – became the legal principle applied to water use in the arid west, defining the development of the agriculture in the region (Pisani 1996). It has been argued that without prior appropriation, the capital needed to build dams and irrigation canals to transform the west into an agricultural region could not have been raised. Although the Reclamation Act of 1902 gave the authority and job of building irrigation projects to the federal government, private capital was used to reclaim most of the irrigated land in the west, and only about one in four irrigated acres used water provided by federally developed projects in the 1980s (Pisani 1996). Moreover, although westerners might have expected the federal government to foster the development of major irrigation projects, most (miners in California, western politicians, developers, and farmers) favored local control over water, where the administration of water rights were best left to individuals organized in irrigation districts (Pisani 1996; Worster 1985).

Utilization of prior appropriation in the western water allocation should not be considered only as a by-product, or the result, of its arid climate, as was first asserted by Walter Prescott Webb, since economic needs and conditions of the region were instrumental in shaping the water rights institutions in the western United States (Dunbar 1983; Webb 1931). Aridity in the west was a crucial limiting factor in the development of western agriculture, but the principle of prior appropriation was first considered in court cases regarding water use in mining in California. In mid-nineteenth-century California, prior appropriation allowed miners on the public lands to have the priority rights both in mineral claims and water with first possession or beneficial use (Pisani 1996). Thus, in some western states and territories, including California and Montana, the principle of prior appropriation was not intended to apply to agricultural lands initially. By the time the Desert Land Act was enacted in 1877, however, lawmakers formally recognized the rights of settlers to use water through the public lands for agriculture and other uses under the prior appropriation rules and procedures.

The Reclamation Act of 1902 was the culmination of a long political struggle to have the federal government help build major water infrastructures in the arid west. Although many westerners wanted federal funds for the irrigation projects, they did

not want the federal control. The result was an act that was a compromise of federal funds and state control with prior appropriation laws. The reclamation iron triangle began to develop in 1889 with the creation of the Committee on Irrigation of Arid Lands in both congress and the states, even before the Act of 1902 (McCool 1994). This committee was made up of mostly pro-irrigation westerners well known for their regional emphasis and bias. Reclamation interests were also successful in sponsoring legislation through the Appropriations Committee, especially due to the chairmanship of Senator Carl Hayden of Arizona, who began serving on the Committee in 1928, became the chair in 1953, and served as the chair until his retirement in 1969 (McCool 1994). Among the many reclamation projects that Senator Hayden won for the west was the Central Arizona Project.

The decision-making process explaining the background and building of the Teton Dam in Idaho is a memorable example of the iron triangle of stakeholders, including interest groups, congressmen, and agencies working together as described by Engstrom (1976) soon after its collapse. To gain support for the Teton Dam project, its major advocates (the Bureau of Reclamation and the Fremont-Madison Irrigation District, represented by Mr. Willis Walker,) provided a significant quantity of favorable expert testimony to Congress and congressional committees and pushed for the building of the dam. These advocates wanted a multipurpose dam with flood, irrigation, and hydroelectric power production purposes. Although flood control was not the major purpose of the dam, the floods in the spring of 1962 and 1963 helped establish the need for a dam on the Teton River. The Army Corps of Engineers, which was also interested in the construction of a dam on the river, however, ruled it out after studying the feasibility of a levee system to help control the spring floods in 1955 (Engstrom 1976).

Since many farmers had "natural flow" water rights on the Teton River, the upstream users could only divert the surplus water after priority water rights were used. Thus, when the dam was built, it could not legally be filled with water unless the runoff was high. Thus, a plan was developed to put in a system of underground wells and pump water from the ground to the surface and into the river to meet the downstream water requirements. The plan called for 100 wells, with an estimated cost of $3,680,000 (Engstrom 1976). The storage was to provide supplemental irrigation water for approximately 114,000 acres in the Fremont-Madison Irrigation District. Phase II of the project was to provide water to 37,000 new acres of land carried by a 30-mile pump canal and 28-mile gravity canal. However, these benefits likely were overestimates, since 3 to 3.5 acre-feet of water could be sufficient to produce crops in the region and 87,000 of 111,000 acres already received an average of 11 acre-feet of water per acre of irrigated land (Engstrom 1976). Unfortunately, after the construction of Phase I was completed early in 1976, the dam was about 70% filled when the north side ruptured and collapsed, flooding the valley all the way to the American Falls Reservoir, killing 11 people and injuring hundreds.

As major water storage infrastructure was developed in the arid west, the water rights institutions within which they operated, in some cases, exacerbated the water distribution situation. For example, many of the storage water rights that are maintained at a great distance from the farms that utilize the water resources utilize

existing riparian corridors to deliver the water resources. Therefore, a detailed knowledge of the river dynamics, taking into account losses from evaporation and groundwater recharge in addition to the ability to translate volumes of storage into dynamic flow measures, is necessary. Similarly, the electrification of the farm and the ability to tap into deepwater wells (which can be hydrologically connected to surface water sources) require a unified water rights system (Xu et al. 2014).

Although the direct (consumptive) impacts of the development of major water infrastructure in the arid western United States tended to be positive, the nonmarket impacts were most certainly negative. The presence of large storage dams, which enabled the spatial and temporal transfer of water resources across seasons, resulted in reductions in water available for ecosystem use and increased the intra-seasonal volatility in water deliveries (Hansen et al. 2014). Particularly in drought years, many streams and rivers in the arid western United States are dewatered for much of the year, reducing the opportunity for anadromous fish migration; this is particularly true for rivers in which the main channel is dammed. These rivers have seen major declines in wild fish stocks throughout the twentieth century (Hansen et al. 2014).

Rhode (1995) analyzes both the demand and supply forces that influenced the intensification of California agriculture from 1890 to 1914. There is a range of explanations for the transformation of California agriculture from extensive to intensive products, such as fruits. These explanations include improvements and changes in transportation with the completion of the transcontinental railroad, the spread of irrigation, and changes in labor market conditions with the increased availability of labor. Rhode, in his systematic investigation of economic forces to explain the shifts in California agriculture, finds that traditional literature overstates the relative importance of the transcontinental railroad in the transformation of California agriculture and instead finds falling interest rates due to decreasing scarcity of capital as well as advances in biological knowledge that resulted in increases in productivity to be relatively important supply-side forces in this transformation.

The Development of the Urban West: Agriculture to Urban Growth

Whereas canals and major water infrastructure in the eastern part of the United States used water resources as a capital input more akin to a technological input, for users in the arid west, water was a raw material input to production. In most cases, and in the whole of the arid western United States, water resources are the constraining input – there has always been far more land available to be irrigated than there is irrigation water available. Similarly, for many of the major metropolitan areas in the arid western United States, major water infrastructure enabled population growth and provides the water needed for urban, commercial, and industrial development. Examples include the greater Sacramento, Los Angeles, San Diego, Las Vegas, and

Phoenix areas, which are home to nearly 10% of the total US population, and all of which receive less than 15 in. of precipitation per year.

The practice of importing vast quantities of water from distant sources in order to meet the consumptive demands of a burgeoning urban population was pioneered several millennia earlier by the Romans. Even in the arid southwestern United States, as recently as 1400 AD, the Hohokam civilization moved water dozens of miles in order to irrigate the desert (Reisner 1986). However, unlike the Romans and Hohokam, who relied on gravity to move the water from a relatively nearby source to site, many of the aqueducts that supply water in the arid western United States rely on costly pumping plants and move the water resources across elevational gradients, over many hundreds of miles. The City of Los Angeles "pioneered" the practice of importing water from a great distance – drawing water from the Colorado River, San Joaquin Delta, and Owens Valley. The Metropolitan Water District of Southern California, founded in 1928, serves more people, over a broader area, than any other water district in the United States. Without the availability of imported water, the existing local sources would only meet the needs of a small fraction of the total population (Erie and Joassart-Marcelli 2000).

In September of 1905, the City of Los Angeles issued and approved a $23 M (approximately $500 M in 2018 dollars) bond to build an aqueduct to deliver water to Los Angeles from Owens Valley, some 223 miles away (Reisner 1986). At the time, Owens Valley, which was agriculturally dominated, had vast water resources but very little agricultural potential – the volume of arable land was limited, the soil quality was poor, the growing season was short, and the cost of transporting agricultural commodities to market was steep (Libecap 2009). The aqueduct itself would take 6 years and upward of 6,000 people to build and in the process would result in 53 miles of tunnels, 120 miles of railroad track, 170 miles of electric transmission lines, and 500 miles of roads and trails (Reisner 1986). Surprisingly, in the early years, the aqueduct produced very little water for Los Angeles, with the lion's share being delivered to higher-valued agricultural fields in the San Fernando Valley instead (Reisner 1986). In terms of total volume, by 1920 the Owens Valley watershed, via the Los Angeles Aqueduct, supplied four times as much water to Los Angeles as the Los Angeles River supplied (Libecap 2009). By 1930 (from 1900), property values in Los Angeles County had increased by 4408% but by only 917% in Inyo County, home to Owens Valley (Libecap 2009). This disparity is reflected in the current marginal value of water for agricultural uses ($15–$25 per acre-foot) when compared to urban uses (over $500 per acre-foot) (Libecap 2009).

Further north in the San Francisco Bay Area, the San Francisco Board of Supervisors, having already experienced a massive population surge during the mining boom of the mid-1800s, was interested in securing a long-range water source to meet the growing needs of the city (Starr 1996). This need was only amplified after the 1906 earthquake and major fire, which destroyed much of the city and led to calls for a high-pressure water system to prevent any future recurrences (Starr 1996). Over 150 miles away in the Sierra Nevada Mountains and, more significantly, entirely within the confines of Yosemite National Park, the supervisors identified the

Hetch Hetchy Valley as the best site to build a reservoir on the Tuolumne River. After years of politicking, in 1923 the O'Shaughnessy Dam was completed – the second highest dam in the United States. It required 500 men and 3 years to complete (Starr 1996). Eleven years and $100 M later, the Tuolumne River waters arrived in San Francisco. All told, the Hetch Hetchy complex would encompass five major reservoirs, four dams, several hydroelectric facilities, and hundreds of miles of tunnels and pipeline (Starr 1996).

Until 1941, the water resources delivered to Los Angeles from the Owens Valley were the only sources of imported water (Libecap 2009). With a growing population, alternative sources were needed. The Colorado River Aqueduct and Storage Project, a $220 M project funded largely by property taxes, was completed in 1941. It supplied the greater Los Angeles area with two thirds of the power that it produced and over four million acre-feet of water (Reisner 1986). Unlike the earlier Owens Valley project, which was gravity fed, Colorado River water brought with it costly pumping requirements due to the need to move the water over great mountain ranges and across great distances (Erie and Joassart-Marcelli 2000). Hoover Dam "had been built to safeguard the future of the entire Southwest" and served as an example to all of the other states and countries that aspired to build truly massive water storage dams. The Colorado River, which the Hoover impedes, is not a huge river, ranking just outside of the top 25 in the United States in terms of total flow (Reisner 1986, pp. 257).

The California Water Plan, which had its genesis in the 1950s, presented a proposal for what would become the largest water project ever built by a state or local government (Reisner 1986). Previous water systems, such as New York's Catskill Aqueduct or the Delaware Aqueduct System, which was completed during the Second World War and included 85 miles of underground tunnels, would be dwarfed in comparison with what was proposed in the California Water Plan. The California Plan, when compared to the Catskill Aqueduct that delivered water to New York City, would deliver four times the water over six times the distance (Reisner 1986). Unlike the Colorado River Aqueduct, the State Water Project, which includes the California Aqueduct and a number of smaller branch aqueducts and canals, was funded by water sales, as opposed to taxes (Erie and Joassart-Marcelli 2000). This is primarily because the water collection and distribution network is spread across the state, with dozens of dams and hundreds of miles of canals, tunnels, and pipelines.

The Electrification of the City and Farm

In the early 1930s, the Pacific Northwest region of the United States only had three million residents, over half of whom lived in rural communities. Of the rural population, over 70% had no access to electricity (Reisner 1986). Reisner (1986) notes that prior to 1933 the Columbia River didn't possess a single dam; by the mid-1970s, the main stem of the Columbia and its tributaries possessed 36 great dams, including 13 "tremendous dams" such as the Grand Coulee, Bonneville, John

Day, and Hells Canyon. As the Grand Coulee filled in the early 1930s, very little of its available hydroelectric power was utilized, which was due largely to the construction of the Bonneville Dam downriver (Reisner 1986). However, by the early 1940s, with wars raging in the Pacific and European theaters, over 90% of the hydroelectric power being produced by the Grand Coulee and Bonneville was going toward defense industries and processing the aluminum needed for weapons and airplanes (Reisner 1986). Reisner (1986, pp. 164) noted "the Axis powers were no match for two things: the Russian winters, and an American hydroelectric capacity that could turn out sixty thousand aircraft in four years." Later, the Grand Coulee would provide inexpensive power to the burgeoning aerospace industry and would help to provide the electricity that was needed to develop the Hanford nuclear production facilities. To say that the Grand Coulee was a large dam was off considerably; according to Reisner (1986, pp. 159) "Hoover was big; Shasta was half again as big; Grand Coulee was bigger than both of them together." The Grand Coulee was the largest dam in the world, both in the total mass of 10.5 million cubic yards of concrete and the crest length of nearly a mile, requiring 130 million board feet of lumber to build. The city that arose to support the workers had more bars and brothels within a 5-mile radius than any other area in the world (Reisner 1986).

The impact of the Rural Electrification Administration (REA), in tandem with the availability of relatively inexpensive hydroelectric power, brought electricity to rural farms and significantly increased crop productivity, output, and land values (Kitchens and Fishback 2015). The dams themselves not only provided flood control and water for irrigation but also increased the economic potential of farms at a distance from riparian water sources by providing electricity that could be used to pump groundwater and process agricultural commodities. Pisani (2002, pp. 204) estimates that the electrification of the farm "added 10 million acres of bench and mesa land into the West's supply of irrigable land." It is difficult to disentangle the REA benefits from the benefits that arose from the major water infrastructure complexes (Kitchens 2014). From the early years of the Reclamation Act, hydroelectric power was seen as a resource that would "revive and expand old industries as well as create new ones" (Pisani 2002, pp. 203). Given the sheer landmass of the arid west, its scattered population, and rugged terrain, a rail system built on electricity as opposed to coal made tremendous sense. Electric rail was believed to be "cheaper, faster, more efficient, and less vulnerable to cold weather and mechanical breakdown than steam-powered trains" (Pisani 2002, pp. 204).

By the early 1940s, many of the nation's largest hydroelectric dams and complexes were completed – including the Grand Coulee, Hoover, and the Wilson and Norris Dams under the Tennessee Valley Authority (TVA). In the early 1930s, farming accounted for only 2.6% of all electricity consumed; by 1940, rural farm electrification had increased by 230% (Kitchens and Fishback 2015). As with most hydroelectric projects, the areas that receive the hydroelectric-produced electricity have some of the lowest, if not the lowest, electricity rates in the country (Kitchens 2014). This is certainly true of the TVA, which included a series of canals, roads, flood control systems, reservoirs, and dams (currently 29 hydroelectric dams) that provided hydroelectricity to many farms and homes that previously didn't have

access to electricity, as well as a broad array of co-benefits such as flood control and navigation improvements (Kline and Moretti 2014; Kitchens 2014). The TVA provided electricity to seven southeastern states, and although the short-term, aggregate economic impacts from the electrification that the TVA did provide, relative to its costs, aren't substantial, the aggregate infrastructure investments, flood protection, and transportation safety resulted in "the model intervention for any nation that sought to develop its water resources" (Kitchens 2014, pp. 390). The fact that the TVA was a place-based investment raises the fear that any localized benefits may be offset by losses elsewhere in the United States (Kline and Moretti 2014). However, over the long run, Kline and Moretti (2014) found that the net present value of the benefits from the TVA is on the order of $6.5–19.2 billion, with notable industrialization and the creation of manufacturing employment opportunities and high-paying manufacturing jobs. Overall, there were minimal indirect effects from the TVA, but the lion's share of the national direct impact, which was an increase in productivity of the domestic manufacturing sector by a third of a percent between 1940 and 1960, came directly from the investments in public infrastructure (Kline and Moretti 2014).

Concluding Thoughts

Water and water management are complex economic resources that have resulted in equally complex physical infrastructure, institutions, and governance. In the history of the US west and western migration, we argue that no natural resource has impacted the economy more than water and the infrastructure and governance that it promoted. The major water infrastructures of the west – dams, canals, and aqueducts – that crisscross the arid land have facilitated this migration by storing, elevating, and transporting water resources. Water in the arid west is *the* constraining input to growth, so major water works and infrastructure were needed in order to provide storage and to allow humankind to reclaim the desert. Federal legislation, including the Homestead and Reclamation Acts, encouraged and facilitated this migration. The institutions that developed in tandem with mining, agricultural, and urban water demand in the arid west bore little resemblance to the riparian water governance of the east. As cities grew in the arid west, urban demands increased, and technological innovations allowed the vast stored water resources to simultaneously create hydroelectric power, which fueled an industrial renaissance and electrified the farm.

In the age of heightened climate variability, major water infrastructure continues to play an important role in mitigating the potential impacts of floods and droughts (Hansen et al. 2011). Greater historical perspective, strengthened by the quantitative approach to economic history, i.e., cliometric research, has been essential in providing insights for the current debate about the effects of climate change and climate variability and how best to respond to it. The expansion of agriculture west of the 100th meridian during nineteenth and twentieth centuries in North America encountered climatic variability that was not predicted or previously experienced

(Olmstead and Rhode 2008). Thus, historical analyses of how those variable climatic conditions were addressed in the past can and will provide valuable information for addressing heightened climatic variability and it impacts. For example, Hansen et al. (2011) show the significant impact of water infrastructure and water management on crop mixes, fallow practices, and agricultural production in the western United States using large-scale, comprehensive data, such as county-level water infrastructure data that is spatially linked to topographic characteristics, historical climate data, and historical agricultural data during the twentieth century. In addition to examining the agricultural land use and crop productivity benefits of water storage infrastructure, the literature has addressed some of the potential ecosystem impacts, such as low levels of stream flow and water supply variability that may have been exacerbated by the water supply infrastructure and water rights governance in Idaho using cliometric techniques and exploiting the long-term patterns of water transfers between cold and warm seasons (Hansen et al. 2014). This more comprehensive and quantitative way to evaluate the long-term impacts of water infrastructure development and its management on agriculture and natural ecosystems in the semiarid regions will continue to help illuminate the debate on the trade-off between the agricultural benefits and ecological impacts and lead to a set of more balanced policies in the future (Hansen et al. 2011). Comparative analyses of crop mixes and agricultural yields, hedonic analyses of the impact of water infrastructure on property values and recreational use values, and estimates of the value of the hydroelectric power that dams provide do not capture their true impacts, and such analyses require better understanding and the incorporation of institutional responses and more rigorous econometric analysis with a longer time frame (Libecap 2011; Hansen et al. 2011). The institutions that emerged in response to the agricultural needs in semiarid regions of the United States are still here today and play a key role in today's water markets by raising the costs of reallocating water to higher-valued uses (Libecap 2011). Prior appropriation and governance of water rights systems provided framework for water allocation, use, and investment, and this framework continues to impact the contemporary use and allocation of water in response to new urbanization and environmental, industrial, and agricultural demands (Leonard and Libecap 2017).

The hydroelectric power enabled by the major water infrastructure provides a renewable and sustainable, carbon-neutral, inexpensive electricity production technology when compared to nonrenewable coal- or natural gas-based electricity technologies. Over a third of the US population lives in states at, or west of, the 100th meridian that also comprise a majority of the food-producing capacity in the United States. However, whereas in the past the focus was on constructing major water storage and transportation infrastructure, looking forward, much of the focus has been on decommissioning and removing this same infrastructure. Environmental awareness and a returned focus on the in-stream values of wild-flowing water, including recreational uses, first foods, and the survival of endangered and threatened native species, have elevated this conversation. Far more dams have been decommissioned than constructed since the last major dams were proposed in the late 1960s and early 1970s (Pisani 2002). In fact, much of the new major water

infrastructure that has been proposed is focused more on rehabilitating the aging inventory of major dams, adding hydroelectric production capabilities to those that lack it, moving water resources away from the major dams, providing alternatives for migrating fish species, and raising the height of existing dams so as to provide storage for additional water resources. As climate-driven conversations emphasize the need for sustainable energy resources, and nonconsumptive uses exert more influence on water management decisions in the arid western United States, the contributions of cliometricians, in an attempt to better understand how we have arrived at the constrained situations that we find ourselves in today, are essential components of the discussion.

Cross-References

- ▶ Agricliometrics and Agricultural Change in the Nineteenth and Twentieth Centuries
- ▶ Institutions
- ▶ Property Rights to Frontier Land and Minerals: US Exceptionalism
- ▶ Railroads
- ▶ The Great Depression in the United States

References

Bogart EL (1913) Early canal traffic and railroad competition in Ohio. J Polit Econ 21(1):56–70. *JSTOR*, JSTOR. www.jstor.org/stable/1819852

Brosnan CJ (1918) History of the State of Idaho. C. Scribner's Sons, New York

Chanda A, Craig LA, Treme J (2008) Convergence (and divergence) in the biological standard of living in the USA, 1820–1900. Cliometrica 2:19. https://doi.org/10.1007/s11698-007-0009-1

Coman K (1911) Some unsettled problems of irrigation 1911. Am Econ Rev 101(1):36–48

Cranmer HJ (1960) Canal investment, 1815–1860. In: Trends in the American economy in the nineteenth century. NBER. Princeton University Press, Princeton, pp 547–570

Dennen RT (1976) Cattlemen's associations and property rights in land in the American West. Explor Econ Hist 13(4):423–436

Dunbar RG (1983) Forging new rights in western waters [Western States (USA)]. Lincoln: University of Nebraska Press

Engstrom J (1976) A policy of disaster: the decision to build the Teton Dam. Rendezvous 11(2):62–74

Erie SP, Joassart-Marcelli P (2000) Unraveling Southern California's Water/Growth Nexus: metropolitan water district policies and subsidies for suburban development, 1928–1996. Calif West Law Rev 36(2):Article 4. https://scholarlycommons.law.cwsl.edu/cwlr/vol36/iss2/4

Fletcher RH (1960) Free grass to fences: the Montana cattle range story. University Publishers, New York

Gates PW (1979) History of public land law development. Arno Press, New York, reprint of his 1968 volume prepared for the Public Land Law Review Commission, Washington, DC

Haines MR, Craig LA, Weiss T (2003) The short and the dead: nutrition, mortality, and the 'antebellum puzzle' in the United States. J Econ Hist 63(2):385–416

Hansen ZK, Libecap GD (2004a) The allocation of property rights to land: US land policy and farm failure in the northern great plains. Explor Econ Hist 41(2):103–129
Hansen ZK, Libecap GD (2004b) Small farms, externalities, and the Dust Bowl of the 1930s. J Polit Econ 112(3):665–694
Hansen Z, Libecap GD, Lowe SE (2011) Climate variability and water infrastructure: historical experience in Western United States. In: Libecap GD, Steckel RH (eds) The economics of climate change: adaptations past and present. University of Chicago Press, Chicago/London, pp 253–280
Hansen ZK, Lowe SE, Xu W (2014) Long-term impacts of major water storage facilities on agriculture and the natural environment: evidence from Idaho (U.S.). Ecol Econ 100:106–118. https://doi.org/10.1016/j.ecolecon.2014.01.015
Hibbard BH (1924) A history of the public land policies. Macmillan, New York
Kitchens C (2014) The role of publicly provided electricity in economic development: the experience of the Tennessee Valley Authority, 1929–1955. J Econ Hist 74(2):389–419. https://doi.org/10.1017/S0022050714000308
Kitchens C, Fishback P (2015) Flip the switch: the impact of the rural electrification administration 1935–1940. J Econ Hist 75(4):1161–1195. https://doi.org/10.1017/S0022050715001540
Kline P, Moretti E (2014) Local economic development, agglomeration economies, and the big push: 100 years of evidence from the Tennessee Valley Authority. Q J Econ 129(1):275–331. https://doi.org/10.1093/qje/qjt034
Leonard B, Libecap GD (2017) Collective action by contract: prior appropriation and the development of irrigation in the Western United States. NBER working paper 22185. http://www.nber.org/papers/w22185
Libecap GD (1981) Bureaucratic opposition to the assignment of property rights: overgrazing on the western range. J Econ Hist 41(1):151–158
Libecap GD (2009) Chinatown revisited: Owens Valley and Los Angeles-bargaining costs and fairness perceptions of the first major Water Rights Exchange. Journal Law Econ Org 25(2):311–338. Available at SSRN: https://ssrn.com/abstract=1476649 or https://doi.org/10.1093/jleo/ewn006
Libecap GD (2011) Institutional path dependence in climate adaptation: Coman's "some unsettled problems of irrigation". Am Econ Rev 101(1):64–80
Libecap G, Hansen Z (2002) "Rain follows the plow" and dryfarming doctrine: the climate information problem and homestead failure in the Upper Great Plains, 1890–1925. J Econ Hist 62(1):86–120
McCool D (1994) Command of the waters: iron triangles, federal water development, and Indian water. University of Arizona Press, Tucson
Nace RL, Pluhowski EJ (1965) Drought of the 1950's with special reference to the Mid-continent. U.S. geological survey water-supply paper no. 1804. Government Printing Office, Washington, DC
North DC (1956) International capital flows and the development of the American West. J Econ Hist 16(4):493–505. *JSTOR*, JSTOR. www.jstor.org/stable/2114694
Olmstead AL, Rhode PW (2008) Abundance: biological innovation and American agricultural development. Cambridge University Press, New York
Pisani DJ (1984) From the family farm to agribusiness: the irrigation crusade in California and the West, 1850–1931. University of California Press, Berkeley
Pisani DJ (1996) Water, land, and law in the West: the limits of public policy, 1850–1920. University Press of Kansas, Lawrence
Pisani DJ (2002) Water and American government: the Reclamation Bureau, national water policy, and the West, 1902–1935. University of California Press, Berkeley
Rae JB (1944) Federal land grants in aid of canals. J Econ Hist 4(2):167–177. *JSTOR*, JSTOR. www.jstor.org/stable/2113882
Ransom RL (1964) Canals and development: a discussion of the issues. Am Econ Rev 54(3):365–376. *JSTOR*, JSTOR. www.jstor.org/stable/1818521
Reisner M (1986) Cadillac desert: the American West and its disappearing water. Viking, New York

Representative Patterson. Appendix to the congressional record, 1879. In: 45th Congress, 3rd Session, p 221
Rhode P (1995) Learning, capital accumulation, and the transformation of California agriculture. J Econ Hist 55(4):773–800
Robbins RM (1942) Our landed heritage: the public domain, 1776–1936, Princeton: Princeton University Press
Starr K (1996) Endangered dreams: the Great Depression in California. Oxford University Press, New York
Turner FJ (1920) The frontier in American history. H. Holt and Company, New York
U.S. Department of Interior, General Land Office. Annual Reports of the Commissioner. Washington DC: GPO, various years (1880–1925)
Webb WP (1931) The Great Plains. Ginn and Company, Boston
Worster D (1985) Rivers of empire: Water, aridity, and the growth of the American West. New York: Pantheon Books
Worster D (2002) A river running west: the life of John Wesley Powell. Oxford University Press, New York/Oxford
Xu W, Lowe SE, Adams RM (2014) Climate change, water rights, and water supply: the case of irrigated agriculture in Idaho. Water Resour Res 50:9675–9695. https://doi.org/10.1002/2013WR014696.2

Part V

Money, Banking, and Finance

Early Capital Markets

Ann M. Carlos and Stephen Quinn

Contents

Abstract

Capital markets before 1750 created enduring ways to fund risky, long-term, large-scale investments. Demand for capital was driven by the expansion of long-distance trade and by conflict between fiscal states. Supply responded with experimentation in political structures, in legal environments, in how firms initially raised funds, and in how investors resold securities. Improvements in these structures allowed investors to better protect their claims against the information and incentive problems inherent in long-term commitments. As a result, more investors could and did participate. Notable innovations such as the

A. M. Carlos (✉)
Department of Economics, University of Colorado Boulder, Boulder, CO, USA
e-mail: ann.carlos@colorado.edu

S. Quinn
Department of Economics, Texas Christian University, Fort Worth, TX, USA
e-mail: S.Quinn@tcu.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_34

Dutch and English East India Companies and the British national debt had a strong influence on the subsequent era of industrialization by demonstrating how to arrange deep secondary markets for equities and bonds.

Keywords
Early modern capital markets · Equity and debt · Joint-Stock companies · Sovereign debt

Introduction

A capital market is a market where buyers and sellers trade in securities such as equities and bonds. Such markets help channel surplus funds from savers to institutions with a demand for long-term investment funds. The capital market consists of primary markets, which deal in the new issue of stocks and bonds, and secondary markets, which provide liquidity for those holding securities through the resale of existing securities. In this chapter we lay out the foundation of this market. Most firms before 1800 were small, and most "early capital" was self-funded within those small firms, but there were major exceptions that drove a demand for external and long-term funding. Examples are long-distance trade, transportation, war, and the military.

This chapter focuses on the development into what we consider today as the modern capital market. The chapter proceeds by defining terms and setting out the basic challenges facing that development. Successes, and failures, are then considered. The chapter concludes by noting how this development created funding technologies put to great use during the nineteenth century.

Concepts

Businesses, even small businesses, need some combination of land, buildings, infrastructure, ships, equipment, and raw materials. Such expenditures commit large sums for long periods with uncertain returns. In the context discussed in this chapter, capital refers to a long-term investment in a business where profits often take years to arrive and at levels hard to foresee. The simplest example is the sole proprietor nurturing a fledgling business with her own time and money. She alone is balancing a variety of incentives because she has invested her own money; she controls decisions, and she will receive eventual profits or losses. As the sole proprietor, she internalizes the nexus of firm-specific trade-offs. At the same time, her entrepreneurial capacities limit the scale of operation. Even if she has the skills to expand, she may prefer not to concentrate too much of her wealth into one endeavor. Capital markets begin when she invites others to make a long-term commitment to her enterprise (Guinnane et al. 2007).

For outsiders, capital investment in someone else's enterprise is perilous. The incumbent owner obviously knows more about the firm than investors do, and her goal might be to take advantage of the less well informed. To paraphrase the famous lemons problem, if the entrepreneur knows the enterprise has poor prospects, she might want to sell out, and she might misrepresent to do so (Akerlof 1970). If such behavior is common, then investors will assume most opportunities are poor, and, as a result, quality opportunities struggle to find funding on agreeable terms. So to encourage a market for capital investment in situations rife with asymmetric information, owners need ways to signal quality, and investors need ways to screen prospects. One method is to share active ownership though partnerships. The new partners gain access to inside information and participate in governance, while the original owner remains invested in ongoing success. Another method is for the owner to borrow. Lenders gain a schedule of interest payments and principal repayment while the owner retains profits. If the debt is not repaid, the lenders gain priority in a bankruptcy. Borrowing can signal firm quality because owners expecting profits might favor debt more than do owners of unprofitable firms (Ross 1977). Debt, however, leaves substantial information asymmetries that lenders can ameliorate through a variety of stratagems: limiting the amount lent, demanding collateral, rolling over short maturities, covenants that limit additional borrowing, etc. To navigate these issues, equity and debt can be arranged and combined in a variety of ways. The point is that the structure of long-term funding is a fundamental way capitalists find imperfect solutions to information and incentive problems.

We can think of capital market development occurring when new structures result in investors having to make less effort to protect their claims. If the sole proprietorship is our benchmark, then a publicly traded, limited liability, joint-stock company has capital investors with much reduced information and control. At the same time, such companies can raise far greater sums than the wealthiest individual can supply. Capital markets developed to allow the paradoxical reduction of investor engagement to coincide with greater capital accumulation. The proximate cause of innovation was a larger demand for capital. For example, long-distance trade to Asia required more money for longer periods than did nearer trade. The innovation that allowed this to happen was the organization of the market such that an investor did not need to know or control what ships or agents did on the other side of the world while being liable for no more than the capital she alone invested. As such, capital market innovation can be thought of as learning to supply imperfect substitutes for investor engagement and knowledge acquisition. These substitutes range from accounting procedures to boards of directors and from limited liability to secondary market liquidity. This chapter is about how markets did this. Markets here mean not just the act of investing but all the institutions, arrangements, and structures supporting the process.

Before proceeding, we must clarify that short-term funding is a different story. Businesses have a persistent need for funding to cover payroll, suppliers, inventory, goods in transit, and customer credit. These needs usually resolve within months and can be covered by ledger credit, letters of credit, bills obligatory, bills of exchange, overdrafts, discounts, promissory notes, commercial paper, repurchase agreements, and the like (Puttevils 2015; Santarosa 2015). These types of instruments are supplied

by firms along an information spectrum. Members of the supply chain lend based on transaction-specific information and repeated interactions. Specialty lenders, like merchant banks and commercial banks, blend industry-specific information with funding from outside investors. Finally, the money market comprises lenders who supply funds while knowing few specifics; outsiders invest because insiders carefully arrange the debt into (secure) instruments insulated from privileged information. Short maturities are the most common way to reduce the need for information. Examples of short-maturity instruments favored by outside investors are bank accounts, bills with multiple endorsements, and repurchase agreements with government debt as collateral (Quinn 2001; Quinn and Roberds 2014).

Gorton (2017) defines such safe assets as giving investors little incentive to produce their own information about the circumstances behind the debt. For example, knowing that 30-day commercial paper is rated AAA is sufficient justification for a money market mutual fund to make a purchase. Similarly, Holmström (2015) defines safe assets as having the quality of no-questions-asked. This means, for example, that a merchant accepts bank money in payment without questioning the balance sheet of the bank. The goal is to make investors collectively ignorant of specifics but confident that others will not take advantage of such ignorance. The result is short-term debt that pays relatively low rates of return and varies little, if at all, in price. The process of risk/price discovery has been carefully and intentionally minimized.

In contrast, financial capital (large-scale funding for long periods) is rarely a safe asset. By definition, equity represents firm-specific risk. The underlying business model needs to be remarkably stable for equity investors to not ask questions. Long-term debt is easier to make safe because it can be insulated by equity, but it still requires firms of sufficient scale to be well known, and it requires the firm to have sufficiently limited leverage in order for equity to be a credible insulator. The more common example of safe long-term debt is the bond of credible governments. While not-for-profit enterprises, governments certainly need large-scale funding for long periods. Investors must consider the political willingness to repay in addition to the fiscal ability, but this is a distinction of degree because politics looms over most all large-scale companies in the early modern era. The Bank of Saint George in Genoa, the Dutch and English East India Companies, and the Bank of England were each intertwined with their respective states.

This chapter focuses on early (pre-1750) examples of how markets found ways to supply capital from increasingly anonymous investors. The examples discussed are not strictly the first example of a particular innovation but the example that has had a lasting impact. The first steps are in primary (initial) investment such as adding active partners, adding sleeping partners, and, eventually, the joint-stock company whose development required secondary (resale) markets, an expansion that brings with it price-based markets, dealers, exchanges, financial newspapers, etc. Finally, the particulars of how a capital market developed (or failed to develop) vary with the sector in question. The chapter will focus on two sectors where demand reached sufficient scale to encourage supply-side innovation: trade and the fiscal state.

Long-Distance Trade

Trade is ubiquitous. It takes place within families when members specialize by task. It happens within communities when some specialize in products for sale or barter with other members of the community and it takes place between communities. By long-distance trading we mean the movement of goods over considerable distances and thus long period of time. Long-distance trades go back millennia with evidence coming, for example, from stone tablets and pre-Columbian archaeological records. Our interest here is not in trade itself but the intersection between trade, especially over long distances, the financing of that trade, and the development of capital markets.

The volume and extent of trade depend on relative supplies and demands for particular commodities which, in turn, are shown by the relative prices in the sending and receiving regions and the transportation costs of moving goods from the lower-price region (higher relative supply) to the higher-priced region (higher relative demand). High unit costs of transportation will limit trade to commodities with higher market value in the importing region. Costs of transportation do not merely capture the physical movement of goods by horses, camels, wagons, or boats but also costs of transshipment, taxes, and tariffs as the goods cross political borders, banditry, theft, and holdup by varied authorities. Thus, independent of the physical costs of moving goods, the political environment and security of legal rights affect the volume of trade; the more uncertain that environment, the lower the volume of trade.

Long-distance trades changed in scope and scale in the fifteenth century with the ocean voyages of the Age of Exploration. Explorers, European and Asian, made amazing voyages of discovery circumventing known continents, reaching continents previously unknown by Europeans and creating new sea routes for regions previously connected by land. These ocean voyages created new opportunities and new markets and changed the financial requirements for these activities. Ocean-going vessels were larger and more expensive than those that traveled along the coast or through the Baltic or Mediterranean; they held more cargo; they had to be armed; and the total voyage time round-trip could take 2–3 years. Each of these characteristics increased the financial requirements. As a result, these changes in the spatial operation of trade led to changes in the financing of the trades and developments in the market for capital.

Precursor Solutions

Concurrent with the increasing financial requirements of long-distance trade, long-distance trade itself posed challenges for the operation of the firm and thus for the financing of the firm stemming from reallocation of risk (Goetzman 2016). To understand the nature of these challenges and the solutions that emerged, we first discuss some of the precursor solutions that allowed for the allocation and reallocation of risk between owners and investors prior to the joint-stock company

and the emergence of the early capital markets. Although the discussion following pertains to the operation of the firm, we are not interested in the firm per se but rather in the structures the firm used to manage its organization and how they reflect the issues of information, monitoring, control, and profits important to external investors.

The majority of firms, today and historically, are sole proprietorships. A sole proprietorship combines information about the firm, monitoring within the firm, control over decisions, and profit/losses from the firm into one capital investor. Thus the size of any such firm is determined by the ability of that single proprietor. As a firm grows, or when the activities of the firm take place in two different locations, the ability of the proprietor to monitor activities diminishes substantially. When a sole proprietorship engages in foreign trade, the owner can travel with his or her goods to one single destination, leaving activities in the home base unattended or the merchant can hire an agent to conduct her business in the foreign location.

Organizing business spatially is complex. When a merchant hires an agent to conduct business on his behalf in a foreign location, he is assigning authority to that person to handle his business activities in that foreign location. The problem facing the merchant/owner is whether the agent will honor the contract and work in the merchant's best interest or cheat the merchant (Greif 1993). Because the agent's activities are taking place in a location removed by time and distance, the merchant will have only incomplete information about the agent's conduct. How does the merchant know that the sale price was low as the agent attests? How can the merchant know if the goods were damaged in transit as the agent attests? Incomplete information and imperfect monitoring make it very difficult for a merchant to verify the extent to which an agent is behaving in the best interest of the owner/firm or behaving opportunistically, because such incomplete information and imperfect monitoring make it easier for an agent to cheat (Carlos and Nicholas 1988, 1990). Given this, a merchant should never hire an agent, and such arm's length activity should not take place. Attenuating an agent's ability to behave opportunistically, or to reduce moral hazard, is essential for any merchant who engages in long-distance trade.

Here we describe just a number of possible solutions.[1] A famous example of merchants' actions to attenuate opportunistic conduct of its agents is the case of the Maghribi traders. Greif (1993) argues that Maghribi traders, who operated across the Mediterranean, handled the agency problem through what he terms a coalition, which laid out the expectations of behavior for agents and the consequences of cheating. Information flowing within the coalition enabled monitoring and the existence of the coalition allowed for a coordinated punishment strategy. Agents who cheated any one merchant faced collective punishment by all members of the coalition, meaning that no members would hire that agent, and if a merchant

[1] An obvious mechanism is the legal system. If an agent defrauds a merchant or does not uphold the conditions of the contract, the merchant can sue in court. Historically, however, while a court system might sometimes be available, issues of locational, legal, and religious jurisdiction could apply.

cheated, all members of the coalition were free to cheat on that merchant. Although solving the agency problem, such a coalition structure depended on the ability to coordinate and to ostracize. As outside opportunities rose, such community-based structures would fail (Greif 2006).

More common than coalitions just described is the use of partnerships, whereby individuals invest jointly in the firm and are jointly liable for any obligations. Partnerships are agreements that generally cease upon the death of any one partner or if a partner wishes to exit. There are examples of this form from Greco-Roman times to the present. In the thirteenth century, a variant emerged in the Western Mediterranean port cities – *commenda* contracts – that encouraged greater trade and mitigated opportunism by overseas agents. Essentially, the *commenda* was a profit sharing contract that limited an investor's losses to the investment made by the investor at the same time as rewarding the overseas agent with a share of the profits and a cap on losses. What came to distinguish this contract form from a partnership was the limiting of liability for the investor, allowing merchants to mobilize capital from anyone within the community, and allowing those with savings to diversify by investing in multiple *commenda* contracts, with the downside risk limited to the investment made. A constraint on the firm, however, was that each trip required its own *commenda* contract, in which all profits were distributed at the conclusion of the voyage (Harris 2009). While *commenda* contracts were used extensively in the Western Mediterranean, salaried agents remained common in English and Baltic trade with the attendant agency issues.

Similar to but differing from a coalition, individual merchants could organize themselves in associations with obligations and benefits. The guild was one such association, often formed to protect the interests of the particular group and enjoying certain privileges granted to it by the authorities (Ogilvie 2014). Of importance is the legal development that saw the guild itself as a legal entity with a perpetual life, independent of the particular members at a point in time. In England, the guild structure and distance trade came together with the Merchant Staplers incorporated in 1319 and the Company of Merchant Adventures of London incorporated in 1407 with a Royal Charter. Another example is the Russia Company (1555). These companies received a set of privileges from the government that could include the rights to trade in certain locations or designation as the sole importer or exporter of a particular commodity. The Hanseatic League in northern Europe is another example of a formal association of merchant guilds and towns. Within these organizations, however, individual merchants operated on their own accounts but did so under the umbrella of the organization rules and obligations. What the institution provided was a trade-related infrastructure of ships, warehouses, lodging houses, and legal protection in the main ports of the trade, which were financed by the dues or payments owed by the members of the organization. While this did not solve the agency problem, it did enhance the flow of information. The umbrella term for these organizations is the regulated company. Thus, while the merchants themselves continued to be sole proprietors or partnerships, the association was infinitely lived and had privileges in its own name.

Joint-Stock Companies

The new long-distance sea routes to Asia, Africa, and the Americas opened new opportunities for both merchants and European governments. Although these sea routes were yet another example of water-based trade, the scale and scope of these new markets were challenging. Capital requirements for ships, cargoes, and crews were greater than more local trade, and for investors the risk was unknown. Additionally, Spanish and Portuguese first-mover advantage in these markets added an additional layer of political complexity and potential for conflict. The Spanish and Portuguese managed their trade as royal/state-run monopolies excluding investors from the market (Rei 2011).

The first challenge facing merchants was the need to mobilize capital to cover the costs of sending a ship to Asia, India, the Americas, or Africa. In the 1590s, both Dutch and English merchants began to organize trade to the East Indies. In 1599, a group of English merchants petitioned the Crown for a charter to trade in the East Indies (Scott 1910). Issued in 1600, the charter was along the lines of a regulated company. It gave a group of insiders, who would be called the East India Company, monopoly rights over English trade to the region (defined in the charter) and monopoly right to import products from that region into England. Of course, the English company was competing with all other European companies and countries and with other English groups that could also be chartered by the Crown (Scott 1910). The actual investment made by these EIC merchants was in the voyage/ship and its cargo (not in the company) with liability limited to the amount invested. Profits depended on the costs of the voyage (cargo, wages, supplies for the voyage), the return of the ship, and the sale of the commodities. What must be remembered is that these voyages to and from China could take 2 years, during which there would be minimal information on the likely success of the voyage. These shares in the voyage were tradeable, but not very liquid, and their price uncertain.

From 1600 to 1659 the English East India Company organized its trade both by selling shares in single voyages and sometimes for multiple sets of voyages, some of which were profitable and some not at all. Indeed, in some years, the company was close to, or actually, insolvent (Scott 1910). A shortcoming of this method of financing was that any net profits had to be fully dispersed to the investors, which left the "company" seeking new investment for each subsequent voyage. In other words, there could be no retained earnings. Whether formally realized or by happenstance, the re-chartering of the company in 1659 provided for a permanent joint stock such that investors owned a share in the company. Each investor's liability was known: limited basically to the equity investment made or requirements laid out in the charter. The ownership of shares also created a political constituency in support of corporate trade rights (Jha 2015).

In Holland, merchants faced the same issues of distance and cost. There, however, the first voyages were full liability partnerships. There too the first ships that returned in 1597 failed to cover the costs of the expedition. As in England, the belief in potential profits encouraged further voyages. Whereas in England, the trade was organized under the umbrella charter of 1600, in Holland, the first years of the trade

were carried out by competing sets of merchants from the major cities. The failure of attempts at local coordination led to coordination at the federal level with the chartering of a public monopoly in 1602, the *Vereenigde Oost-Indische Compagnie* (VOC), or the Dutch East India Company. It was to be managed by the six local chambers which, in turn, appointed a governing board of seventeen delegates (the *Heeren* XVII) to manage the business. Each of the major cities was represented, and the capital stock was specified in the charter. Shareholders committed capital for 10 years, but the initial charter (set for 21 years) required a full-distribution in 1612, at which time shareholders could decide to pull out or reinvest. In this first decade the company faced financial constraints, which the distribution of assets in 1612 made difficult to resolve. Unlike the East India Company of this period, the VOC played a role as an arm of the government with forts, factories, and Asian shipping assets. How these complexities would be realized in 1612 led to a decision to give the VOC an indefinite life, but with the same fixed capital stock. While the shares in the VOC were traded from its outset, the constraint on the size of the capital stock limited trading potential (Dari-Mattiacci et al. 2017).

From the mid-sixteenth century, there was experimentation in the organization of English trade to Africa. Starting from partnerships with exclusive grants to land areas, to constructing forts on the coast, various groups pursued the Africa trade, which focused on gold, ivory, pepper, and redwood with the slave trade as a smaller part of many of the early voyages. While the length of these voyages was shorter, the trade was volatile. Depending on the ability to acquire goods, to evade Portuguese and other European vessels and privateers, the outcome could be highly profitable or the investors could experience sizeable losses. The result, as described by Scott (1910), was a series of experiments prior to the chartering of the Governor and Company of the Royal Adventurers of England trading into Africa (1662–1672), which went bankrupt within 10 years. Its charter was ended and the debts taken over and written down by the newly chartered Royal African Company of England in 1672. This company was a joint-stock company with a perpetual life and capital stock of £100,000 sold in shares of £100 face value. The charter included limited liability for the investors. Two years earlier in 1670, a charter had been issued to the Governor and Company of Adventurers of England trading into Hudson's Bay, also with shares of £100 face value and limited liability for investors.

From the middle of the sixteenth century, merchants experimented with different ways to mobilize capital for the long(er) distance trades – private trade, partnerships, quasi-regulated companies, and the joint-stock company. The joint-stock company sold transferable shares in a perpetually lived company with locked-in capital. By the middle of the seventeenth century, the joint-stock structure, whereby investors purchased shares in the enterprise, was widely used in many countries in Europe. Scott, in his three-volume work, documents the range of these companies across multiple sectors in England, Scotland, and Ireland. Those companies with a royal, or after 1689 a parliamentary, charter generally had limited liability and tended to be larger in terms of their capital stock. The charters laid out the organization of the company, the face value of a stock, and the level of shareholding needed to be an officer or serve on the company board. These were modern joint-stock organizations.

Secondary Stock Markets

The existence of a joint-stock company does not imply the existence of a secondary stock market. If each initial investor was a long-term passive investor, there would be no stock market because no investor would be selling. Stock markets contribute to economic growth because they provide a mechanism to mobilize savings; they provide liquidity and risk diversification; they allow for the acquisition of information. The existence of a stock market eases the tension between the needs of the firm for long-run financing, especially in the case of high-return projects, and an investor's demands for liquidity (Levine and Zervos 1996). The difficulties the East India Company faced during its early decades in trying to finance individual voyages or sets of voyages (as did other such long-distance trades) speak to this tension.

For early capital markets there needed to be a place or person where trading could occur. In the last quarter of the seventeenth century a number of local but interrelated markets for shares arose in London and in Amsterdam. One location was the company head office, where potential buyers and sellers could locate one another. Coffee houses, located in Exchange Alley and Lombard Street close to the Royal Exchange in London, became another focal point, especially Jonathan's and Garraway's.[2] Social standing, gender, and living outside the metropole circumscribed access to company houses or coffee shops. Well described by Anne Murphy (2009), the market was quickly served by an array of middlemen, brokers, and solicitors willing to act on behalf of individuals. Some of these agents could themselves be women (Laurence 2008). Regardless of the location or entity involved in the transaction, the transfer of a share had to be documented at the company office to legalize the transfer of the property right and also to ensure that the company would know to whom to pay a dividend.

For a capital market to function efficiently, investors need information. To the extent that markets work efficiently, information will be revealed through price and price changes. In the aftermath of the Glorious Revolution and the loosening of restrictions on the press, financial broadsheets, such as John Castaing's *Course of the Exchange* (1689–1823), emerged (Neal 1991). Such broadsheets listed the market prices of various financial assets. The *Course of the Exchange* was published on Tuesday and Friday, listing the price for that day and for the preceding 2 days. The periods when trading was closed, such as Sundays or prior to making a dividend, were also noted. These financial broadsheets were widely distributed not just in London but also mailed to interested parties in England and elsewhere, in particular, Amsterdam (Koudijs 2016; Carlos and Neal 2011). Trading prices were also listed on boards in the various coffee houses along Exchange Alley and at the company headquarters.

Shares, trading locations, price information, financial broadsheets, property rights, and limited liability made it possible for the long-term opportunities in the

[2]Indeed, the London Stock Exchange would be built on the site of Jonathan's coffee house.

overseas trades to access capital. Market liquidity made it possible for investors to invest in a company because they knew they could "quickly, cheaply, and confidently sell their stake" when needed (Levine and Zervos 1996: 327). The confluence of these forces provided a mechanism through which firms/institutions could access large amounts of capital from multiple small investors.

Fiscal State

In the early modern era technological advancement and increasing scale caused military expenditure to rise relentlessly (Hoffman 2015a). To pay for these costs fiscal states had to navigate the politics of creating permanent taxation (Hoffman 2015b). To expand the tax base, fiscal states struggled to centralize and standardize revenue authority and create political mechanisms to monitor expenditure (Dincecco 2015). The slowly emerging result was an increasingly large and credible stream of future revenues, with states using capital markets to turn these future expected streams into money for current warfare. Substantial government spending on other public goods, such as education and sanitation, did not arise until well after the Napoleonic era (Hoffman 2015b).

It was not, however, simple to convert expected taxes into current funding. Like any enterprise, a government might lack the funds to repay, or it might decide not to repay. For capital investments by firms, reliable enforcement of contracts through courts of law can compel the unwilling to pay. Investors thus focus on the ability to repay either through expectations of profit or bankruptcy. A sovereign, however, lacks a third party who can coerce repayment, so investors must decide if they trust the state to repay. The longer the time to repayment, the more time there is for expectations or ability or willingness to degrade. Unlocking future tax revenue for current uses requires a sovereign to convince investors that the future revenues are secure and that the commitment to share them is credible. In short, the whole arrangement is as much political as economic. The political being the decisions to assess taxes and then redistribute the funds to creditors.

Monarchies

The interplay of ability and willingness is evident in monarchical efforts to secure multiyear borrowing. The monarchs of early modern France and England typically had the right to collect a set of ordinary taxes but had limited administrative capacity by which to do so. A tax farm is when a sovereign sells the right to collect one of those taxes. The tax farmer pays an amount certain to the monarch for an amount variable that s/he can seek to collect. The sovereign can use bidding to ascertain how much a tax could supply, and the resulting contract informs perspective lenders of fiscal ability. The sovereign can signal an enhanced willingness to repay by pledging the farmed revenues to repay specific loans. The pre-assignment of revenue streams reduces room for the monarch to overcommit collateral. It also makes default more

painful through injury of the monarch's general credibility. Modifying a sovereign's willingness to repay works on the margins, because a sovereign can only be induced to repay rather than compelled to repay.

To further signal ability and willingness, a sovereign can combine debt collection with tax collection: a syndicate buys collection rights and lends in anticipation of those collections. The disciplinary innovation here is that the combined creditor-farmers can disrupt or threaten to disrupt both revenue collection and credit supply should the government default, so they are willing to lend more than they would have otherwise (Johnson 2006). In the early modern era, the monarchs of France and England moved to such "cabal" tax farms to fund the increasing cost of warfare (Johnson and Koyama 2014). The monarchs had to use tax farm syndicates to gain multiyear credit because their unsecured pledges were insufficient; the alliance of tax collection and anticipatory lending eased the rationing of credit only as far as the threat could be maintained. And the threat was needed. For example, from 1598 to 1655, French kings broke the leases on two-thirds of tax farms (Johnson 2006). Moreover, the arrangement's capacity to lend increased with scale, so both France and England invested in the centralization of revenue administration. For example, Louis XIV merged tax farms into the General Farms in 1681, a quasi-corporation that routinely renegotiated terms with kings from a defendable bargaining position of being "too big to fail" (Johnson 2006).

In both nations farm syndicates were a result of monarchs doing their best to borrow independently of their representative bodies. For example, when French kings summoned the Estates General in 1576 and in 1789, the assemblies did not support new taxes because of opposition to redistribution and because giving the king what he wanted removed the need for the assembly (Stasavage 2015; White 1995; Hoffman et al. 2000). In contrast, with the restoration agreement of Charles II in 1660, parliament became willing to supply the monarch with extraordinary taxes for limited periods. As a result, Charles ended the two great farms in 1671 (customs) and 1683 (excise). The arrangement became more explicit after the Glorious Revolution of 1688. Among other issues, parliament agreed to fund the king's war in exchange for the powers of routine assembly and routine reauthorization of the king's power to collect taxes (North and Weingast 1989). In these negotiations, sovereign debt both redistributes from tax payers to creditors and is a key determinant of political power. In England, assemblies, and the Crown eventually found a compromise, and farm syndicates ended. In France, they did not.

Spain avoided tax farms becoming loan syndicates. When the kingdom's assembly (Cortes) agreed to make a tax permanent, the king could pledge the stream of revenue to fund an annuity. While sometimes tax farms did collect these taxes, the farms did not control lending on those revenues. Instead, the annuities were sold to the nobility, and the nobility formed the assembly, so repayment was credible. Also, the Cortes resisted creating new permanent taxes because withholding such authorization was the assembly's primary source of negotiation power (Drelichman and Voth 2014). Unlike other European monarchs, Spanish monarchs did have independent revenues, especially silver fleets from Latin America, and they did borrow in advance of those revenues, but they did so from Genoese bankers.

Although the Spanish king, Phillip II, repeatedly defaulted on such Crown loans, he eventually negotiated terms with the bankers. In the long term, the bankers did well because they could determine if Phillip's default was from a legitimate inability to pay, i.e., a silver fleet failed, versus a strategic unwillingness to pay (Drelichman and Voth 2014). Like farm syndicates, the Italian bankers could credibly threaten to disrupt future borrowing if the monarch was able but unwilling to repay. Spanish kings were, however, careful not to violate commitments sanctioned by the Cortes.

Republics

Centuries before the monarchs of England, France, and Spain learned to arrange multiyear borrowing, the city-republics of Italy unlocked the supply of long-term capital. Like for the monarchies, the critical source of domestic credit was the merchant class, because they had money and wanted to diversify their investment portfolio. Unlike the monarchies, merchants held substantial political power in cities like Florence, Genoa, and Venice, so political will was expected to remain focused on repayment (Stasavage 2015). Initially, this was done through substantial elite participation in loans, and "Citizens thus became true lender-taxpayers, rather than lender-investors" (Pezzolo 2007: 4).

For example, in 1262, the oligarchy running Venice compelled creditors to restructure and consolidate short-term debts into perpetual annuities backed by dedicated revenues managed by the city (Fratianni and Spinelli 2006). When subsequent wars created more fiscal need, additional compulsory loans were created. While such loans sound predatory, the elites of Venice used this instrument because it required their political consent, it limited the use of direct taxation that would fall on them, it avoided the domination of lending by a faction, and it created a useful demarcation of citizenship (Pezzolo 2007). As a result, most people of at least modest wealth held perpetual bonds in the form of a ledger entry at the state loan office. A local resale market developed. "Government credits could be transferred by sale, by testamentary bequest, or by contract of gift or dowry; they could be lent, pledged as surety for nearly any kind of transaction, including bank loans, or used as money in payment of obligations" (Mueller 1997: 458). This only made the credits more valuable, and that helped Venice borrow even more. By the mid-1400s, however, new compulsory loans began to exceed the elite's ability to supply. As a result, people had to sell old credits at diminished prices to pay for new credits and political support for compulsory loans dissolved. In 1528, Venice switched to voluntary loans.

The Republic of Genoa did not have as stable an oligarchy as did Venice. Noble factions quarreled with each other and with an ascendant merchant class, so maintaining the political will to repay creditors was less assured. To acquire long-term credibility for its compulsory annuities, Genoa conceded control of revenue collection to tax farms, whereas Venice maintained direct administration of its taxes. Also, Genoa hit the limits of its long-term borrowing a century before Venice did. In 1407 Genoa initiated a voluntary restructuring that swapped old annuities for

shares in a new holding company called the *Casa di San Giorgio* (Bank of St. George) by which the state got a reduction in interest rates. Creditors got a company that increased their bargaining power by consolidating negotiations, tax collection, and debt management into one entity. This super tax farm syndicate "reduced creditors' fears of government defaulting on its obligations, and ultimately lowered the cost of debt to Genoa" (Fratianni 2006: 487). It also deepened the resale market for sovereign debt by repackaging it as a homogeneous share of the company. In modern parlance, San Giorgio securitized Genoa's sovereign debt. The result was that capital investors needed to know little because the leadership of the company were political elites who collected information and negotiated terms on behalf of a board and the rest of the shareholders. Leadership even assessed when circumstances legitimately prevented repayment, in which case debt could be restructured or even forgiven (Fratianni 2006: 496). In total, individual investors in San Giorgio owned a liquid security about which they needed little information, a remarkable innovation for the 1400s. This security encouraged institutions like the church and charities to become substantial investors and resulted in Genoa, on a per capita basis, borrowing substantially more than any other Italian state (i.e., three times Venice) from 1500 to the French Revolution (Chilosi 2014).

The first republic to cover a territory encompassing competing cities was the Dutch Republic. During the early years of the revolt from Spain (1568–1648), individual Dutch cities continued the tradition of borrowing on annuities. As early as the thirteenth century, French cities such as Amiens, Rheims, Tournai, and Troyes had gained commune status from the French kings. Each used that corporate status to borrow in its town name. The status was credible because French lawyers agreed that the King had the authority to grant communal status, at the same time the King's authority to impose modifications to local arrangements was limited by tradition (Tracy 2003: 19). In exchange for communal status, the cities supplied military funding to the king. The commune-annuity arrangement spread into the cities of the Habsburg Low Countries, and later the northern cities issued annuities to fund their rebellion. The arrangement did not work in southern France because either the French king's "rights as suzerain were not so clearly defined" (Tracy 2003: 19) or the city was under English rule and suffered from political divisions.

During the revolt, the Dutch cities had to determine how to coordinate their fiscal efforts. In 1574 the cities decided to standardize taxation authority at the level of the province – the Dutch Republic being comprised of seven provinces (Fritschy 2003). In this way tax obligations were made uniform for those in cities and the countryside. The cities, however, did not centralize the collection and processing of taxes. Instead, they continued using local tax receivers to borrow advances on the local obligations to provincial taxes. The same receivers then later repaid those advances with locally collected funds. The cities controlled the provincial government; keeping taxes local assured urban citizens that rates were necessary, that funds were spent wisely, and that advances would be repaid (Gelderblom and Jonker 2011). The local sequestration of Holland's fiscal process had the intended effect of standardizing local rates while keeping local control of the flow of funds. The national government had only a minor fiscal role.

An unintended effect was to give localities freedom regarding how to supply long-term sovereign debt. While the province might favor annuities repayable at the government's discretion, locals preferred bills due within a year, partly because the transfer of annuities was cumbersome and paid a tax/duty while bill's matured quickly without a tax (Gelderblom and Jonker 2011). As a result, Holland's long-term capital debt became mostly a continuous rolling over of short-term bills.

Massive use of short-term debt had two related consequences. Borrowing rates were lower than annuities, but lenders might withdraw their support and not roll over their loans. This happened in 1672 when France coordinated a three-sided invasion of the Netherlands. Holland had to suspend conversions and impose forced loans (Gelderblom and Jonker 2011). The conclusion of that war in 1678 allowed the situation to restabilize. Holland successfully created safe sovereign debt that was functionally equivalent to a time deposit with the local tax receiver. Like a bank deposit, investors were largely ignorant of fiscal operations. "The Estates published no financial information whatsoever; only the receivers knew how much money was to be raised or still needed to be subscribed, so they could play investors against each other" (Gelderblom and Jonker 2011: 24). Like a bank, the "no-questions-asked" status of Holland's bills could suddenly collapse into a run if investors decided that their ignorance made them vulnerable.

As ever in the early modern era, the reliability of Holland's system depended on politics, and politics depended on war. The wars of 1687–1697 and 1701–1713 placed massive fiscal demands on Holland, and under this pressure the province turned on its bill creditors. In 1687 Holland began to tax interest payments on bills (Gelderblom and Jonker 2011). In effect, the tax used the amount owed on bills as a proxy for a person's wealth. At the same time, the province borrowed new funds that were not subject to the tax while limiting access to investors who held the old bills and paid the tax. All this was possible because local tax receivers controlled who had access to new debt issues. Receivers threatened to cut off access to any government debt to locals who cashed out old bills to avoid the tax. Receivers encouraged rollovers by controlling access to the new, more lucrative investment. Receivers could manipulate local investors precisely because the province did not have a centralized process.

Joint-Stock Sovereign Debt

While the Dutch manipulated illiquid forms of sovereign debt, their ally Britain experimented with melding the liquidity of trading companies with sovereign annuities. The proof of concept was the public offering for the Bank of England. With the defeat of allied forces by the French in 1690, the navy needed funds to rebuild, but trust in the government was such that borrowing at any reasonable interest rate was not possible. The limited liability, joint-stock company structure was used as the vehicle by which funds would be channeled to the government. The Bank of England, chartered in 1694, raised £1.2 million in only 12 days in a world in which the annual wage for a laborer was roughly £20. In simple

purchasing power, this is the equivalent of £160.1 million pounds in 2017 (Measuring Worth). Not only did the IPO demonstrate the wealth available in London but also the willingness of individuals to buy into the stock market (Carlos and Neal 2006; Murphy 2009).

Although called a bank, the initial Bank of England was, in effect, a pool of sovereign debt funded by securities: shares and bonds (Kleer 2017). The Bank of England only slowly learned to issue circulating banknotes. Additionally, the early Bank of England did not manage tax collections nor represent all sovereign creditors, so it had much less threat potential than Genoa's *Casa di San Giorgio*. However, the government did get to place long-term debt at a low rate. Investors got liquidity, and the collective political threat of upset shareholders should the state default (Murphy 2013). The Bank soon discovered the power of supplying currency and a trade-off evolved (Broz and Grossman 2004). In 1697, the Bank of England agreed to swap new shares in the bank for short-term sovereign debt held by savers that was in arrears. The debt was simultaneously converted into annuities. In return, the Bank of England was granted a monopoly over corporate banking in England. Five times over the next century the Bank of England absorbed new annuities in exchange for being allowed to keep its monopoly, and the amount of annuities tied to the Bank of England's charter grew tenfold from its origin. The Bank of England, however, never held more than a quarter of the national debt. But the model was one that was replicable. In exchange for corporate privileges, the New East India Company lent £2 million to the government in 1698, and the United East India company lent £1.2 million more in 1708 (Scott 1910). This is in contrast to the VOC, whose capital stock was fixed in perpetuity.

The path to the great financial bubbles of 1720 begins with the use of the corporate form to convert existing debt into shares (Neal 1991). In 1711 the Tory government chartered the South Sea Company (SSC) to rival the Whig's Bank of England. While not a bank, and hardly even a trading enterprise, based on the experiences of the Bank of England and the East India companies, the South Sea Company was able to undertake a very successful swap of £9.1 million worth of sovereign debt for new shares in the company because people wanted liquidity.[3] As a result, the SSC suddenly held more government annuities than the Bank of England and the East India Company combined. The Treasury, which was responsible for selling annuities to the public, decided to use these companies to restructure other annuities held by the public. Benefits accrued to all parties: the government paid the companies a lower rate of interest, companies received a stable flow of interest payments, and the shareholder now held a highly liquid asset.

[3]£9.1 million in 1711 is £1,219 million in 2018 in terms of purchasing power parity or $1,1618,868,570 at current exchange rates. Measuring Worth "How Much is That" accessed June 15, 2018.

The British experience showed the power of aligning long-term sovereign debt with corporations in exchange for various trade and banking privileges. The Scottish promoter John Law took the lessons to Paris and convinced the Regent to follow the example and then go one better. From 1717 to 1720 in France, Law monopolized currency (Bank Royale), the collection of taxes (Comptroller General), and foreign trade (Mississippi Company). He used the combination of powers to encourage people to swap sovereign debt for Mississippi Company shares.

This expanded concept jumped back to London where the treasury had the South Sea Company and the Bank of England bid for the right to absorb the half of the national debt still held directly by public investors through a giant debt-for-equity swap. The South Sea Company won the contract and was indebted to the government for over £7 million pounds (Scott 1910). South Sea Company share price rose dramatically because the company supplied subscription credit but also because investors expected that the successful company would dominate public finance like Law's system in Paris. Most of the available annuities were converted into shares as the South Sea Bubble inflated over the summer of 1720.

John Law's scheme in France began to implode in May of 1720. Law was printing so much currency to maintain a Mississippi Company share price floor that he suspended convertibility of currency into coins. The suspension, however, was politically unacceptable, with the result that Law had to revert to a policy that was unsustainable. In July his bank failed and the stock's value collapsed.

Meanwhile, in London, credit creation could not keep pace with subscription margin calls on earlier rounds of credit, and South Sea Company share prices began to fall in late August. But by then, the conversions were largely complete, so the Treasury did not suffer. However, many investors who had swapped government debt for equity were unable to repay credits. Personal failures increased for those who had purchased at the height of the market, leading to calls for government to take responsibility for these losses. Much has been made of these losses but not everyone lost; those who sold at the height of the market or gained over the Bubble period tended to be hidden (Carlos and Neal 2006). In 1723 the political calls for action resulted in a compromise in which half of the holdings of South Sea Company stock were split into two – one half to continue to be held as equity and the other half to convert to annuities to be housed in the Bank of England. Despite the Bubble, the capital market in London continued to grow through subsequent decades of the nineteenth century. Both debt and equity became a highly liquid and safe form of sovereign debt. In the decades to follow, the Treasury took the final step of routinely and directly issuing annuities that traded like stock. The annuities were recorded and transferred at the Bank of England using the same types of ledgers as equity shares. Indeed, people called these perpetual annuities stocks. This market provided the British government with a mechanism to meet the financial exigencies of war. But more importantly, the market was and continued to be available and accessible to even the smallest of investors. It was this accessibility that provided the liquidity not only for smoothing expenditure needs but also for companies making large capital investments in canals, railroads, and many other infrastructure projects.

Conclusion

Capital markets involved a great deal of experimentation, often in desperate times. By 1800, however, a replicable system of limited liability, long-term securities had been worked out. Investors needed neither to participate in management nor to closely monitor the firms. A great deal of useful information flowed through the prices of stocks and bonds on the secondary market. Indeed, one of the milestones in cliometric analysis was when Larry Neal (1991) showed that capital markets were highly correlated in the eighteenth century. A substantial literature has followed that uses statistical analysis to highlight the sophistication of information processing in early modern capital markets: some recent examples being Chilosi (2014), Petram (2014), Koudijs (2015), and Wandschneider (2015). In the modern era, the corporate securities technology developed by Dutch and English trading companies would jump sectors again and play an essential role in funding the large, industrial giants of railroads, steel, automobiles, etc. And the sovereign version of the corporate bond would go on to become the safe asset that anchors the financial system of the industrialized world.

A capital structure using corporations and securities, however, was limited in the early modern era and remains a questionable fit for smaller firms. For example, Hoffman et al. (2002) showed how Parisian notaries intermediated capital lending without secondary markets or even price competition. Again, a substantial literature has followed showing how capital could be created through personal connections and rolling over short-term loans. For example, see Temin and Voth (2013), Hoffman et al. (2015), and Gelderblom and Jonker (2016). While securities-based finance has become a massive industry, it is important to remember that a good deal of capital creation still occurs through more private networks.

References

Akerlof G (1970) The market for 'lemons': quality uncertainty and the market mechanism. Q J Econ 84(3):488–500

Broz JL, Grossman RS (2004) Paying for privilege: the political economy of Bank of England charters, 1694–1884. Explor Econ Hist 41(1):48–72

Carlos AM, Neal L (2006) The micro-foundations of the early capital market: Bank of England shareholders during and after the south sea bubble, 1720–1725. Econ Hist Rev 59(3):498–538

Carlos AM, Neal L (2011) Amsterdam and London as financial centers in the eighteenth century. Financ Hist Rev 18:21–47

Carlos AM, Nicholas S (1988) Giants of an earlier capitalism: the early chartered companies as an analogue of the modern multinational. Bus Hist Rev 26(3):398–419

Carlos AM, Nicholas S (1990) Agency problems in early chartered companies: the case of the Hudson's Bay Company. J Econ Hist L(4):853–875

Chilosi D (2014) Risky institutions: political regimes and the cost of public borrowing in early modern Italy. J Econ Hist 74(3):887–915

Dari-Mattiacci G, Gelderblom O, Jonker J, Perotti EC (2017) The emergence of the corporate form. J Law Econ Organ 33(2):193–236

Dincecco M (2015) The rise of effective states in Europe. J Econ Hist 75(3):901–918

Drelichman M, Voth H-J (2014) Lending to the borrower from hell. Princeton University Press, Princeton

Fratianni M (2006) Government debt, reputation and creditors' protections: the tale of San Giorgio. Rev Finance 10:485–504
Fratianni M, Spinelli F (2006) Italian city-states and financial evolution. Financ Hist Rev 10:257–278
Fritschy W (2003) A 'financial revolution' reconsidered: public finance in Holland during the Dutch Revolt, 1568–1648. Econ Hist Rev LVI(1):57–89
Gelderblom O, Jonker J (2011) Public finance and economic growth: the case of Holland in the seventeenth century. J Econ Hist 71(1):1–31
Gelderblom O, Jonker J (2016) Direct finance in the Dutch Golden Age. Econ Hist Rev 69(4):1178–1198
Goetzman WN (2016) Money changes everything: how finance made civilization possible. Princeton University Press, Princeton
Gorton G (2017) The history and economics of safe assets. Annu Rev Econ 9:547–586
Greif A (1993) Contract enforceability and economic institutions in early trade: the Magribi traders' coalition. Am Econ Rev 83(3):525–548
Greif A (2006) Institutions and the path to the modern economy: lessons from medieval trade. Cambridge University Press, Cambridge
Guinnane T, Harris R, Lamoreaux NR, Rosenthal J-L (2007) Putting the corporation in its place. Enterp Soc 8(3):687–729
Harris R (2009) The institutional dynamics of early modern Eurasian trade: the *commenda* and the corporation. J Econ Behav Organ 71(3):606–622
Hoffman P (2015a) Why did Europe conquer the world? Princeton University Press, Princeton
Hoffman P (2015b) What do states do? Politics and economic history. J Econ Hist 75 (2):303–332
Hoffman P, Postel-Vinay G, Rosenthal J-L (2000) Priceless markets: the political economy of credit in Paris, 1660–1870. University of Chicago Press, Chicago
Hoffman P, Postel-Viney G, Rosenthal J-L (2002) Priceless markets: the political economy of credit in Paris, 1660–1870. University of Chicago Press, Chicago
Hoffman P, Postel-Viney G, Rosenthal J-L (2015) Entry, information, and financial development: a century of competition between French banks and notaries. Explor Econ Hist 55:39–57
Holmström B (2015) Understanding the role of debt in the financial system. Bank for International Settlements, working paper no 479
Jha S (2015) Financial asset holdings and political attitudes: evidence from revolutionary England. Q J Econ 130:1485–1545
Johnson N (2006) Banking on the king: the evolution of the royal revenue farms in old regime France. J Econ Hist 66(4):963–991
Johnson N, Koyama M (2014) Tax farming and the origins of state capacity in England and France. Explor Econ Hist 51:1–20
Kleer RC (2017) Money, politics and power: banking and public finance in wartime England, 1694–96. Routledge, New York
Koudijs P (2015) Those who know the most: insider trading in eighteenth-century Amsterdam. J Polit Econ 123(6):1356–1409
Koudijs P (2016) The boats that did not sail: asset price volatility in a natural experiment. J Finance 71(3):1185–1226
Laurence A (2008) The emergence of a private clientele for banks in the early eighteenth century: Hoare's Bank and some women customers. Econ Hist Rev 61(3):565–586
Levine R, Zervos S (1996) Stock market development and long-run growth. Work Bank Econ Rev 10(2):323–339
Measuring Worth https://eh.net/howmuchisthat/
Mueller RC (1997) The Venetian money market. Johns Hopkins University Press, Baltimore
Murphy A (2009) The origins of the English financial markets. Cambridge University Press, Cambridge
Murphy A (2013) Demanding 'credible commitment': public reactions to the failures of the early financial revolution. Econ Hist Rev 66(1):178–197
Neal L (1991) The rise of financial capitalism. Cambridge University Press, Cambridge

North DC, Weingast B (1989) Constitutions and commitment: the evolution of institutions governing public choice in seventeenth century England. J Econ Hist 49:803–832
Ogilvie S (2014) The economics of guilds. J Econ Perspect 28(4):169–192
Petram L (2014) The world's first stock exchange. Columbia Business School, New York
Pezzolo L (2007) Government debts and credit markets in Renaissance Italy. Working paper, Department of Economics, University of Venice Ca' Foscari
Puttevils J (2015) Tweaking financial instruments: bills obligatory in sixteenth-century Antwerp. Financ Hist Rev 22(3):1–25
Quinn S (2001) The Glorious Revolution's effect on English private finance: a microhistory, 1680–1705. J Econ Hist 61(3):593–615
Quinn S, Roberds W (2014) How Amsterdam got fiat money. J Monet Econ 66(1):1–12
Rei C (2011) The organization of Eastern Merchant Empires. Explor Econ Hist 48(1):116–135
Ross SA (1977) The determination of financial structure: the incentive signaling approach. Bell J Econ 8:23–40
Santarosa VA (2015) Financing long-distance trade: the joint liability rule and bills of exchange in eighteenth-century France. J Econ Hist 75(3):690–719
Scott WR (1910) The constitution and finance of English, Scottish and Irish joint-stock companies to 1720, 3 vols. Thoemmes Press, Bristol, England
Stasavage D (2015) What we can learn from the early history of sovereign debt. Explor Econ Hist 59:1–16
Temin P, Voth H-J (2013) Prometheus shackled: Goldsmith Banks and England's financial revolution after 1700. Oxford University Press, Oxford
Tracy JD (2003) On the origins of long-term urban debt in medieval Europe. In: Boone M, Davids K, Janssens P (eds) Urban public debts. Brepols Publishing, Turnhout, pp 13–24
Wandschneider K (2015) Landschaften as credit purveyors – the example of East Prussia. J Econ Hist 75(3):791–818
White EN (1995) The French Revolution and the politics of government finance, 1770–1815. J Econ Hist 55(2):227–255

Origins of the U.S. Financial System

Richard Sylla and Robert E. Wright

Contents

Abstract

Before the Revolution, British North American colonists generally circulated foreign specie (full-bodied silver and gold coins) rated in local monies of account and paper bills of credit to settle open account balances, but during monetary stringencies they resorted to country pay and other exigencies. Formal financial intermediaries were few, and for the most part urban and nonprofit.

During the Revolution, governments financed rebellion by issuing dollar-denominated Continentals and other fiat paper monies. Much more currency was emitted than demanded at the prewar price level, so the paper money depreciated until it became virtually worthless. Americans then reverted to specie, and the rebel governments turned to foreign loans and in 1782 to the nation's first joint-stock commercial bank, the Bank of North America.

R. Sylla (✉)
Stern School of Business, New York University, New York, NY, USA
e-mail: rsylla@stern.nyu.edu

R. E. Wright
Social Science, Augustana University, Sioux Falls, SD, USA
e-mail: robert.wright@augie.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_72

Shortly after the Revolution, two more commercial banks appeared in seaport cities, and a few other business corporations formed, but generally entrepreneurs held back, fearful of the continuation of recession and the return of civil war. Only with the passage of the new Constitution and the installment of a new federal government, led by the venerable George Washington and financial reformer Alexander Hamilton, did the investment floodgates open. By 1795, the new nation had a fully formed modern financial system – including strong public finances and debt management, a uniform unit of account, a central bank, a commercial banking system, efficient securities markets, a network of insurance companies, and economic incentives and legal rights to innovate – that drove America's nineteenth century economic growth and development.

Keywords
Alexander Hamilton · Banks and banking · Financial institutions and markets · Insurers and insurance · Lender of last resort · Money and monetary systems · Securities markets

Introduction

Early U.S. economic history furnishes what is perhaps the best example of financial modernization leading to modern economic growth. When the members of the new federal government under the Constitution took office in 1789, the country had virtually none of the key components of modern financial systems. Half a decade later it had all of them. The term "financial revolution" is apt.

At roughly the same time, economic growth also accelerated to its modern rate, i.e., sustained increases in Gross Domestic Product (GDP) per person of 1% or more per year. The United States – not, as long assumed, Great Britain with its "first" industrial revolution – may have been the first country to benefit from modern growth rates (Lindert and Williamson 2016). After its financial revolution of the 1790s, it took the United States less than a century to succeed Great Britain as the richest and foremost industrial economy.

The United States was not the first or the last country to experience finance-led growth. The Dutch Republic at the turn of the seventeenth century experienced financial modernization and went on in that century to become the leading world economy. A century later, after its Glorious Revolution of 1688, Great Britain also had a financial revolution; it succeeded the Dutch Republic as the leading economy of the eighteenth century and the first part of the nineteenth. Perhaps the most interesting case, and a later one, is that of Japan, which after centuries of isolation, opened up to the world starting in the 1850s. During the 1870s and 1880s, the Japanese modernized their financial system as the Dutch, the British, and the Americans had done earlier. The growth of the Japanese economy then accelerated to modern levels and by the early twentieth century Japan far surpassed its Asian neighbors in economic development (Rousseau and Sylla 2003).

During their long colonial era from the early seventeenth century to the late eighteenth, Britain's North American colonies that became the United States experienced economic growth, but only at pre-modern rates. Living standards were high compared to the rest of the world because land and other natural resources abounded compared with the relatively small, though rapidly growing, population. But over time, living standards increased little if at all. Expansion resulted from a rapidly increasing population – the increase was about 3% per year – that maintained, not increased, its relatively high level of income. In other words, both the population and economic output expanded at roughly 3% per year, an exceptional rate in world history up to that time, but not growth in per capita (person) terms.

Pre-modern financial arrangements constrained colonial growth. Much of the specie that colonists earned from their exports of commodities and shipping services was expended on needed imports. The dearth of specie led to a search for innovative substitutes, but both economic realities and imperial regulations often frustrated their usefulness. British interference with colonial financial and monetary experiments and the resulting constraints on colonial economic growth, along with Britain's mercantilist trade policies and taxation without representation, became prime causes of the American revolt, which began in the wake of the long French and Indian War (1754–1763) and reached a world-altering climax in the 1770s.

Money and Banking in Colonial America

For legal and practical reasons, British North American colonists minted few coins themselves. Full-bodied foreign gold and silver coins (specie) generally did not abound in domestic circulation because colonists preferred consuming imported goods over maintaining copious cash balances (Michener 2003). A colonist about to part with a Portuguese gold coin (a Johannes, "Joe," or "Jo") penned a poem about the peripatetic nature of specie:

> We are come to the unhappy parting hour. Lately I received you into my house as a Traveller, and almost a Stranger; you was welcome. . . . I must tell you, I am now obliged to sell you to a Merchant, don't think I do it of choice. . . . I hoped you might have remained an inhabitant of the country, that I might have receiv'd some visits from you, but now I expect you will have a quick dispatch to Boston, or New York; immediately take ship and I shall see you no more. . . . Think not hard of me for putting you under this sentence of Banishment, necessity knows no Law. Farewell, my friend Jo ("The Following is a Speech, Made at the Delivery of a Gold Piece, Call'd a Johannes, to a Merchant for Debt," *Connecticut Gazette,* 23 January 1768).

Colonists chose to remit full-bodied coins (i.e., with precious metal content equal to the face value of the coin) to pay for imported goods because they possessed cash substitutes that were generally adequate to their needs. Contrary to myth, colonists rarely engaged in barter – the direct exchange of one nonmoney good for another, though final payments often consisted of goods or services. Under the "country pay" system, for example, colonial legislatures made baskets of local agricultural products

legal tender, sometimes only for public debts but sometimes for private debts as well, at specific specie prices. Single commodities, including tobacco, beaver skins, and wampum also served as money at various times and places, sometimes by fiat and sometimes by common consent. Tobacco also spurred the creation of a representative paper money known as tobacco notes, which provided holders with legal rights to specific amounts of inspected and warehoused tobacco (Michener 2003).

The open account system was much more flexible and honest than country pay, which could be used to trick unsuspecting or ignorant creditors. Sometimes called bookkeeping barter, the open account system, which lasted into the twentieth century in rural areas, entailed the exchange of goods, services, and money between economic entities over long periods, from months to years to decades (Flesher 1979). In essence, every household and business firm performed banking functions by keeping track of the market value in specie of purchases, sales, loans, and repayments over time, with periodic settlement with counterparties made in cash (coins, bills of credit, or tobacco notes) when a balance grew too large.

Colonists typically recorded economic values in terms of local pounds (£), shilling (s), and pence (d) with 12d to the s and 20s to the £. By rating foreign coins, especially the Spanish milled dollar, each colonial government or merchant association set the value of its local pound at some fraction of the Mother Country's pound sterling. New York eventually rated the dollar at 8 shillings New York money, so the New York pound was worth 2 and one half dollars (20s/8s = $2.50). Pennsylvania rated the dollar at 7 shillings and 6 pence Pennsylvania money, so the Pennsylvania pound was worth 2 and two-thirds dollars (20s/7.5s = $2.67). In late colonial New England, the dollar was rated at 6s (20s/6s = $3.33). In Britain, the dollar was worth about 4s 6d sterling, or $4.44 per pound sterling (20s/4.5s = $4.44) (Michener 2003).

The differences in local coin ratings were reflected in exchange rates, i.e., the prices of sterling-denominated bills of exchange. To purchase £100 sterling, New Yorkers in peacetime typically paid £174 to £182 New York currency, while Pennsylvanians typically paid £163 to £171 Pennsylvania currency, and late colonial New Englanders paid £127 to £133 Massachusetts currency. The intercolonial difference was due to the differential rating of the silver dollar in each colony compared to the sterling rating of 4 s 6d per dollar. The intracolonial range reflected the transaction costs inherent in shipping and insuring specie. When bills of exchange became too dear, colonists who owed debts in Britain shipped specie instead; when bills of exchange became too cheap, which occasionally happened, colonists with credit in Britain imported specie from abroad instead (Michener 2003).

Bills of exchange were short-term credit devices with a foreign exchange component, a sort of international paper check denominated in another currency, typically pounds sterling (but sometimes Dutch guilders or German marks). Typically, a drawer (or maker, an individual with funds or credit in England or elsewhere overseas) would sell the bill, which could be completely epistolary or a printed form with key blanks filled up by hand, to a purchaser, who remitted the bill to the payer (the banker/merchant) in London, Liverpool, Amsterdam, Hamburg, etc. in triplicate (or quadruplicate) via different ships. Only the first bill presented to the

payer was valid. Bills of exchange could be sold or otherwise transferred to another party only by endorsement (signing over ownership on the back) and the payer did not always honor (pay) them.

Bills of exchange provided a substitute to coin in international trade but not in domestic circulation. There, a very different paper instrument with a similar sounding name, bills of credit, substituted for cash money. Issued by colonial governments via mortgage loans or directly for goods and services, bills of credit were printed bearer instruments typically denominated in local pounds, shilling, and pence. Often, but not always, a legal tender for all debts public and private, bills of credit were typically backed by loose promises to redeem them for taxes in the future, not by specific mortgages, caches of specie, or other assets. Only Maryland backed an emission of bills of credit with tangible assets and a credible commitment to redeem them at a specific time. Everywhere else, bills were essentially unbacked, so their value in the marketplace was determined first by the local coin rating and second, if bills displaced all specie from domestic circulation, ostensibly by the quantity of bills circulating. Attempts to cliometrically test the quantity theory of money in the colonies, however, have been flummoxed by the cross-border circulation of bills, incorrect measurement of the total value of bills outstanding, and the difficulty of measuring specie flows (Michener 2003, 2015).

Writing after the Revolution about bills of credit in the Middle Colonies (New York, New Jersey, and Pennsylvania), "Eugenio" explained that they maintained their value because

> the people voluntarily and without the least compulsion threw all their gold and silver, not locking up a shilling, into circulation concurrently with the bills; . . . If any one doubted the validity or price of his bill, his neighbor immediately removed his doubts by exchanging it without loss into gold or silver. If any one for a particular purpose needed the precious metals, his bill procured them at the next door, without a moment's delay or a penny's diminution (*New Jersey Gazette,* 30 January 1786).

In other words, the inhabitants of the Middle Colonies, not formal financial institutions, maintained the convertibility of local bills of credit into specie and vice versa as part of their open account or bookkeeping barter system (Michener 2003). Such formal financial institutions that did exist in the Middle Colonies were charities or mutual insurers that made long-term investments in mortgages and ground rents (perpetual interest-only mortgages) (Roney 2014).

The monetary situation in early colonial South Carolina and mid-century New England was very different. In those places, military exigencies induced the colonial governments to issue far more bills than were needed for domestic circulation at prevailing prices, freeing all the coins in domestic circulation to be sent abroad in exchange for manufactured goods. While exchanging small pieces of metal for consumption goods and productivity-enhancing tools provided a temporary boon, with nothing to anchor the value of bills of credit they began to depreciate. Although it is tempting to call such episodes inflationary, that would be anachronistic and misleading because prices denominated in specie did not change appreciably.

Instead, bills of credit lost value vis-à-vis specie and goods until reforms were initiated. In South Carolina, that meant an exchange rate of about £700 South Carolina currency for £100 sterling. Nominal exchange rates in New England were also quite high until the region, led by Massachusetts, underwent a series of currency reforms in the early 1750s that essentially put the region back on a specie standard (Brock 1975).

The first attempts to create private forms of money took place in South Carolina and New England as entrepreneurs sought to create specie substitutes that held their value better than fiat bills of credit did. In the former, Sir Alexander Cumings ran a private bank that issued paper notes redeemable, at least at first, upon demand in specie. During a diplomatic mission among the Southeast's American Indians in 1729, Cumings, who had seen John Law's system in operation in France in 1719, issued his promissory notes via three different loan offices. After establishing his creditworthiness, he began to sell bills of exchange as well and when he left the colony, quite publicly, all seemed well. Three weeks later, however, the bills of exchange he had drawn came back protested (unpaid) due to the "Character of the Drawer." Noteholders busted into his "treasury" building only to find in it "some empty Boxes, old Iron, and other Rubbish." Contemporaries estimated Cumings' haul at £15,000 sterling (*Boston Weekly News Letter*, 30 July 1730; [London] *Echo*, 16 September 1730).

Ironically, Cumings' initial success spurred the creation of a competitor, a bank with 25 principals who each chipped in £2000 South Carolina currency (approximately £286 sterling). After a rocky start, the merchants lent £40,000 to £50,000 worth of their notes to the public and redeemed them on demand in South Carolina bills of credit. A third bank and several individual merchants also apparently were able to circulate private notes in 1730s South Carolina, but documentation remains thin (Michener 2017).

Also in response to the depreciation of fiat bills of credit, a group of Boston merchants in 1733 printed and lent at 6% interest £110,000 in private paper notes, redeemable in specie at 19 s an ounce of silver in three installments over the course of a decade. The action caused a stir, but the attorney general of Great Britain eventually ruled the notes were legal because the Bubble Act and other legal barriers did not (yet) apply to the colonies. Because Rhode Island successfully continued to pump new bills of credit into circulation throughout New England, the value of fiat money paper money dipped enough to induce people to hoard the merchants' notes, which soon disappeared from circulation (Michener 2017).

In Connecticut in 1732–1733, another group, styled the New London Society United for Trade and Commerce, tried and failed to get their paper notes to pass current, apparently because the projectors were not considered creditworthy. The legislature revoked its charter within a year. In New Hampshire in 1735, a group of prominent merchants copied the business plan that Boston merchants had used 2 years earlier to issue notes successfully. Merchants and lawmakers quickly agreed to make the New Hampshire notes illegal in Massachusetts, which soon killed the scheme, leaving the noteholders to bear the full loss (Michener 2017).

In 1740, two more private banking schemes were floated in Boston, one that would have issued notes backed by mortgages on real estate and another backed by deposits of silver. Political rivalries prevented either scheme from moving forward with much speed and the extension of the Bubble Act to the colonies in 1741, which made explicit reference to both the land and silver banks, effectively squelched both enterprises, along with two additional schemes, one established in April 1741 in Essex County and another in Middlesex County. Extension of the Bubble Act to the colonies also forced the South Carolina banks to wrap up their affairs, which by then apparently were negligible anyway (Michener 2017).

In the aftermath of the French and Indian War, British authorities outlawed the issuance of new, legal tender bills of credit, mandated high taxes that rapidly called in the bills of credit issued during the war, and placed colonial trade under very strict supervision. As a result, the colonial money supply (coins and bills of credit) dropped precipitously, causing a major macroeconomic disturbance tied directly to the agitation over the Stamp Act. Money became so scarce that interest rates soared, property values plummeted, and default rates on loans soared.

Desperate for some means of liquidating debts, the residents of Bucks County, Philadelphia, resorted to circulating squirrel scalp bounties, epistolary IOUs worth 1d each made out by justices of the peace to pay the bounty on squirrels, which colonists considered agricultural pests of the first order. Merchants in Philadelphia, where debts were larger and squirrels less numerous, opted instead to form a bank that would issue private, asset-backed notes. Jealousies at home and legal pressures from London, however, quickly killed the enterprise, along with a similar scheme floated in New York (Michener 2017).

Revolution

The postwar monetary stringency helped lead to the Revolution by inducing colonists to resist the Stamp Act. As an anonymous colonist explained to a correspondent in 1768:

> It is not the Stamp Act or New Duty Act alone that had put the Colonies so much out of humour tho the principal Clamour has been on that Head but their distressed Situation had prepared them so generally to lay hold of these Occasions. . . . money being plenty, both in Specie and Bills of Credit [during the war] People were not afraid of entering into deep Engagements equivalent to our Circulation. . . . [In addition to the use of warships to] cramp trade by stoping [sic] and detaining Merchant ships and pressing their Men . . . Another capital Greivance [sic] and Inconvenience to these middle Colonies is being restrained from issuing Notes or Bills of Credit, which is no other than a Bank held by the public who are engaged for the Credit of it and reap the Emoluments of it; to recount our Distress for want of this Medium were Endless or by what unaccountable Policy G. Britain acts in the Restriction ("New Jersey Currency Question," Record Group 23, Box 4, New Jersey Historical Society, Newark, N.J.).

Resistance to the Stamp Act set off a long chain of events that ended with the Declaration of Independence. At first, the rebel governments financed the

Revolution the same way that most had financed wars throughout the eighteenth century, by issuing fiat paper money. During the Revolution, that meant state-issued bills of credit and also dollar-denominated Continentals issued by Congress, the nascent national government. The paper money issues at first refreshed economies long desiccated of cash, but as the war dragged on and the unbacked money in circulation swelled in nominal value, depreciation soon set in. Inflation, i.e., a rise in the specie price of goods, also occurred but paled in comparison to the depreciation of paper monies, which is shown in Table 1, a "scale of depreciation":

While the war had the salubrious effect of reducing reliance on bookkeeping barter and inducing more Americans to reckon economic value in terms of dollars, the massive depreciation and repudiation of federal and state-issued bills of credit, the third instance of bills of credit losing almost all value in a generation (New England's currency reforms of 1750 and colonial bills were the first two), soured Americans on fiat paper money for the better part of a century. They could have reverted to the use of specie and open accounts but the government's need for revenue suggested the more fiscally-savvy approach of establishing a commercial bank that would issue bank notes and deposits convertible into specie and would make short-term loans to governments and businesses.

Credit for the nation's first commercial bank, the Bank of North America (BNA), must be shared between Alexander Hamilton, a financially precocious young officer on General Washington's staff, and Philadelphia merchants Robert Morris and Thomas Willing, two of the merchants behind the abortive 1760s bank scheme. Hamilton laid out the policy grounds for the bank in two letters in 1780 and 1781. In the first letter, written most likely in early 1780 to an unknown addressee who may have been a member of Congress, Hamilton developed a plan to restore confidence in American currency, which had been destroyed by excessive issues of fiat currency and the near hyperinflation that followed. His plan called for Congress to obtain a

Table 1 Depreciation of Continental currency, 1777–1781

Month	1777 Continentals per one specie dollar ($C/S)	1778 $C/S	1779 $C/S	1780 $C/S	1781 $C/S
Jan.	1.5	4	8	42	75
Feb.	1.5	5	10	45	80
March	2	5	10	50	90
April	2.5	5	16	60	100
May	2.5	5	20	60	150
June	2.5	5	20	65	250
July	3	5	21	65	400
Aug.	3	5	22	70	500
Sept.	3	5	24	72	600
Oct.	3	5	28	73	700
Nov.	3	6	36	74	800
Dec.	4	6	40	75	1000

Source: Isaac Hite Papers, Virginia Historical Society, Richmond, VA

large foreign specie loan and use the proceeds to capitalize, along with private investors, a bank – which he calls the Bank of the United States. The bank, by backing its notes and deposits with specie, would create a new and trusted form of money to replace the discredited fiat issues of state and national governments. It would be jointly owned by the national government and private investors, who would share in its profits, and would lend to businesses to promote commerce, and perhaps most importantly, to the national government so that it could carry on the War of Independence (Sylla and Cowen 2017, Chap. 1).

Over the course of following year, Congress sent John Laurens, a fellow army officer and close friend of Hamilton, to France to seek the loan. It also tried to reform its discredited fiat currency by trying to call in old bills and replacing them with new ones at a ratio of 40 old to one new. The reform proved ineffective, inducing a frustrated Congress in early 1781 to appoint Robert Morris to be its Superintendent of Finance, an appointment Hamilton had recommended to a congressman in September 1780.

Hamilton used the occasion of Morris's appointment to write him a long letter in April 1781. He told Morris, who likely needed no convincing, that bringing order to America's chaotic finances was more important than winning a few battles if the Americans were to prevail in their War of Independence. After making rough estimates of Congress's revenues and expenses, he calculated a large deficit that pointed to the need for the foreign loan he had called for a year earlier. He reiterated his earlier argument that the best use of such a loan was to capitalize a national bank. Hamilton cited historical examples of how such banks had strengthened governments that sponsored and used them in their public finances, while also promoting economic growth by furnishing credit to merchants and entrepreneurs. Now, however, he outlined in great detail what the charter of such a bank might contain. Hamilton concluded by saying that his proposed bank would enable the Americans to prevail over the British, who were tiring of the long and inconclusive war, and that the growth of the American economy after independence would make it relatively easy to manage the new nation's large war debts and even redeem them in a matter of decades (Sylla and Cowen 2018, Chap. 3).

Shortly after his letter to Robert Morris, Hamilton left Washington's staff to become a line infantry commander in the Continental Army. In October 1781, he led his troops on a daring nighttime assault on a British fortification at Yorktown. The operation's success made the British position untenable, leading British General Cornwallis to surrender his army. The Yorktown victory proved decisive. As Hamilton had predicted in the Morris letter, the British were tired of the war, and a new government in London entered upon the peace negotiations that recognized American independence in 1783.

While Hamilton was fighting in 1781, Morris told him that he shared his ideas on a national bank. The Superintendent of Finance persuaded Congress by the closing days of 1781 to charter the Bank of North America. Based in Philadelphia, it was on a smaller, more realistic scale than the bank envisioned in Hamilton's letter to Morris. Still, the BNA had trouble attracting sufficient subscriptions from private investors. To complete the bank's contemplated capitalization, Morris had to

purchase a majority of its shares for the national government. He was able to do that using the proceeds of a specie loan from France, which John Laurens successfully negotiated in 1781.

As Hamilton predicted, the new bank helped the nation to finance itself during the war's long closing period. By 1783, a return of confidence allowed Morris to sell the government's shares to private investors, and the BNA became an ordinary commercial bank.

Soon after peace became official in 1783, merchants in New York and Boston sought to reprise the success of the Bank of North America, which the government had granted a monopoly for the duration of the war. Hamilton, by 1784 a rising lawyer in New York, helped that year to draft the constitution of the Bank of New York, an institution in continuous operation ever since (but under the name Bank of New York Mellon since its 2007 merger). The three commercial banks of the 1780s in Philadelphia, New York, and Boston were a portent of what was to come, but their overall economic impact was limited. The Bank of New York was an especially conservative lender at first because the state of New York would not grant it a corporate charter, potentially rendering its stockholders open to unlimited liability if the institution failed. Moreover, the overall economy remained soft in the 1780s as it remained unclear if the new nation could remain united, independent, and solvent.

Constitution and Financial Revolution

Despite the achievement of independence, the 1780s unfolded the weakness of the U.S government under the Articles of Confederation became increasingly manifest. The national government lacked powers of taxation. Instead, it made requisitions of revenue from the states, which often failed to meet them. Lacking revenue, the Confederation Congress could not meet its large debt obligations incurred in the war, and was forced to meet interest payments with more debt.

States had taxing authority and hence were in better shape to meet their debt obligations, which some did. But the states also bickered with one another over trade and territorial claims, and some turned to fiat money in excess, in some cases to inflate away their debt burdens at the expense of creditors.

A taxpayers' revolt, Shays's Rebellion, arose in Massachusetts in 1786, indicating that the new nation, barely a decade old, might be coming apart. Looming over all this domestic dissention, fears that European powers would take advantage of American weakness to reassert their claims prompted nationalist leaders led by Washington, Hamilton, and James Madison to look for a better way.

Their solution was to replace the Articles of Confederation with a new Constitution that greatly expanded the powers of the new federal government it authorized, albeit with many built-in checks and balances designed to prevent abuse of the expanded powers. The Constitution created separate executive and judicial branches of the federal government; under the Confederation, these had been subsumed by the legislative branch: Congress.

The new government had taxing powers concurrent with those of the states along with exclusive powers to tax imports and ship tonnage. In the financial and monetary areas, the Constitution authorized Congress to use federal revenues to pay the debts of the nation, granted it authority to incur further debt on the credit of the United States, and gave it powers to coin money and regulate its value.

To address perceived problems and make its intent perfectly clear, the Constitution explicitly removed the power of states to coin money, issue bills of credit, make anything other than gold and silver a legal tender in debt payments, and pass any laws impairing the obligation of contracts.

Washington became the first president under the Constitution in April 1789. In September of that year, after Congress created the Treasury Department, he appointed Hamilton, his former principal aide-de-camp and comrade in arms, to be the first Secretary of the Treasury (i.e., finance minister). Hamilton, who had a modern understanding of finance and its possibilities far in advance of any other American of his era, proved to be the right man in the right place at the right time. Since his letters on finance of 1780 and 1781, Hamilton had spent a decade refining his plans to modernize America's financial system. As Treasury Secretary, he was in a position to implement the plans. He did so in a whirlwind series of reports – on public credit, on a national bank, on a mint, and on manufactures – prepared at the request of Congress, which then debated his proposals, accepted most of them, and enacted laws to implement them.

The first was the Report on Public Credit of January 1790. It called for federal assumption of the states' war debts into an enlarged national debt, and funding the entire debt including the large arrears of interest at par. "Funding" meant pledging specified tax revenues, mostly those derived from customs and tonnage duties, to servicing and eventually redeeming the debt. Under the plan, a variety of existing federal and state securities constituting the old domestic debt were to be voluntarily exchanged for a package of new Treasury bonds – a 6% bond, a 6% "deferred" bond that paid zero interest for 10 years and then became a 6% bond in 1801, and a 3% bond.

The reason for offering a package of bonds to holders of old debt was to reduce the effective interest rate on the debt from 6%, the original rate promised to investors, to about 4%. The Treasury was empty when Hamilton took office, and it would be several years before tax revenues would prove sufficient to service the debt and fund ordinary government operations. Offering debt holders a reduced rate of interest was effectively a "haircut" in modern terms, but Hamilton made a number of arguments in the Public Credit Report on why they should accept the lower interest payments. He hinted that raising taxes high enough to pay debt holders the 6% they thought they were owed could lead to more taxpayers' revolts like Shays's Rebellion, and that would not be in debt holders' interest. He also offered the debt holders some valuable considerations in return for their accepting the haircut, notably call protection. The deal included a provision that the government could redeem not more than 2% of its outstanding debt in any 1 year, meaning that debt holders were protected from the government taking advantage of lower market rates down the road to borrow at lower rates and redeem all or much of its 6% debt (Sylla 2011; Sylla and Cowen 2018, chap. 7).

Most of the 6% bonds were to pay the original principal of the old federal debt. Other 6 s, some of the deferred 6 s, and the 3% bonds were to cover the assumed state debts and the large arrears of interest on the entire debt that had accumulated since the war. Overall, after the assumption of state debts, the domestic debt of the country stood at some $63–64 million. The voluntary conversion of old debt to new was essentially accomplished by 1795 when Hamilton stepped down from his cabinet position.

The United States also had accumulated a large foreign debt of about $12 million, including arrears of interest, mostly as a result of loans from its ally, France, during the War of Independence. These were paid according to the terms of the original contracts, mostly by taking out new loans in Europe, chiefly Amsterdam, during Hamilton's tenure as Treasury Secretary, 1789–1795. By the time he stepped down, the entire French debt had been repaid, using the proceeds of Dutch loans that matured in the early nineteenth century (and were repaid then).

Congress, after much contentious debate and some notable side deals (Hamilton gained the support of southern-state congressmen for the assumption of state debts, to which most of them were opposed, by agreeing in June 1790 to locate the permanent national capital on the Potomac River, closer to their homes), Congress enacted the essence of Hamilton's debt restructuring plan in the summer of 1790. The three new bonds representing the national government's debt began to be issued in the fall of 1790, and quarterly interest on them commenced in January 1791. Interest on bonds issued to replace assumed state debts began a year later.

As the new bonds were issued, they immediately began to be traded in new and newly energized securities markets organized in Philadelphia, New York, Boston, and eventually other cities. They also began to serve as collateral for loans from banks, and even to pass as a form of near money among merchants, investors, and others. Hamilton had foreseen these developments in his Report, and no doubt was pleased to witness them. This was the birth of the U.S. Treasury debt market, which in time became the largest and most liquid market in the world (Sylla 1998).

Hamilton met the interest payments due on the enlarged national debt promptly, which quickly bolstered the formerly badly damaged credit of the United States government. It was not an easy task because tax revenues remained insufficient to cover the costs until 1793–1794. Hamilton met the deficiencies by borrowing from domestic banks (including, as of 1792, the newly established national bank, the Bank of the United States [BUS], which was a key element of his overall plan, as discussed below) and foreign bankers. Despite the dicey financial situation, Hamilton exuded competence, control, and confidence. This charmed public creditors, who experienced large capital gains. Before 1790, when the federal debt was essentially in default, old U.S. and state debts sold in sporadic markets for small fractions of their face value. By the summer of 1791, the 6% rose to par, with proportionate gains in value for the deferred and 3% bonds. But the new wealth seemingly created out of thin air and the speculation taking place in securities markets infuriated Hamilton's political opponents, who resented his ability to implement his program and assumed his true intentions were nefarious, if not corrupt. The national unity of 1789 would give way to political party formation and competition by 1791 (Sylla 2011).

Debt restructuring and the establishment of public credit was the first key element of Hamilton's plan. The second was the founding of a national bank, the BUS. Hamilton had called for such a bank in 1780 and 1781, and Robert Morris presided over the creation of a limited version of such an institution, the Bank of North America, in 1781–1782. Hamilton's proposal for a much larger national bank, the BUS, came in his Bank Report of December 1790. The report discussed at some length the pros and cons of banking in general, which was a new and controversial subject for most Americans, and referenced history to show that the leading nations of Europe had established such banks, and found that they became sources of government strength and economic prosperity. Among the advantages to the economy would be an augmentation of the money supply by the bank's issue of currency notes and deposits convertible into gold and silver, and an expansion of credit by the bank's lending to creditworthy private borrowers.

The Report touted the advantages of a national bank for aiding the Treasury's financial management and for supporting public credit. Much of the bank's initial capital would consist of recently issued public debt, which would support the government bond market, and the bank would be a source of loans to the Treasury, which very much needed them (although Hamilton did not emphasize this) given its limited revenues in relation to the large financial obligations it had taken on in restructuring the national debt (Sylla 2011).

The Bank Report went on to outline a constitution consisting of 24 articles for the bank, which became the basis of the charter Congress would enact in February 1791 for the first Bank of the United States. The bank would be a very large corporation (in comparison to the few existing banks and corporations then existing) with a capital of $10 million, divided into 25 thousand shares each with a par value of $400. The U.S. government would subscribe for a minority ownership position, 20% of the shares, paying for them with a loan from the bank that would be repaid over 10 years, and it would share in the bank's profits to the extent of its ownership stake. These design features ensured that the bank would get off to a fast start and indicate that it had some responsibilities to be the federal government's bank, although to maintain the bank's independence, the government was to have no role in its management. By having the federal government share in the bank's profits, Hamilton also sent a message to state legislatures: you too can profit by chartering banks and tying them to state public finance. That the states got the message quickly is evident in the rapid expansion of banks chartered by state governments, which both invested in banks and taxed them (Sylla et al. 1987; Sylla 1998).

Private investors would own 80% of the bank's shares, and their subscriptions could be paid one quarter in specie and three quarters in the 6% Treasury bonds recently issued. Hamilton cleverly designed the bank to support the national debt and the debt to capitalize the bank. The private shareholders would elect the bank's directors, who had to be U.S. citizens. Finally, the charter authorized the bank to open branches, called Offices of Discount and Deposit, throughout the United States. State-chartered banks were confined to operating in the states that chartered them, so this provision was important for giving the country a version of nationwide banking in its earliest decades.

Both houses of Congress quickly approved the bank. But a minority in opposition raised the issue of the bank's constitutionality. Among the opponents who deemed the bank to be unconstitutional was James Madison, and he was joined in that view by cabinet officers Thomas Jefferson and Edmund Randolph. President Washington, mulling over whether to sign or veto the bill, asked Hamilton to respond to the opponents, which he did in a powerful, far-ranging defense of the bank and its constitutionality. The president was persuaded by Hamilton's opinion, and signed the bill (Sylla and Cowen 2018).

The BUS's initial offering of shares (actually rights, called scripts, to buy shares in several later installments) took place in July 1791. It sold out quickly, and a speculative bubble and collapse in them occurred over the following weeks, but was contained with deft interventions by Hamilton (Sylla et al. 2009). In December 1791, the bank opened for business at its home office in Philadelphia, and it began to open branches in other cities and states a few months later.

The third key element of Hamilton's plan came in his Mint Report of January 1791. Its main purpose was to define the U.S. dollar as the unit of account and monetary base of the country. Before the Mint Report, the United States had virtually no coinage of its own. Instead it relied on a variety of foreign coins, the most common of which were Spanish-empire minted silver pesos, which Americans since colonial times accumulated in trade and called dollars (Michener and Wright 2005). Not all of these dollars contained the same amount of silver, a point Hamilton addressed in his report. To be a respectable nation, Hamilton argued, the United States needed to have its own unit of account and mint. Should it be based on gold or silver? Hamilton set forth the pros and cons of each metal, in the end opting for bimetallism and a dollar defined in terms of both silver and gold.

Well aware of Gresham's Law, which predicted that if the mint ratios of gold and silver differed from their ratio in markets, the one undervalued at the mint would not be minted, he reasoned that a fairly long-term stability of the market ratio of gold and silver of about 15 to 1 might allow bimetallism to work for the United States. He studied a variety of Spanish dollars to determine that on average they contained 371.25 grains of pure silver, which standard he adopted as the silver dollar of the United States. The 15 to 1 ratio therefore implied that the U.S. gold dollar would contain 24.75 grains (371.25/15) of pure gold. Underlying opting for bimetallism was Hamilton's hope that it would result in a larger money supply and offer more stimulation to economic growth. The rest of his report dealt with technical issues of coinage such as the names and denominations of coins, the amount of alloy in coins, and the organization of the mint and how it would be financed (Sylla and Cowen 2018, Chap. 11).

Congress adopted the essence of the Mint Report in 1792, and launched a mint in Philadelphia. Hamilton hoped that it would supply sufficient U.S. coins within a few years so that foreign coins would no longer need to be legal tender at U.S. ratings. In that he would be disappointed. The mint failed for decades to supply enough U.S. coins to dispense with the use of foreign ones. Only in the 1850s, after the major gold discoveries in California, would the United States end its reliance on foreign coins as legal tender. Until that time, the U.S. money supply would consist of

a few U.S. coins, a larger amount of foreign coins rated in dollar terms, and bank notes and deposits (Wright 2005).

The last of Hamilton's great state papers of 1790–1791 was the Report on Manufactures, delivered to Congress in December 1791. It was his longest report and perhaps in the long run his most influential effort because it mapped out the policies by means of which an undeveloped or underdeveloped economy with the support of enabling governmental policies might develop and grow. It said little about finance, other than to provide arguments and evidence that the restoration of public credit and the national bank resulting from the adoption of his policies were already working to alleviate the young nation's deficiencies of capital and credit that some considered reasons to avoid developing U.S. manufacturing. Congress took no direct action in response to the report, although less than a year later, for reasons unconnected to the report itself, it adopted most of the report's recommended revenue measures.

As Treasury Secretary, Hamilton had the direct authority to recommend and implement policies to establish public credit, launch a central bank, and define the U.S. dollar as the nation's monetary unit. He relied on others to implement the rest of his plan, which called for promoting the development of the U.S. banking system, its securities markets, and its corporations. But his policies induced them to do so.

The creation of $64 million of prime domestic federal debt securities and $8 million of Bank of United States shares in the hands of the public, for example, fostered the development of vibrant securities markets. They sprang up almost immediately in Philadelphia, New York, and Boston, and soon spread to other cities such as Baltimore and Charleston (Wright 2008). The three Treasury bonds and BUS shares became the national market securities in each city market, where they were later joined by the bonds of state governments, and the shares of state-chartered banks, insurance companies, and other corporations (Davis 1917; Wright 2014). Americans were not the only ones to use these markets. Foreign investors were large purchasers, such that by the first years of the nineteenth century they owned more than half of the U.S. national debt and more than half of the shares of the BUS (Sylla et al. 2006). This, too, was part of Hamilton's plan for attracting foreign capital to the United States. Cliometric investigations have uncovered the securities listings and price histories of these early markets, and documented that they grew over time (Rousseau and Sylla 2005; Sylla 2005).

Similarly, Hamilton's BUS, by far the largest corporation of the early United States, prompted states for a variety of reasons to charter an increasing number of banks of their own, as well as insurance companies and other corporations. The United States, which had only three state banks in operation in 1790, had 28 by 1800, more than a hundred by 1810, more than 300 by 1820, and nearly 600 by the mid-1830s (Sylla 1998). In total, some 300 corporations including banks were chartered in the 1790s, and that was just the beginning (Davis 1917). In subsequent decades, as again shown by cliometric research, thousands and tens of thousands more corporations were chartered by state legislatures, and eventually by administrative agencies of state governments (Sylla and Wright 2013; Wright 2014, 2015).

Hamilton's launch of the U.S. financial revolution was complete by 1795, the year he resigned as Treasury Secretary and returned to New York to practice law. His policies had given the country a modern, articulated financial system with several key institutional components, namely strong public finances and public debt management, a central bank (the BUS) with nationwide branches, the U.S. dollar as a new national currency, an expanding banking system, active securities markets and two nascent stock exchanges in Philadelphia and New York, and a growing number of corporations, both financial (e.g., banks and insurance companies) and nonfinancial (e.g., transportation and manufacturing companies). Virtually none of these components existed before 1790. The new financial system financed the country's territorial expansion (e.g., the Louisiana Purchase in 1803), its transportation and industrial revolutions of the early nineteenth century, and its wars. Financial development, as in other places and times, led U.S. economic development and growth.

There were, to be sure, bumps and setbacks. One came in 1792 in the form of the new nation's first financial crisis, which was intimately bound up with Hamilton's financial policies. Speculators fueled by expanding credit, including that of the brand-new BUS, attempted to corner the market in U.S. 6% bonds, which subscribers to BUS stock needed to complete their purchases of BUS shares. The bonds' prices rose well above par early in the year, only to come crashing down to below par in the early weeks of March after one notable speculator defaulted on his obligations, triggering other defaults and panic sales of securities. Hamilton sensed what was happening early in the year, and called for banks to gradually reduce their credit creation. Instead, banks, notably the BUS, stepped on the brakes, contributing to the panic and crisis.

At that point, Hamilton stepped in and managed the crisis much like a modern central banker would. He urged solvent banks to keep up their lending, and promised the banks that their loans to finance merchants' tax payments would not lead to Treasury withdrawals of cash. He organized open-market purchases of substantial amounts of government debt to inject liquidity. He hatched and saw implemented a plan whereby distressed securities dealers were encouraged to collateralize their holdings for bank loans at values Hamilton specified instead of dumping them on the market at fire-sale prices to obtain liquidity. He coupled the plan with a sort of repo guarantee, saying that if a bank making such loans got stuck with the collateral, he would take it off the bank's hands at the values he had specified. These and other measures alleviated the crisis, which blew over in a matter of months with minimal damage to the economy (Sylla et al. 2009).

A greater, if temporary, setback to the financial revolution came in 1811, when Congress failed to renew the 20-year charter of the BUS by the narrowest of margins. It was not because the central bank had failed to do its job of stabilizing the financial system and regulating the banks. All evidence points to it having done a good job of that. The problem was political, not financial or economic. By 1810, states had chartered more than a hundred banks, which both competed with the BUS in many markets and were regulated by it. Some of these banks and the state legislatures which chartered them and invested and taxed them, reasoned that if the BUS was not rechartered, they would get rid of a competitor and a regulator, and once it was gone

they would get the banking business of the federal government, which would no longer have its own bank. It was a win, win, win political opportunity, and it led to the defeat and demise of the first BUS in 1811 (Sylla 2008).

Hamilton was not there to fight this battle. He had died after a duel with Vice President Aaron Burr 7 years earlier, in 1804. It was left to new Treasury Secretary, Albert Gallatin, to fight the battle. Gallatin fought hard and gallantly to preserve the BUS, but in the end he failed.

War of 1812 and Advent of the Second Bank of the United States

The timing of the demise of the first BUS was unfortunate. In mid-1812, barely a year after the bank closed its doors and began to wind up its affairs, the United States declared war on Great Britain in response to Britain's continued predations on American shipping and impressments of crew members of American ships into the British navy. Without a central bank, the country embarked on a war with one arm (finance) tied behind its back. The war itself proved inconclusive; when it ended, the status quo ante bellum was simply restored. But as the war unfolded, financial embarrassments were as numerous as military failures. Even before the war ended in early 1815, national leaders realized that killing the first BUS had been a major mistake, and they moved to create a new version of it.

The absence of a central bank during the War of 1812 had two negative consequences for U.S. finance. First, the government could not borrow from a bank that did not exist. Treasury Secretary Albert Gallatin, anticipating war, had hoped up to 1811 to recharter an enlarged version of the first BUS, one with $30 million of capital instead of the first BUS's $10 million, and to obligate the new bank to lend half of its capital to the government on demand (Edling 2014). In the event of war, short-term loans from the central bank would facilitate a quick mobilization of military resources, just as in 1794, when a BUS loan of $1 million financed the mobilization of forces that put down the Whiskey Rebellion. Central bank loans could be repaid with the proceeds of long-term government borrowing via bond issues as a war continued and wartime finance became regularized. Congress rejected Gallatin's plan in 1811, complicating his task of financing the war a year later.

The second negative consequence of not having a central bank came in 1814 when banks outside New England suspended specie payments at the time of a British invasion of the Chesapeake. Instead of having one national currency managed by a central bank, which also transferred the government's funds throughout the nation (and even internationally) to places where they were needed to make payments, the country had multiple local currencies of varying rates of exchange with the specie dollar. Without BUS notes constituting a national currency, the government was forced to accept local bank notes of varying values in payment of taxes and as subscriptions to its debt issues. That sometimes meant that although the government had a surplus of local currencies in some places, it was short of acceptable means of payment in others. War finance became needlessly complicated and expensive (Edling 2014).

Without a central bank, and in the later stages of the war an inadequate national currency, the government relied heavily on issues of short-term Treasury notes to relieve the money shortage and pay war expenses. The notes bore interest, were payable to the bearer, and were to be redeemed 1 year later. They substituted for an absent national currency. More importantly, they became reserves for the state banks that proliferated during the war era, and served to stimulate an inflationary monetary expansion. The government issued a total of $37 million of Treasury notes during the war, but because of redemptions, the maximum outstanding at any one time was about half of that (Edling 2014).

Even more important than Treasury notes was the issue of long-term bonds. Five loan acts during the war authorized issues of bonds totaling $73 million at face value. But the bonds were not easily sold, and they brought in proceeds that were well below the face value of the bonds. Before the war began, the first loan act in March 1812 authorized $11 million of borrowing. Subscriptions from banks and individual investors totaled only $6 million. Another loan act of early 1813 fared even worse. The government asked for $16 million, but buyers initially took only $4 million. In response, the government reopened the books with a novel approach: the bonds would be auctioned to the highest bidders. That worked. The loan was over-subscribed, but the bonds sold at a discount, 88% of par value. Wealthy individuals acted as nascent investment bankers, purchasing large amounts of the bonds with the intention of reselling them at a profit to smaller investors. A few months after the $16 million loan, a third act authorizing $7.5 million achieved similar results, with investors paying 88.25% of par for the bonds (Edling 2014).

The year 1814 proved more difficult for Treasury finance. A $10 million loan in April filled at 88% of par, but only because a dubious investor made a large subscription on which he later partly defaulted. Around the same time, the British subdued Napoleon in Europe, increasing the chances that they would send more forces to America, prolong the war, make a favorable outcome for the United States less likely, and increase the likelihood that more loans would be needed. Fearing this, investors in the April loan forced Treasury Secretary George Campbell (Gallatin had stepped down to participate in peace negotiations with Britain in Europe) to agree that if another loan later in 1814 were sold below 88% of par, the price of their bonds would be similarly reduced. The investors were prescient. The Treasury requested a new loan of $6 million in August, at the very time British forces were invading the Chesapeake and burning U.S. government buildings in Washington, D.C. The bottom fell out of the bond market, and less than half of the loan was filled at a price of 80% of par. That meant the investors in the April loan had their price reduced from 88 to 80 (Edling 2014).

A new treasury secretary, Alexander Dallas, replaced Campbell. He called for tax increases to service old debts, and a new and vastly larger BUS, capitalized at $50 million, to be the principal source of new loans. In the interim, the Treasury relied on new issues of Treasury notes to pay most of its bills. Fortunately, the war ended at the start of 1815, leaving U.S. government finances in a mess, but relieving the pressures of financing a continuing war. Breathing sighs of relief, national leaders turned attention to undoing the mistake of 1811 by creating a new central bank.

Congress's bill to charter the second BUS became law in April 1816. The new bank, with a capital of $35 million, was essentially an enlarged version of the first BUS. Like its predecessor, its main office was in Philadelphia, and it could open branches throughout the nation. The government again took 20% ($7 million) of the bank's stock; the government had the right to appoint 20% of the directors, a privilege it did not have with the first BUS. Private investors could pay for their $28 million in shares in three installments by tendering 25% of the cost in specie and 75% in U.S. government securities, which were accepted at par even though selling at discounts from par in securities markets. The resulting demand for $21 million of government debt to pay for bank shares led to a rapid rise in federal bond prices, with yields falling from more than 7% to less than 6%. In effect, the second BUS as it was organized absorbed about a sixth of the outstanding U.S. public debt of more than $120 million, an amount greatly enlarged by the War of 1812 borrowing (Knodell 2017). The rapid restoration of the U.S. government's credit, which had become tattered in the war, was a principal achievement of the bank's founding.

The restoration of public credit, however, was not a main reason for founding the bank. Congressional leaders and the Treasury were more interested in restoring a national currency convertible into specie. Such a currency had been lost when, with the first BUS gone, banks outside of New England suspended specie payments in mid-1814. Convertibility was soon restored in the older Atlantic seaboard states, but the West was another story. There the wartime inflation gave way to a postwar boom in land sales, much of which was financed by the credit expansion of state banks, whose interest in, and capability of, maintaining convertibility was highly suspect. When the new BUS pressed western banks to convert their notes to specie by presenting them, it encountered resistance. So it stopped accepting them. That forced the Treasury to revive its wartime practice of selecting state banks to become depositories for Treasury receipts of tax payments and proceeds of land sales. In effect, the second BUS transferred currency risk to the Treasury (Knodell 2017). The Treasury was not happy about the BUS's lack of cooperation, and even unhappier when some of the banks it had selected as depositories for public funds in a form the BUS had declined to accept failed. The BUS had protected itself from those defaults, but by contracting credit in the western land boom it had paved the way for a financial panic in 1819. Westerners, including Andrew Jackson, would remember the Panic of 1819 and the BUS's role in leading up to it when the issue of renewing the BUS's 20-year charter came up in the early 1830s.

After this shaky start, the second BUS settled down to provide the nation with the currency and financial stability that its predecessor, the first BUS, provided from 1792 to 1811. That was particularly true after Nicholas Biddle, a Philadelphia patrician, became its president in 1823. Several notable achievements marked Biddle's presidency. One was establishing a large branch network throughout the country, which helped stabilize U.S. currency and credit everywhere, but especially in the West, where so many state bank failures occurred around the Panic of 1819. Besides smoothly transferring the government's money from where it was (mostly in established eastern financial centers) to where it was needed (often on the rapidly settling western frontier), Biddle's BUS promoted economic expansion and national

economic integration by operating an efficient system of domestic monetary exchanges. A western merchant, for example, could easily use BUS facilities to order goods from an eastern supplier by obtaining a credit from the local BUS branch in the West, and the money would be paid to the supplier by an eastern BUS branch. The goods would be shipped west and sold, allowing the western merchant to repay his credit from the local BUS branch. The entire transaction went smoothly and at a low cost. In the absence of the BUS, such transactions could have required more costly specie shipments or the facilities of private bankers who would have charged more because they lacked the economies of scale achieved by the BUS (Knodell 2017).

A second achievement of the BUS was increased confidence in the currency. The BUS became the principal repository of the nation's specie reserves. When more specie came into the country than went out, BUS reserves rose; when more went out than came in, they fell. In that way, the BUS cushioned a growing number of state banks from large fluctuations in their own reserves. With more stable reserves, the state banks could lend more and create more notes and deposits. Nevertheless, they were restrained from excessive expansion by the BUS, which received state bank liabilities (notes and checks written on deposits) and presented them to the state banks for payment. Banks and entrepreneurs could (and did) complain about BUS restraints on credit expansion, but Knodell (2017) finds no evidence that economic growth was constrained. A positive result was growing confidence in bank currency, for which there is substantial evidence. The money-holding public had a choice of forms in which to hold money, which could be divided between specie and bank liabilities. Over the course of the 1820s and early 1830s, the public chose to hold increasing amounts of money in the form of bank liabilities and to reduce holdings of specie. Americans economized on specie, an expensive form of money, in favor of the cheaper bank money (notes and deposits), in which the second BUS had fostered increased confidence (Engerman 1970; Knodell 2017).

The BUS managed the foreign monetary exchanges of the United States along with internal domestic exchanges. Large amounts of U.S. debt and equity securities were held in Europe, necessitating payments of interest and dividends in foreign financial centers. The BUS had business arrangements with foreign bankers that facilitated these payments. By the late 1820s, the principal of the U.S. national debt was being paid off as well, much of it to foreign investors. The BUS managed all these foreign payments without disrupting the domestic financial system. That, too, increased Americans' confidence in bank money.

In 1825, Britain suffered a major banking crisis with numerous failures of financial institutions. The United States, which had numerous trade and financial ties with Britain, escaped the crisis. Some financial historians credit the second BUS with taking actions to prevent the British crisis from spreading to the United States (Hammond 1957; Govan 1959); others point to instances in which the BUS appeared to be looking out for itself instead of taking actions to stabilize the U.S. financial system (Temin 1969; Knodell 2017). The telling fact in these disputes would seem to be that the U.S. avoided a serious crisis and a long period of stable economic expansion continued (Sylla 2013). On balance, therefore, the BUS acted as a stabilizer.

Despite its economic and financial successes, the second BUS, like the first, became a victim of banking politics. Whig politicians, seeking to embarrass the populist Democrat president Andrew Jackson in 1832, an election year, introduced a bill to recharter the second BUS early, 4 years before the bank's federal charter expired in 1836. The bill passed both houses of Congress, but was vetoed by Jackson, and the bank's backers could not muster the supermajorities needed to override the veto. The bank hoped for a reversal of veto, but when it contracted credit in response to the government's diversion of deposits of revenues to other banks, it lost favor with the public. On the expiration of its federal charter in 1836, the BUS reconstituted itself as a Pennsylvania state bank with a charter that cost it dearly. A few years later, after forays into investment banking, it failed.

The United States would be without a central bank for seven decades, from 1836 to 1914. The price was not reduced economic growth, but decreased economic and financial stability. Financial crises and economic downturns became more frequent during the seven decades without a central bank (Sylla 2013). European nations in those decades introduced new central banks and converted old national banks to central banks. As a result, Europe had far fewer crises. The panic of 1907, which afflicted the United States but largely bypassed Europe, was the last straw. Congress responded with authorizations for emergency currency issues and the creation of a National Monetary Commission, which recommended a new central bank. It came in 1914, in the form of the Federal Reserve System, which now has lasted for more than a century.

Financial Sector Development and Growth to 1836

As Fig. 1 shows, despite the trials and travails of U.S. central banking, the number of commercial banks in operation grew quickly in the first decades of the nation's existence. New entrants vastly outnumbered exits, the first of which occurred with the bankruptcy of three banks controlled by Boston speculator Andrew Dexter in early 1809 (Kamensky 2008). Some early exits were due to voluntary liquidation rather than outright failure; mergers came later.

Early commercial banks had three major sources of funds (types of liabilities): notes, deposits, and owner equity (capital). Occasionally, a troubled bank borrowed from another bank or a rich individual, but such instances were rare. Uses of funds (types of assets) included specie reserves, secondary reserves (generally government bonds), notes of and deposits in other banks, physical property (banking house; land seized for debt), and short-term loans, known as discounts, wherein the bank literally purchased a promissory note, bill of exchange, or other short-term evidence of business debt for its discounted present value. At 6% annual interest, for example, a bank would pay $94.34 today for a $100 note with exactly a year to run ($PV = FV/(1 + i)^n$, where PV = present value or price; FV = future, face, or principal value; i = interest rate; n = number of compounding periods. So in this case, PV = 100/1.06 or $94.34 rounded up to the nearest penny) (Bodenhorn 2000, 2003).

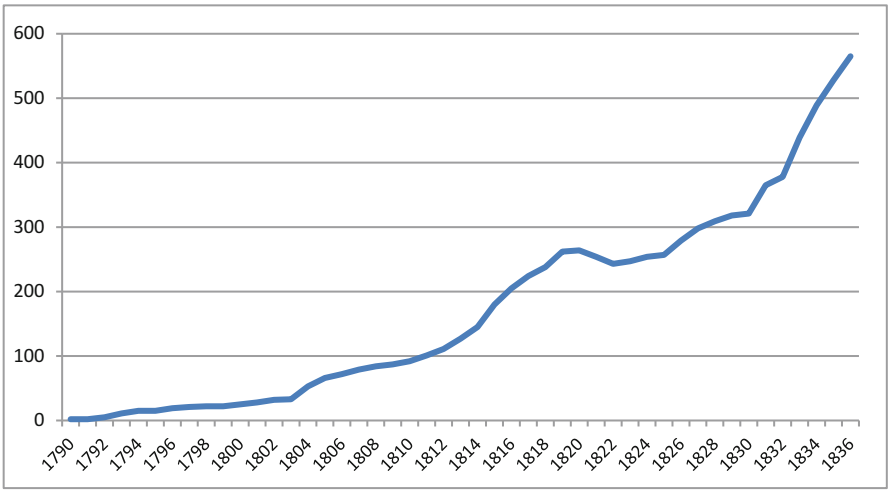


Fig. 1 Number of U.S. Commercial Banks in Operation, 1790–1836. (Source: Weber 2005)

Like banks today, early commercial banks faced interest rate, credit, liquidity, and capital adequacy risks. If interest rates increased, the market value of their secondary reserves dropped. If they discounted notes willy-nilly, defaults cut directly into capital, which, if insufficient to cover losses, meant bankruptcy. Bankers also had to keep sufficient specie reserves on hand to meet depositor withdrawals and noteholder redemptions, so instead of lending money that would fall due in a few, large globs, banks tried to spread repayments out evenly over time so they always had new cash flowing in. They also held liquid assets like government bonds as secondary reserves that could be quickly and easily sold should they need cash to meet unexpected exigencies (Bodenhorn 2000, 2003).

Political opponents and groups that wanted to charter new banks sometimes portrayed early commercial banks as the exclusive clubs of a few wealthy merchants or, in New England, related industrialists (Lamoreaux 1994). Many early commercial banks, though, lent to significant numbers of artisans, farmers, retailers, and manufacturers as well as to large wholesalers. That said, commercial banks were generally not open to consumers, i.e., small depositors and individual borrowers, especially those seeking mortgages. Their needs were met by person-to-person markets, like those for mortgages and ground rents, or new forms of financial intermediaries including lombards, mutual savings banks, joint-stock savings banks, building and loans, and marine, fire, and life insurers (Perkins 1994; Murphy 2010).

Lombards were for-profit financial intermediaries that lent small sums to consumers. The first, the New York Lombard Association, formed in April 1824. Before the end of 1836, eight lombards had formed, two in Pennsylvania and the rest in New York. Their total legally authorized capital ranged from a minimum of $1.5 million to a maximum of $2.6 million (Wright 2015).

The first mutual savings bank formed in the United States was organized in Philadelphia in December 1816 and was quickly emulated in Boston and elsewhere. The first incorporated joint-stock savings bank, the New Orleans Savings Bank, received its charter from Louisiana in March 1827. By the end of 1836, 151 savings banks had formed across the nation, 26 joint-stock, 3 hybrid (a mix of mutual and joint-stock), and the balance mutual. Regardless of their ownership structure, all accepted small deposits, made large mortgage loans, and bought public securities. The joint-stock and hybrid savings banks had a cushion of equity capital that averaged about $75,000 but paid for the privilege in the form of stockholder dividends. Mutuals built up cushions over time out of accrued profits (Wright 2015, 2017).

Fire, marine, and life insurers also invested in mortgages and public securities, including government bonds and corporate stocks and bonds. There were 620 insurance companies incorporated nationwide by the end of 1836. Most were joint-stock, but 103 were mutuals and 39 took a hybrid mutual-joint-stock form. Insurers aided businesses by spreading and pooling risks like shipwrecks and warehouse fires, but also by inducing insureds to mitigate risks through the adoption of best practices, like fire drills, and technologies like fire-resistant construction and reinforced hulls (Wright 2015, 2017).

All the individual lenders, savings banks, and insurers could not satiate the demand for mortgages, so in 1831 the first building and loan association in the United States appeared in Frankford, just north of Philadelphia (Rilling 2001). Building and loans dominated the residential mortgage business throughout much of the country in the late nineteenth and early twentieth centuries before being replaced by savings and loans during the Great Depression (Mason 2004).

Conclusion

From its humble colonial origins, where most households provided basic banking services, to the horrors of wartime hyperinflation, the U.S. financial system emerged in modern form soon after ratification of the Constitution in 1788, thanks in large part to the policies of Treasury Secretary Alexander Hamilton, who unleashed the power of private investors by restoring their confidence in the national currency (a specie-defined dollar) as well as in public credit. From the solid foundation established in the early 1790s, which included a central bank that on occasion acted as a lender of last resort and helped to maintain the national government's credit, financial innovators built a system of commercial, investment, private, and savings banks and fire, marine, and life insurers.

Although imperfect, the system was far from primitive. It did much to supply the money and credit that stimulated rapid U.S. economic growth and development in the nineteenth century, including the agricultural, market, transportation, and industrial revolutions (Rousseau and Sylla 2005; Wright 2017). A detailed, quantitative comparison of the U.S. and British systems as they developed over the four decades from 1790 to 1830 concludes that the United States had the better, more modern system overall (Sylla 2009). Britain's deserved reputation as the financial leader of

the nineteenth century came because its system improved after 1825, while the U.S. system retreated after Jackson's veto of the bill to recharter the second BUS (Sylla 2008).

References

Anonymous (nd). Boston Weekly News Letter, 30 July 1730; [London] *Echo*, 16 September 1730

Anonymous (nd) The following is a speech, made at the delivery of a gold piece, Call'd a Johannes, to a merchant for debt. Connecticut Gazette, 23 January 1768

Anonymous (nd) New Jersey Gazette, 30 January 1786

Anonymous (nd) New Jersey currency question. Record Group 23, Box 4, New Jersey Historical Society, Newark

Anonymous (nd) Isaac Hite papers. Virginia Historical Society, Richmond

Bodenhorn H (2000) A history of banking in antebellum America: financial markets and economic development in an era of nation-building. Cambridge University Press, New York

Bodenhorn H (2003) State banking in early America: a new economic history. Oxford University Press, New York

Brock L (1975) The currency of the American colonies, 1700–1764: a study in colonial finance and imperial relations. Arno Press, New York

Davis J (1917) Essays in the earlier history of American corporations, 2 vols. Harvard University Press, Cambridge

Edling MM (2014) A Hercules in the cradle: war, money, and the American state, 1783–1867. University of Chicago Press, Chicago

Engerman SL (1970) A note on the economic consequences of the second Bank of the United States. J Polit Econ 78(4):725–728

Flesher DL (1979) Barter bookkeeping: a tenacious system. Account Hist J 6(10):83–86

Govan TP (1959) Nicholas Biddle: nationalist and public banker, 1786–1844. University of Chicago Press, Chicago

Hammond B (1957) Banks and politics in America from the revolution to the Civil War. Princeton University Press, Princeton

Kamensky J (2008) The exchange artist: a tale of high-flying speculation and America's first banking collapse. Viking, New York

Knodell JE (2017) The second Bank of the United States: 'central' banker in an era of nation-building, 1816–1836. Routledge, New York

Lamoreaux N (1994) Insider lending: banks, personal connections, and economic development in industrial New England. Cambridge University Press, New York

Lindert PL, Williamson JG (2016) Unequal gains: American growth and inequality since 1700. Princeton University Press, Princeton

Mason DL (2004) From buildings and loans to bailouts: a history of the American savings and loan industry. Cambridge University Press, New York

Michener R (2003) Money in the American colonies. In: Whaples R (ed) EH.Net Encyclopedia. [Internet]. [updated 2011 Jan 13; cited 2017 May 24]. Available from: http://eh.net/encyclopedia/money-in-the-american-colonies/

Michener R (2015) Redemption theories and the value of American colonial paper money. Financ Hist Rev 22(3):315–335

Michener R (2017) Comment on Perkins's 'conflicting views on fiat currency . . .' University of Virginia working paper

Michener R, Wright RE (2005) State "currencies" and the transition to the U.S. dollar: clarifying some confusions. Am Econ Rev 95(3):682–703

Murphy SA (2010) Investing in life: insurance in antebellum America. Johns Hopkins University Press, Baltimore
Perkins EJ (1994) American public finance and financial services, 1700–1815. Ohio State University Press, Columbus
Rilling D (2001) Making houses, crafting capitalism: builders in early Philadelphia, 1790–1850. University of Pennsylvania Press, Philadelphia
Roney JC (2014) Governed by a spirit of opposition: the origins of American political practice in colonial Philadelphia. Johns Hopkins University Press, Baltimore
Rousseau PL, Sylla R (2003) Financial systems, economic growth, and globalization. In: Bordo MD, Taylor AM, Williamson JG (eds) Globalization in historical perspective. University of Chicago Press, Chicago
Rousseau PL, Sylla R (2005) Emerging financial markets and early US growth. Explor Econ Hist 42:1–26
Sylla R (1998) U.S. securities markets and the banking system, 1790–1840. Fed Reserve Bank St. Louis Rev 80(3):83–103
Sylla R (2005) Origins of the New York Stock Exchange. In: Goetzmann WN, Rouwenhorst KG (eds) The origins of value: the financial innovations that created modern capital markets. Oxford University Press, Oxford/New York
Sylla R (2008) The political economy of early US financial development. In: Haber S, North DC, Weingast B (eds) Political institutions and financial development. Stanford University Press, Stanford
Sylla R (2009) Comparing the UK and US financial systems, 1790–1830. In: Atack J, Neal L (eds) The origins and development of financial markets and institutions. Cambridge University Press, Cambridge
Sylla R (2011) Financial foundations: public credit, the national bank, and securities markets. In: Irwin D, Sylla R (eds) Founding choices: American economic policy in the 1790s. University of Chicago Press, Chicago
Sylla R (2013) Entral banking and the incidence of financial crises. Financ Hist 108:20–23
Sylla R, Cowen DJ (2018) Alexander Hamilton on finance, credit, and debt. Columbia University Press, New York
Sylla R, Wright RE (2013) Corporation formation in the United States, 1790–1860: law and politics in comparative contexts. Bus Hist 55(4):653–669
Sylla R, Wallis JJ, Legler JB (1987) Banks and state public finance in the new republic, 1790–1860. J Econ Hist 47:391–403
Sylla R, Wilson JW, Wright RE (2006) Integration of trans-Atlantic capital markets, 1790–1845. Rev Financ 10:613–644
Sylla R, Wright RE, Cowen DJ (2009) Alexander Hamilton, central banker: crisis management and the lender of last resort during the US panic of 1792. Bus Hist Rev 83:61–86
Temin P (1969) The Jacksonian economy. Norton, New York
Weber W (2005) Early state banks in the United States: how many were there and when did they exist? Federal Reserve Bank of Minneapolis working paper 634
Wright RE (2005) The first Wall Street: Chestnut Street, Philadelphia and the birth of American finance. University of Chicago Press, Chicago
Wright RE (2008) One nation under debt: Hamilton, Jefferson, and the history of what we owe. McGraw-Hill, New York
Wright RE (2014) Corporation nation. University of Pennsylvania Press, Philadelphia
Wright RE (2015) US corporate development [Internet]. http://repository.upenn.edu/mead/7/
Wright RE (2017) Financing U.S. economic growth, 1790–1860: corporations, markets, and the real economy. In: Rousseau P, Wachtel P (eds) Financial systems and economic growth: credit, crises, and regulation from the 19th century to the present. Cambridge University Press, New York

Cliometrics and Antebellum Banking

Hugh Rockoff

Contents

Abstract

The three decades before the Civil War in the United States provides a remarkably rich source of informative experiences for the student of banking and finance. The Second Bank of the United States was wound up after the famous "Bank War." Inflows of gold and silver and banking panics disrupted the financial system. And the states followed a wide array of models when they chartered and regulated banks including the famous free banking laws. This paper describes the evolution of scholarly thinking about this era and the many contributions made by cliometricians.

Keywords

Antebellum banking · Second bank of the United States · Bank war · Free banking · Gold rush · Andrew Jackson · Panic of 1837 · Panic of 1857

H. Rockoff (✉)
Department of Economics, Rutgers University, New Brunswick, NJ, USA
e-mail: Rockoff@econ.rutgers.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_59

Introduction

This paper reviews the work by cliometricians on banking in the United States during the 30 years before the Civil War. I have used the term antebellum to describe this period, although the term is somewhat flexible. Sometimes it is used to refer to the whole period from the War of 1812 to the Civil War, and at other times to just the two decades before the Civil War. One reason for limiting the coverage of the essay to this period is pragmatic: another essay in this volume will cover banking up to 1830. But 1830 to 1860 also forms a coherent historical epoch. The demise of the Second Bank of the United States in the mid-1830s led to a wide variety of experiments with banking systems. Cliometricians have therefore found this period to be especially fertile for testing ideas about the effects of alternative banking arrangements.

Although monetary economics has long been quantitative, banking history, to a surprising extent, has not. American economic historians have often written about banking history while making very little use of quantitative data or methods. Fritz Redlich's (1947) *The Molding of American Banking: Men and Ideas* is still required reading for anyone who wants to understand the history of banking during the antebellum era. But it contains no charts, tables, or equations. Bray Hammond's (1957) brilliant *Banks and Politics in America from the Revolution to the Civil War* was awarded the Pulitzer Prize for History in 1958. It is still one of the key texts on antebellum banking, yet it also contains no charts, tables, or equations. Clearly, there was room for studies that made more use of the quantitative record. Here I will describe how a generation of cliometricians enriched the work of these pioneers. First, however, we need to be clear about what a bank was and how it functioned during the antebellum era.

The Antebellum Bank Balance Sheet

Banks are financial intermediaries that issue short-term liabilities and invest in longer-term assets. The customers of banks gain because they value the liquidity from holding an asset that can be quickly turned into cash. The price they pay for liquidity is the low rate of interest they receive. The bank profits from creating liquidity because the longer-term assets they buy pay a higher rate of interest than the short-term liabilities they issue. Table 1 shows a stylized balance sheet of an antebellum bank.

Most items will be familiar to someone familiar with modern bank balance sheets. Deposits were subject to check, although to make a long-distance payment a depositor would have to purchase a draft on a bank familiar to his or her counterparty (James and Weiman 2010). For example, someone living in Springfield, Ohio, would not normally be able to make a payment in New York by drawing a check on our hypothetical Springfield bank. Instead, they would purchase a draft on a New York bank. The capital, surplus, and loan loss reserve accounts were also similar to accounts on a modern balance sheet. In some dodgy cases, the capital would be "paid" with loans from the bank itself, so that the true amount of capital invested in the enterprise was exaggerated.

Table 1 Balance sheet of the first cliometric bank of Springfield, Ohio, July 4, 1840

Assets		Liabilities and capital	
Reserves	$10	Notes	$100
Loans	$200	Deposits	$100
Securities	$100	Loan loss reserves	$50
Real estate	$50	Surplus	$50
Bank building and furnishings	$40	Capital	$100
Total	$400	Total	$400

Banking theory at the time held that ideally these loans should be short-term bills that originated in "real" transactions. The famous "real bills doctrine" can be traced back at least to Adam Smith who gave the idea its name. For example, a miller draws a 90 day bill to buy wheat from a farmer; and the farmer discounts the bill at a local bank. The bill is a safe investment for the bank because it is "two-name" paper and because it had originated in a real transaction – a sale of wheat – so that the miller would be able to pay the bill when the flour made from the wheat was sold. It was liquid because it would be paid in 90 days. These bills, however, because of their safety and liquidity, often paid low rates of interest and banks frequently made longer-term loans. And of course, nominally short-term loans could be renewed again and again, in effect becoming long-term loans. Many of these loans in the antebellum era, as Lamoreaux (1994) showed, were "insider loans." The bank would make loans to shareholders who often were trustees or administrators of the bank. Obviously there was a danger that insiders would exploit and weaken the bank by making excessive loans to themselves at unrealistic rates. But Lamoreaux shows that the system often worked surprisingly well: by lending to insiders the bank solved the problem of securing adequate information about the worth of the borrower.

The securities of the banks would normally be government bonds, often bonds issued by the state or municipality where the bank was located. Under the free banking laws, discussed in more detail below, banks were required to hold bonds to protect the value of the notes they issued. The reserves of the bank were, first of all, specie (gold and silver coins). For the first part of our period, they were usually silver, but gold became important after gold was discovered in California. Banks also considered deposits in other banks to be part of their reserves.

Bank Notes

Bank notes are the least familiar and the most controversial item on the antebellum balance sheet. Private Banks at that time issued paper currency that circulated from hand to hand like modern Federal Reserve notes. This was true not only in the United States but around the world. Today, only a few privately owned banks still issue notes intended to circulate from hand to hand. Typically, these notes promised to pay legal tender on demand when presented at the bank's office. In New England, as will be discussed below, country bank notes could be redeemed at the Suffolk Bank in Boston.

Many economic historians have condemned private note issues. Why? For one thing, banks sometimes failed and people who held the notes were left holding notes of little or no value. Some of the note holders may have accepted notes passively as part of a minor transaction. In situations of this sort, looking into the condition of the bank and deciding whether or not to accept a note would have been impractical. Moreover, with so many banks issuing large numbers of notes, it has been claimed that counterfeiting was made easy. If someone offered a note, how was the poor merchant to know whether it had been issued by a strong bank, a weak bank, a bank that had already failed, or whether, indeed, it was a counterfeit? Mihm (2007), consistent with much earlier writing, maintains that this was a disastrous state of affairs curable only by the nationalization of the currency. Bank note reporters that listed discounts to be applied to notes from distant banks as well as descriptions of the many altered (10s changed to 100 s) and counterfeit notes circulating were regularly published. The Bank note reporter certainly helped the merchant in deciding whether a note should be accepted in payment. But using them took time, and in the nature of things the information they provided was imperfect. It was indeed a strange state of affairs to someone familiar with today's safe and universally accepted notes. But at the same time the cost of using the currency for most users is easily exaggerated. In Chicago, one would typically use Chicago notes. If someone arrived with New Orleans notes they could exchange them, although the New Orleans notes would be discounted by 2% or 3%. It was a fee on the same order of magnitude that one might pay today at an ATM machine. There were many counterfeits and altered notes one had to be wary of, but not very long ago merchants kept lists of phony credit card numbers. It was an inconvenience, but in my judgment the social costs were limited.

The lists of discounts in the bank note reporters have provided grist for the cliometrican's mill. In one of the first studies, Gary Gorton (1996) showed that the discounts on notes issued by newly established banks were high, but that they then declined as note brokers gained confidence in the soundness of the bank. Gorton argued that this seasoning of notes was an important factor discouraging bad banking; wildcat banking as it was known. Bodenhorn (1998) reinforced Gorton's point by showing that the notes of banks that were about to fail went to high discounts. Jaremski (2011) continued the study of the discounts showing, among other findings, that the discounts reflected evaluations of the stability of state banking systems. All in all, these studies show that the banknote discount market was much like modern markets for other types of financial assets.

The requirement that notes and deposits be redeemable in specie set limits on how much exchange rates could vary between regions. For example, if a $1 note issued by a Louisiana bank found its way to Chicago, it would be discounted. It might sell, for example, for $0.97 in notes issued by Chicago banks, "current money" as it was called, but the price would be unlikely to fall, say, to $0.50 because anyone could buy the note and send it to New Orleans where it could be redeemed (exchanged) for $1.00 coin. A profit opportunity of that magnitude would not last for long. The price of the bank note would rise. Nevertheless, Shambaugh (2006) shows that there was enough flexibility in interregional exchange rates that some states could and did follow somewhat independent monetary policies.

The United States became a true multiple currency area with flexible rates during the Civil War. The South, as is well known, had a separate currency; and during the War and after until 1879, the Eastern part of the United States was on the greenback standard while the Pacific Coast remained on gold, and the rate of exchange between the two regions varied. National Banks in the two regions even issued different types of bank notes. Notes of National Banks in the East were redeemable in Greenbacks; Notes of National Banks on the Pacific Coast were redeemable in gold (Greenfield and Rockoff 1996).

The Structure of the Banking System in the Antebellum Era

The banking system during the antebellum era was fragmented into a wide variety of institutional and regulatory regimes. At the beginning of the period, the United States had a "central bank," the Second Bank of the United States. Central bank is in quotes because there is some debate about the extent to which it resembled a modern central bank. It was a reincarnation on a larger scale of the First Bank of the United States. Like its predecessor, it was modeled to some extent on the Bank of England and the large Scottish banks, the Bank of Scotland, and the Royal Bank of Scotland. It had branches throughout the country, and a dominant role in the payments system. But it is not clear that it ever operated as a lender of last resort or that it took responsibility for macro-economic variables such as the price level or rate of exchange. The Second Bank of the United States was forced to give up its federal charter in 1836, a loser in the great "Bank War" between President Andrew Jackson and Nicholas Biddle, the president of the Second Bank. Jackson led the Democratic Party which was strongly supported on this issue by state banks angered by competition from the Second Bank, and by Western entrepreneurs and farmers hoping for lower interest rates. Biddle was backed by the Whigs who were strongly supported by Eastern commercial interests. Politics, of course, is always complicated. Frontiersman Davy Crockett broke with Jackson on the Bank and other issues; no doubt from principle, although there were also his difficulties in repaying a loan from the Bank which was ultimately forgiven (Catterall 1903). The Second Bank struggled on for a short time under a charter issued by Pennsylvania, but failed in 1841.

The establishment and regulation of banks then fell to the states. They followed a surprisingly diverse range of models. Some states continued chartering banks one by one through legislative acts, some created publicly owned monopolies, some allowed free banking (more on this below), and some even prohibited banking. Indiana created a system of independent but mutually cooperative banks that has been praised as a model for achieving stability (Calomiris 1990). There were also savings banks, mutual savings banks, private banks of various sorts, and trust companies. Hammond (1948) and Bodenhorn (2003) survey the landscape. All of these institutions deserve study, but for the most part, cliometricians have focused on the travails of the commercial banks and the Second Bank of the United States.

The End of the Second Bank and the Jacksonian Inflation

The First Bank of the United States, like other central banks in that era, had a renewal provision in its charter. The renewal provision served several purposes: it assured critics of the Bank that they would have a chance to get rid of the bank if it did not live up to the claims of its advocates, and it provided an occasion for extracting some money from the Bank in exchange for renewal. In 1811, when the renewal date was reached, the charter, after a bitter controversy, was not renewed. The United States was left during the War of 1812 without a central bank. After the War, the deranged state of the currency led to calls for a Second Bank of the United States. And in February 1816, a Second Bank of the United States was established, in many respects simply a larger version of the First Bank.

Many traditional banking historians praised the Second Bank for its management of the financial system. Jane Knodell (2017) used cliometric methods to reinforce the traditionally favorable view while arguing that at points the traditional historians went too far. Knodell found that the Second Bank had indeed played an important role in reducing the costs of interregional payments, and that the demise of the bank was a backward step in this regard. After its demise, the financial system had to wait until the issue of the greenbacks and the establishment of the National Banking System during the Civil War to achieve a low uniform price for domestic exchange (Weiman and James 2007). While the costs of making interregional payments remained high after the demise of the Second Bank, it does appear that regional interest rates converged, leading to the conclusion that over the long run the capital market was integrated during the late antebellum era (Bodenhorn and Rockoff 1992; Bodenhorn 1992). Despite some earlier claims, however, Knodell shows that the Second Bank differed in important ways from a modern central bank. It did not act as lender of last resort or set monetary policy to maximize macro-economic variables.

The most important question concerning the Second Bank, at least from the point of view of political history, was whether Andrew Jackson's successful attack on the Second Bank produced a bank-driven inflation. The argument was that the demise of the Second Bank allowed banks to increase their lending, creating a flood of paper money and deposits that produced rising prices and speculative excesses, particularly real estate speculation. And that in turn led ultimately to the revulsion and inevitable collapse: the Panic of 1837, and the depression that followed. In terms of the balance sheet shown in Table 1, the argument was that banks had been constrained by fear of attacks by the Second Bank. The Second Bank, it was claimed, would collect the notes of banks that lent too much given their resources and present the notes for redemption. Fear of this kind of attack kept the banks in line. Freed of this constraint, banks issued notes and created deposits in order to make more interest-earning loans.

The first cliometrician to challenge this view was George Macesich (1960). Macesich's paper began life in Milton Friedman's famous Workshop on Money and Banking at the University of Chicago. Macesich compiled data on the stock of money, the monetary base (gold and silver coins held by banks or in general circulation), price indices, exchange rates, and imports and exports for the period

1834–1845. His main findings were that changes in the stock of money were produced mainly by fluctuations in the balance of payments, and that changes in money were highly correlated with changes in prices, confirming the quantity-theory explanation of the inflation.

Peter Temin (1968) built on Macesich's study. Then, in *The Jacksonian Economy* (1969), a book that became one of the classic texts of the "New Economic History," he offered a new interpretation of the underlying causes of the increase in the stock of money; an interpretation that he showed had many important implications for political as well as economic history. Some of the monetary data that Temin (who used a somewhat different methodology from Macesich) put together as well as data on prices and output are shown in Table 2. Temin's construction of the antebellum stock of money, I should note, built on the pioneering work of Fenstermaker (1965). I have made the table easier to read by dividing each observation by the value of the series in 1829 and multiplying each resulting observation by 100.

It is obvious that there was a very rapid increase in the stock of money shown in the first column of data. The stock of money rose a remarkable 13.8% per year between 1829 and the peak in 1836. The second column shows the ratio of the total stock of money to the amount of coin in the economy, what is commonly known as the "money multiplier." The money multiplier is determined by the extent to which the public trusts the banks and holds notes and deposits rather than coin and the extent to which the banks create notes and deposits from their reserves. The more notes and deposits banks create for each dollar of reserves, other things equal, the higher is the money multiplier. As can be seen in Table 2, the multiplier did rise from 1829 to 1831; but then it declined. If we look at the whole period from 1829 to 1836, the behavior of the money multiplier suggests, and an examination of bank reserve ratios confirms, that the increase in the stock of money was not fueled by the willingness of banks to create more notes and deposits for each dollar of coins held in reserve. The explanation for the increase in the stock of money then is that it was fueled by an increase in the

Table 2 Money, prices, and output in the 1830s

Year	Money	The ratio of money to coin (the money multiplier)	Prices (the GDP deflator)	Real GDP
1829	100	100	100	100
1830	109	112	98	110
1831	148	162	97	121
1832	143	152	99	129
1833	160	129	98	137
1834	164	106	98	135
1835	234	119	104	144
1836	263	119	111	150
1837	221	83	112	150
1838	229	87	110	152
1839	205	81	113	164

Sources: Carter et al. (2006): series by column Cj22, Cj23, Ca13, and Ca9

amount of coin in the economy. The finding that increases in the money multiplier did not explain the increase in the stock of money, Temin pointed out, contradicted the traditional story that the banks had run wild once they were freed from the watchful eye of the Second bank. Temin then looked at where the additional coin was coming from, and showed that part of the story was an influx from Mexico and a decrease in the traditional outflow to China. The decreased flow to China in turn, Temin contended, was a consequence of the deterioration of the Chinese balance of payments resulting from the expansion of the opium trade. In a related paper, Rockoff (1971) argued that the Mexican influx was the result of capital flight. To economic historians this was exciting stuff. Traditional US-centric stories proved to be mistaken. And the key to overturning the traditional interpretation was to look at the numbers. Previous generations of historians, the cliometricians claimed, had found lots of stories that seemed to support the case against Jackson and his attack on the Bank of the United States, but they had missed the key underlying forces.

The third column of Table 2 shows the GDP deflator. Evidently, inflation during the Jacksonian inflation was rather moderate by today's standard, averaging less than 2% per year between 1829 and 1836. Inflation may have seemed worse to the Jacksonian generation than it would to ours because it followed a decade of moderate deflation. It is also possible that in talking about inflation contemporaries had asset prices in mind, such as the prices of land and slaves – the latter rising about 9% per year (Carter et al. 2006, series Bb210) between 1829 and 1836 – that we donot include in GDP. The last column of Table 2 shows that while inflation was moderate by modern standards, real output rose rapidly, averaging an increase of more than 5% per year from 1829 to 1836.

Temin also reexamined the Panic of 1837, which brought the Jacksonian inflation to a close. Again, he stressed the role of international developments, in this case the decision by the Bank of England to raise its policy lending rate. The Bank of England had not been ignored in traditional accounts, but those accounts had focused on American developments: real estate speculation, President Jackson's "specie circular" (an order requiring payment for public land in specie), and the spectacular failures in March 1837 of Hermann, Briggs, & Co., a cotton broker in New Orleans, and J.L. and S. Joseph, an investment house in New York. Subsequently, Rousseau (2002), based on a careful review of regional bank reserves and specie flows, showed that the distribution of the federal budget surplus to the states had a far larger impact than historians had previously thought.

The Panic of 1837 was followed by further panics in 1839, 1854, and 1857. The Panic of 1839 appears to have been triggered by the suspension of specie payments by the United States Bank of Pennsylvania, the successor to the Second Bank of the United States, in October 1839. John Wallis (2001) showed that the Second Bank had gone overboard funding infrastructure improvements, mainly canals, which proved unprofitable.

The Panic of 1854 was mainly a regional disturbance. Banks in Cincinnati were badly hit, and national financial markets were disturbed, but the financial system recovered quickly. The Panic of 1857, however, was another story. This panic was triggered by the failure of the Ohio Life Insurance and Trust Company in August.

Bank runs and suspensions throughout the country followed. The economy slowed and there appears to have been severe distress among some groups, although judgments must be tentative because statistics on unemployment, national income, and so on, despite heroic efforts by cliometricians, are subject to substantial margins of error. The available statistics show that real GDP per capita fell about 0.2% between 1853 and 1854, but fell 2.3% between 1856 and 1857. An index of industrial production rose 1.94% between 1853 and 1854 and another 0.50% between 1854 and 1855, but fell 2.21% between 1856 and 1857, and fell another 6.09% between 1857 and 1858.

Despite the lack of data on a modern scale, cliometricians have made considerable progress in understanding the panic. Calomiris and Schweikart (1991) examined the spread of the crisis and showed that states that permitted branch banking or other forms of cooperative banking systems did better than states with unit banking systems. Kelly and Ó'Gráda (2000) and ÓGráda and White (2003) analyzed the records of the Emigrant Industrial Savings Bank in New York, where Irish immigrants saved, and showed that the response of these immigrants to panics reflected their financial sophistication and social ties that in some cases stretched back to Ireland.

The Panic of 1857 had political as well as economic consequences (Huston 1987). In the North, the Panic strengthened the hand of the newly formed Republican Party. The Panic, Republicans claimed, was the fault of the established parties. Their policies, such as a high tariff, would help to restore prosperity. In the South, the Panic strengthened the hand of the radicals who wanted to secede from the Union. After all, the secessionists said, the recession was less severe in the South than in the North because of the South's near monopoly of cotton production. Whatever suffering the South felt was the result of being tied to the North and its Wall Street speculators. We would be better off if we separated from those people.

Free Banking and Wildcat Banking

With the Second Bank gone after its defeat in the Bank War, it was left, as noted above, to the states to regulate banking. Many plans were tried, but the "free banking law" was the plan for establishing and regulating banks that has attracted the most attention from cliometricians. Under a free banking law, anyone could establish a bank anywhere within the boundaries of the state provided the bank met certain basic requirements, such as capital, note issue, and reporting. The most important of these requirements was that any notes that were issued by the bank had to be backed by government bonds. The exact rules about the backing for notes varied from state to state. Typically, there was a designated list of eligible bonds that included, understandably, the bonds of the state issuing the free bank charter but that often included municipal bonds, bonds of other states, and federal bonds as well. The laws also described how the bonds were to be valued for the purpose of issuing notes. The bonds backing the notes were held by a state official. If the bank issuing the notes went bankrupt, the state was required to sell the bonds and reimburse the holders

of the notes. Thus the system, it was hoped, would provide both the benefits of competition to borrowers and depositors while providing for the safety of noteholders. Note holders, it was thought, were different from other bank creditors because they were passive creditors of a bank. A depositor chose to put their funds into a particular bank, and so if the bank failed, it could be said that they had not done proper diligence and should suffer the loss. On the other hand, a noteholder may have simply accepted the note in as part of a small transaction and could not be expected to investigate the character of the bank issuing the note. The baker can decide where to deposit his money. But whether to accept a bill in the course of selling a loaf of bread is another matter. The baker could consult a "bank note reporter" to see if the note was on a list of bad notes, but the time and inconvenience of doing thorough diligence for currency used in small transactions was prohibitive.

Under the traditional system of chartering, promoters had to get a charter from the state government as a specific legislative act. The traditional system had obvious weaknesses. There was an incentive for would-be bankers to bribe legislators to get the charter they wanted; and an incentive for existing banks to lobby against the issue of new charters. Anna Schwartz (1947), in the first published paper in her distinguished career, explored how these incentives played out in Pennsylvania at the end of the eighteenth and beginning of the nineteenth centuries. Bodenhorn (2006, 2008) explored corruption and politics in bank chartering as they evolved in New York during the antebellum era. Calomiris and Haber (2014) view the free banking law, more generally, as the product of what they dub the "game of bank bargains," in which bankers bargain with the state for privileges. In this case would be bankers won some important privileges, for example, the right to enter new markets as they saw fit, but in exchange the state got a stronger market for its bonds, created by the requirement that free bank notes be backed by bonds. States, of course, had many other ways of extracting resources from banks. Sylla et al. (1987) show that levies on banks became a major source of revenue for state governments before the Civil War. Free banking laws, it should be noted, were also part of a larger movement toward free incorporation of all types of businesses. Indeed, free banking, for all that it reflected the interests of bankers and the governments that created and regulated banks, also reflected the spirit of an age that emphasized opening political and economic opportunities traditionally reserved for the rich and well-connected to the common man. Free banking was especially attractive in regions of new settlement. These states and territories were filled with new, small towns that hoped to grow into great metropolises, and they badly wanted banks because they believed that banks would be engines of growth.

Something called free banking – sometimes similar to the U.S. model, sometimes very different – was adopted in many other countries in the nineteenth century. The Scottish case in particular has often been explored in part because it was celebrated by Adam Smith in the *Wealth of Nations*. Space does not permit an examination of those cases here. Selgin and White (1994) provide a good guide to the literature available when they wrote and put it within the context of the long debate about the appropriate role of competition in banking. Briones and Rockoff (2005) also summarize a good chunk of the historical record.

The free banking laws have been of great interest to cliometricians. In several states, the passage of a free banking law produced a rash of new banks and often, not long after, many of those banks failed. The phenomenon was dubbed "wildcat banking." The origins of the term are uncertain. One interesting although possibly apocryphal story is that the banks were located in remote areas to forestall attempts to redeem their notes – wildcat country. In other words, the banks appeared to be simply note-issuing machines set up to buy eligible bonds rather than intermediaries intended to provide credit for the local community. The story about free banking producing wildcat banking has been important for economists because of the lesson it might hold for the more general question of the limits of *laisse-faire* as a policy for banking. For example, David Alhadeff (1962) claimed that entry into banking, historically, was restricted in the United States because of the danger of overbanking if entry was easy; with some of the worst abuses happening during, as he (1962, 248) put it, in "the 'wildcat banking' period." An exchange in the *Journal of Political Economy* is of special interest because it reveals the opinions of two of the leading monetary economists of the era. Allan Meltzer (1967), a leader of the monetarist school, proposed reducing regulation of banks to increase competition and improve their performance. In commenting on Meltzer's proposal James Tobin (1967, 508), a leader of the Keynesian school of macroeconomics who received the Nobel Prize in 1981, answered Meltzer in part by pointing out that "The United States had a history of wildcat banking that no one would wish to repeat."

Perhaps I am taking too much credit for myself, but I believe that Rockoff (1974, 1975a) inaugurated the cliometric study and revision of the traditional view of wildcat banking. I argued that, first, wildcat banking consisting of a rash of banks that simply bought bonds and issued notes was a relatively rare phenomenon that afflicted only some of the states that adopted free banking laws, and that in other states, where the law was more carefully drawn, notably New York, it had worked well. Indeed, in some cases, such as Massachusetts, where the bond requirements for issuing notes were very demanding, little banking was done under the law. Second, that it was usually a failure to require adequate security for notes that led to a rash of wildcat banks; for example, a decision to accept bonds at par when they were selling below par. And third, that the overall losses from wildcat banking were small. I had expected to be criticized by defenders of the traditional view that free banking had been a disaster. For that reason, following the lead of Robert Fogel (1964), I tried to make sure that my estimates of the amount of wildcat banking, and the resulting losses were upper bound estimates. Surprisingly, most of the subsequent literature criticized me from the other side by arguing that the incidence of wildcat banking and the resulting losses were even lower than I had suggested. Thus, most of the subsequent literature, although critical of my work, strengthened the case for believing that most wildcat banking was a minor phenomenon. I summarized my reaction to the literature that emerged in the decade and a half after my papers in Rockoff (1991).

Arthur Rolnick and Warren Weber were among the first to collect additional evidence and present a view of free banking and its troubles that was more sanguine than mine. In important papers in the Federal Reserve Bank of Minneapolis

Quarterly Review and the *American Economic Review,* Rolnick and Weber (1982, 1983) examined data for four free banking states – New York, Indiana, Wisconsin, and Minnesota – and showed that while those free banking systems "had a significant number of problem banks," they also had a large number of successful banks. They argued that it was misleading "to characterize the overall free banking experience as a failure of *laisse-faire* banking" (Rolnick and Weber 1983, 1090). In a subsequent paper in the *Journal of Monetary Economics* (1984), they provided more evidence and argued that most of the problems faced by the free banking systems were due simply to shocks that reduced the value of the bonds that the free banks held, rather than any inherent instability produced by free entry. Kahn (1985), however, re-examined the methods employed by Rolnick and Weber and concluded that free banks, especially in their first years of operation, had suffered higher rates of failure than Rolnick and Weber had calculated.

Economopoulos's (1988, 1990) studies of Illinois, New York, and Wisconsin provided additional evidence supporting the claim that wildcat banking was rare. He found that few free banks entered when profit opportunities were exceptionally high. He did find, however, that in Illinois free bank losses were high, even if they were not due to wildcat banking. Hasan and Dwyer (1994) took another look at some of the free banking states and argued that sometimes bank runs that undermined free banking systems were ignited by shocks that originated outside the state where the banks were located. Dwyer (1996) summarized much of the earlier literature, added some new facts of his own, and reinforced the view that whatever the problems experienced by antebellum banks, the incidence of wildcat banking was rare.

A more skeptical note about the experience in the free banking states was sounded by Jaremski (2010). Using a large data set created by Rolnick and Weber and made available on the Internet (a model contribution to the discipline), Jaremski argued that it was the undiversified portfolios of the free banks that caused their problems rather than bond price declines, because banks in chartered-banking states did not fail in the same numbers as banks in free banking states as a result of the decline in bond prices. To be sure, regulation left many would-be bankers in free banking states with little chance of getting into the game except by purchasing bonds and issuing notes. But they could have chosen not to take such a risky course. Lehman Brothers was hurt by a fall in the value of subprime mortgages beyond its control, but they can still be castigated because they chose to invest a large part of their portfolio in assets ultimately backed by subprime mortgages.

Two states, New York and Michigan, represent the extremes under free banking. In New York the banking system flourished and New York City continued its march to its preeminent place in world finance; in Michigan, free banking got off to a disastrous start and proved to be the source of many of the famous stories, real and apocryphal, about wildcat banking. Although Kahn (1985) had emphasized the high rate of bank failure in some free banking states; he also drew attention to the success of free banking in New York. Haupert (1991, 1994) explored the New York system in detail and argued that one reason for its success was the efforts banks made to establish a reputation for soundness, for example, by holding large amounts of reserves. Bodenhorn and Haupert (1995, 1996) explored another aspect of the New York experience: the "note issue

paradox." After the Civil War, it appears, at least according to some straightforward calculations, that National Banks did not take full advantage of the profit opportunities made possible by their privilege of issuing bank notes. This paradox – that they left money on the table – was made famous by Friedman and Schwartz (1963) and explored by a number of economic historians. Bodenhorn and Haupert (1995, 1996) made similar calculations for the antebellum free banks in New York. They concluded that various legal restrictions meant that at the margin it paid for these banks to make new loans with deposits rather than notes.

Michigan was actually the first state to adopt free banking. It did so in 1837, months ahead of New York. Rockoff (1985) drew attention to the Michigan experience and presented some arguments about why the significance of the losses in Michigan had been exaggerated. Recently, Dove et al. (2014), on the basis of newly collected data, have gone further in minimizing the amount of wildcat banking in Michigan.

One conclusion that emerges clearly from the studies of free banking in the United States before the Civil War is that wildcat banking was at most a rare phenomenon. One indicator that contemporaries understood this is that the National Banking Act, passed during the Civil War, was simply another free banking law, although one that provided a higher level of security for noteholders. National Bank notes had to be backed by federal government bonds. Riskier state and municipal bonds were not eligible. It is usually said that the National Banking Act was modeled on the successful free banking law of New York, but it should be mentioned that Salmon Chase, the secretary of the Treasury who helped create the National Banking system, was the former governor of Ohio, another state with a successful free banking law.

Other studies of free banking laws (Rockoff 1975b; Ng 1988; Economopoulous and O'Neill 1995; Bodenhorn 2000; Chabot and Moul 2014) have focused on a potentially positive part of the case for free banking: whether freedom of entry improved the supply of banking services and possibly accelerated economic development. It is a difficult question to answer. Obviously, it is hard to imagine economic growth without an expansion of financial intermediation, just as it would be hard to imagine economic growth without an expansion of the transportation network. But showing that removing or adding legal restrictions on banking would affect the rate of growth of the financial sector, and then that better or increased intermediation would affect the rate of economic growth overall, is a tall order given the limited amount of data and the absence of clear natural experiments. Yes, some states had free banking laws while others had individually chartered banks, but as Economopoulous and O'Neill (1995) point out, the passage of free banking laws in some states led to liberalization of bank chartering rules in other states, so that comparisons of different states will fail to distinguish clearly the effects of changes in regimes. Moreover, it appears that there was a high degree of integration in financial markets during the antebellum period so that a dearth of financial intermediation in one region might be offset by an influx of finance from other areas. And the free banking laws required that notes be backed by government bonds, usually of the state where the bank was located. This meant that the seigniorage procured by issuing bank notes was funneled into the coffers of the state government rather into the local economy.

Despite these difficulties, cliometricians have tried to delineate the effect of various forms of banking on economic growth. Rousseau and Sylla (2005), although focused on an earlier period, should be mentioned here because of their unusually clear statement and test of the idea that a deep and efficient financial system is a prerequisite for rapid economic growth; growth in their view was "finance led." Rockoff (1975a) argued that the data suggested that liberal banking regimes provided better services and encouraged the use of bank money. Bodenhorn (2000), focusing on the later antebellum period, summarizes the literature and performs a battery of additional tests to explore the power of the finance-led growth model. He concludes that finance and banking clearly were required for economic growth; and more cautiously that the system bank regulation, while appearing unnecessarily fragmented, allowed individual states to find regulatory systems suited to their own economies. Jaremski and Rousseau (2013), on the other hand, concluded that free banks did less well in generating economic growth than chartered banks. In related work, Atack et al. (2014) described a "virtuous cycle" in that banks encouraged railroad building and railroads then encouraged banking.

The relationship between railroads and banking in the Middle West brings to mind Fogel's (1964) studies of the railroad, one of the foundations of cliometrics. To my mind, Fogel's studies provide a helpful analogy for thinking about the role of banks. The economic growth of the United States required some form of transportation: trails or roads or canals or railroads. But Fogel showed that the difference between the economy as it was with railroads and an economy as it would have been with only canals and roads improved with the money spent on railroads would have been positive but surprisingly small. The same may be true of banking. Some form of banking was necessary for economic growth but the difference between the systems in place and the next best alternative that would have developed in their absence might not have been transformative. I once asked Robert Fogel the following: "if building the railroads does not explain the rapid economic growth of the United States in the nineteenth century, what does? His answer was that it was a "full court press." Similarly, banks undoubtedly contributed to economic development, and banking laws affected how they did so, but we should not forget that many other forces were also pushing the USA forward.

Although free banking was the outstanding feature of antebellum banking as far as banking historians have been concerned, other systems have come in for scrutiny. The banking system of Louisiana, usually referred to as the "Forstall System" after its founder Edmund J. Forstall, garnered considerable praise from earlier generations of economic historians (Helderman 1931; Redlich 1947; Hammond 1957; Neu 1970). The system was put in place after the Panic of 1837. It prohibited the entry of new banks, required high specie reserves, and restricted lending to short-term paper as emphasized by the real bills doctrine. George Green's (1972) *Finance and economic development in the Old South; Louisiana banking, 1804–1861,* which is still the main source for information about the functioning of the Forstall system, provided a different view. Although much heralded for its legal restrictions, Green's evaluation is somewhat critical, arguing that the system retarded economic growth and that the structure of the system cannot be credited with assuring that it effectively surmounted the Panic of 1857, which was generally less severe in the South.

Even more attention has been paid to the Suffolk System of New England. Trade patterns in New England meant that notes issued by New England country banks tended to collect in Boston. Typically, the notes were purchased there at a discount, frustrating consumers and country bankers. The solution was to designate one bank, the Suffolk Bank, as the redemption center for New England notes. The Boston banks set this system up and funded it by purchasing stock in the Suffolk Bank. Country banks were required to maintain deposit balances at the Suffolk, but in exchange the country bank notes were purchased at par. The system seemed to work well, pleasing the public because New England notes always circulated at par through out the region. The country banks, however, chafed under the system because they thought the deposit requirements were too severe, and eventually set up a rival. However, it was in operation only a few years before the Civil War (Lake 1947; Mullineaux 1987). Calomiris and Kahn (1996) looked at dividend payout ratios and related variables and showed that the Suffolk System was not simply a means of exploiting country banks, but rather one that created benefits for the system as a whole. Rolnick et al. (1998) and Bodenhorn (2002) however, argued that the country banks did bare an excessive share of the costs of maintaining the payments system. Moreover, while the Suffolk System produced (perhaps unfairly in terms of the distribution of the costs) many benefits for bank customers in New England, Smith and Weber (1999) argued, primarily on the basis of a model that they constructed, that the claim that it solved all of the potential problems of a privately issued currency would be going much too far.

Although chartered commercial banks were the most important, there were a number of similar institutions that were active during the antebellum years. For one thing, as documented by Richard Sylla (1976), there were many private bankers. And many of them evaded laws that prohibited private banks issuing notes. The most famous of the private bankers was George Smith, who's Wisconsin Marine and Fire Insurance Company issued "certificates of deposit" and checks in denominations as low as one dollar that circulated from hand to hand as cash. Smith's remarkable career as a private frontier banker is described in Smith (1966). At its peak in 1852, Smith's various banks were issuing nearly $1.5 million in notes while another private banker was issuing almost $2 million. Together, these two alone added about 2% to the paper currency (Sylla 1976). There were also a variety of savings banks, trust companies, bill brokers, and investment banks. Indeed, there were financial institutions that would now be labelled shadow banks. Just as we think of the failure of Lehman Brothers as the trigger for the Panic of 2008, before the Civil War people thought of the failure of J.L. and S. Joseph as the trigger for the Panic of 1837 and the failure of the Ohio Life and Insurance Company as the trigger for the Panic of 1857. Today both would be counted as shadow banks (Rockoff 2018). With the exception of the Ohio Life, there has been little cliometric work on the private bankers, in part because records their activities are not easily assembled, if they still exist at all.

Although most of the regulations that affected banks are to be found in charter provisions or the clauses of the free banking laws, there were other parts of the legal environment that influenced banking. One important set of legal restrictions was to be found in the usury laws. Rockoff (2009) was impressed by the relatively high

maximum rates of interest and low penalties for violating them in the West, and argued that this pattern reflected interregional competition for capital. An often-expressed view of economists is that because there is generally no active enforcement of usury laws – typically they were invoked only as a defense when a borrower was sued for nonpayment – they were honored in the breach. Bodenhorn (2007) found some evidence for this in New York State. Benmelech and Moskowitz (2010), however, found that the usury laws did influence economic activity. And they argued that the usury laws, although detrimental to economic activity in general, often favored the interests of the elite who supported them.

The Gold Inflation of the 1850s

The true sources of the Jacksonian Inflation were hard to see and many contemporary observers and latter day historians blamed the banks. It took the cliometricians to show that the big story was an increase in the monetary base produced by foreign flows of specie. The inflation of the 1850s was another story. It was obvious that the discovery of gold in California and subsequently in Australia produced an increase in the monetary base and an increase in the stock of money that in turn produced inflation. The banks expanded their loans, notes, and deposits, but that was because bank reserves were rising. Some relevant data is shown in Table 3. As in Table 2, all of the time series have been converted into indexes to make the table easier to read by dividing each observation by the value of the series in 1849 and multiplying by 100. These series are not as reliable as modern series for the same variables; but they are probably sufficient for some broad conclusions. The first column shows that the stock of money in the United States rose rapidly in the 1850s; but the second column shows that the money multiplier did not. In other words, as in the 1830s, banks were not creating more and more dollars of deposits and notes relative to the amount of coins; the increase in the stock of money was driven by the increase in the monetary base. The third column shows that, as in the 1830s, the inflation was mild by today's standard. Prices rose a bit more than 2.5% per year between 1849 and their peak in 1857. Some asset prices, the price of land and in the South the prices of slaves, were rising more rapidly. Indeed, slave prices rose about 6% per year between 1849 and 1857 (Carter et al. 2006, series Bb210). Real GDP, moreover, was rising rapidly, at 5% per year over the same period.

In some areas of economic history the use of quantitative methods and formal economic theory began with the "New Economic History" in the 1960s. But in monetary history the use of these methods is much older. One of the first quantitative studies of the effects of the gold discoveries was by the outstanding nineteenth century economist William Stanley Jevons (1863). Various aspects of the gold discoveries were explored by cliometricians during the early years of the "New Economic History." Martin (1973) examined the consequences of the gold discoveries for the institutional framework of the monetary system. Stevens (1971) reexamined the data on the stock of money. And Temin (1974) analyzed the effect of the gold discoveries in the context of the relationships among gold flows,

Table 3 Money, prices, and output in the 1850s

Year	Money	The ratio of money to coin (the Money Multiplier)	Prices (the GDP deflator)	Real GDP
1849	100	100	100	100
1850	114	94	102	104
1851	129	97	100	112
1852	143	93	101	122
1853	160	102	100	135
1854	161	99	109	141
1855	169	104	112	142
1856	182	108	110	149
1857	151	89	113	150
1858	173	106	106	154
1859	179	116	107	162

Sources: Carter et al. (2006): series Cj7, Cj8, Ca13, Ca9

international capital flows, and macroeconomic trends. The episode appears to be a good natural experiment revealing the relationship between money and economic activity. For that reason, one might have expected even more research on this episode. Perhaps the clarity of the story has discouraged research by economic historians hoping to score a hit by overturning the conventional wisdom. One issue that might be fruitfully explored is Jevons's contention that prices in more competitive markets responded more quickly to the increase in the stock of money than prices in monopolistic markets because information was disseminated more rapidly in competitive markets.

Conclusion

The antebellum years have proved a fertile source of questions and data for cliometricians. Indeed, with the exception of the Great Depression, it may well be that there is no other period that has more to offer the financial cliometrician than the three decades before the Civil War. There were increases in the monetary base that produced substantial increases in the stock of money with important macroeconomic effects: an influx of silver in the 1830s and the discoveries of gold in the 1850s. There were banking panics in 1837, 1839, 1854, and 1857, which in some ways bear a striking resemblance to the Panic of 2008. And there was a wide array of experiments with ways of chartering and regulating banks, including the famous experiments with "free banking." Cliometricians have made considerable progress in all of these areas by carefully amassing empirical data and employing increasingly sophisticated theoretical models and econometric methods to analyze the data. They have analyzed the impact on prices and production of the surges in the stock of money and the role of the banking system in transmitting those surges. They have analyzed the causes of the banking crises. And they have delineated the

circumstances in which free banking produced something that might be called wildcat banking. What of the future? Has all the silver and gold been mined? We can't be sure, but my guess is that cliometricians will continue to find the antebellum era an important source of information and ideas about how banking systems work.

Cross-References

- ▶ Central Banking
- ▶ Origins of the U.S. Financial System
- ▶ Payment Systems
- ▶ The Antebellum US Economy
- ▶ The Cliometric Study of Financial Panics and Crashes

References

Alhadeff D (1962) A reconsideration of restrictions on bank entry. Q J Econ 76(2):246–263

Atack J, Jaremski M, Rousseau P (2014) American banking and the transportation revolution before the civil war. J Econ Hist 74(4):943–986

Benmelech E, Moskowitz T (2010) The political economy of financial regulation: evidence from U.S. state usury laws in the 19th century. J Financ 65(3):1029–1073

Bodenhorn H (1992) Capital mobility and financial integration in antebellum America. J Econ Hist 52(3):585–610

Bodenhorn H (1998) Quis Custodiet Ipsos Custodes? East Econ J 24(1):7–24

Bodenhorn H (2000) A history of banking in antebellum America: financial markets and economic development in an era of nation-building. Cambridge University Press, New York

Bodenhorn H (2002) Making the little guy pay: payments-system networks, cross-subsidization, and the collapse of the Suffolk system. J Econ Hist 62(1):147–169

Bodenhorn H (2003) State banking in early America: a new economic history. Oxford University Press, Oxford/New York

Bodenhorn H (2006) Bank chartering and political corruption in antebellum New York: free banking as reform. In: Glaeser EL, Goldin C (eds) Corruption and reform: lessons from America's economic history, pp 231–257

Bodenhorn H (2007) Usury ceilings and Bank lending behavior: evidence from nineteenth century New York. Explor Econ Hist 44(2):179–202

Bodenhorn H (2008) Free banking and bank entry in nineteenth-century New York. Finan Hist Rev 15(2):175–201

Bodenhorn H, Haupert M (1995) Was there a note issue conundrum in the free banking era? J Money Credit Bank 27(3):702–712

Bodenhorn H, Haupert M (1996) The note issue paradox in the free banking era. J Econ Hist 56(3):687–693

Bodenhorn H, Rockoff H (1992) Regional interest rates in antebellum America. In: Goldin C (ed) Strategic factors in American economic history: a volume to honor Robert W. Fogel. University of Chicago Press, Chicago

Briones I, Rockoff H (2005) Do economists reach a conclusion on free-banking episodes. Econ J Watch 2(2):279–324

Calomiris C (1990) Is deposit insurance necessary? A historical perspective. J Econ Hist 50(2):283–295

Calomiris C, Haber H (2014) Fragile by design: the political origins of banking crises and scarce credit. Princeton University Press, Princeton
Calomiris C, Kahn C (1996) The efficiency of self-regulated payments systems: learnings from the Suffolk system. J Money Credit Bank 28(4):766–797
Calomiris C, Schweikart L (1991) The panic of 1857: origins, transmission, and containment. J Econ Hist 51(4):807–834
Carter S, Gartner S, Haines M, Olmstead A, Sutch R, Wright G (2006) Historical statistics of the United States, millennial edition. Cambridge University Press, Cambridge
Catterall R (1903) The second bank of the United States. University of Chicago Press, Chicago
Chabot B, Moul C (2014) Bank panics, government guarantees, and the long-run size of the financial sector: evidence from free-banking America. J Money Credit Bank 46(5):961–997
Dove J, Pecquet G, Thies C (2014) The Michigan free Bank experience: wild cat banking or interference with contract? Essays Econ Bus Hist:3247–3279
Dwyer G (1996) Wildcat banking, banking panics, and free banking in the United States. Fed Reserve Bank Atlanta Econ Rev 81(3):1–20
Economopoulos A (1988) Illinois free banking experience. J Money Credit Bank 20(2):249–264
Economopoulos A (1990) Free bank failures in New York and Wisconsin: a portfolio analysis. Explor Econ Hist 27(4):421–441
Economopoulous A, O'Neill H (1995) Bank entry during the antebellum period. J Money Credit Bank 27(4):1071–1085
Fenstermaker JV (1965) The development of American commercial banking, 1782–1837. Kent State University, Kent
Fogel R (1964) Railroads and American economic growth: essays in econometric history. Johns Hopkins Press, Baltimore
Friedman M, Schwartz A (1963) A monetary history of the United States, 1867–1960. Princeton University Press, Princeton
Gorton G (1996) Reputation formation in early bank note markets. J Polit Econ 104(2):346–397
Green G (1972) Finance and economic development in the old south: Louisiana banking, 1804–1861. Stanford University Press, Stanford
Greenfield R, Rockoff H (1996) Yellowbacks out West and greenbacks Back east: social-choice dimensions of monetary reform. South Econ J 62(4):902–915
Hammond B (1948) Banking in the early west: monopoly, prohibition, and laissez faire. J Econ Hist 8(1):1–25
Hammond B (1957) Banks and politics in America from the revolution to the civil war. Princeton University Press, Princeton
Hasan I, Dwyer G (1994) Bank runs in the free banking period. J Money Credit Bank 26(2):271–288
Haupert M (1991) Investment in name brand capital: evidence from the free banking era. Am Econ 35(2):73–80
Haupert M (1994) New York free banks and the role of reputations. Am Econ 38(2):66–77
Helderman L (1931) National and state banks: a study of their origins. Houghton Mifflin, Boston
Huston J (1987) The panic of 1857 and the coming of the civil war. Louisiana State University Press, Baton Rouge
James J, Weiman D (2010) From drafts to checks: the evolution of correspondent banking networks and the formation of the modern U.S. payments system, 1850–1914. J Money Credit Bank 42(2/3):237–265
Jaremski M (2010) Free Bank failures: risky bonds versus undiversified portfolios. J Money Credit Bank 42(8):1565–1587
Jaremski M (2011) Bank-specific default risk in the pricing of bank note discounts. J Econ Hist 71(4):950–975
Jaremski M, Rousseau P (2013) Banks, free banks, and U.S. economic growth. Econ Inq 51(2):1603–1621
Jevons W (1863) A serious fall in the value of gold ascertained, and its social effects set forth. Edward Stanford, London
Kahn J (1985) Another look at free banking in the United States. Am Econ Rev 75(4):881–885

Kelly M, O'Gráda C (2000) Market contagion: evidence from the panics of 1854 and 1857. Am Econ Rev 90(5):1110–1124
Knodell J (2017) The second bank of the United States: "central" banker in an era of nation-building, 1816–1836. Routledge, New York
Lake W (1947) The end of the Suffolk system. J Econ Hist 7(2):183–207
Lamoreaux N (1994) Insider lending: banks, personal connections, and economic development in industrial New England. Cambridge University Press, Cambridge
Macesich G (1960) Sources of monetary disturbances in the United States, 1834–1845. J Econ Hist 20(3):407–434
Martin D (1973) 1853: the end of bimetallism in the United States. J Econ Hist 33(4):825–844
Meltzer A (1967) Major issues in the regulation of financial institutions. J Polit Econ 75(4):482–501
Mihm S (2007) A nation of counterfeiters: capitalists, con men, and the making of the United States. Harvard University Press, Cambridge, MA
Mullineaux D (1987) Competitive monies and the Suffolk Bank system: a contractual perspective. South Econ J 53(4):884–898
Neu ID (1970) Edmond Jean Forstall and Louisiana banking. Explor Econ Hist 7(4):383–398
Ng K (1988) Free banking laws and barriers to entry in banking, 1838–1860. J Econ Hist 48(4):877–889
O'Gráda C, White E (2003) The panics of 1854 and 1857: a view from the emigrant industrial savings Bank. J Econ Hist 63(1):213–240
Redlich F (1947) The molding of American banking: men and ideas. Hafner Publishing Company, New York
Rockoff H (1971) Money, prices and banks in the Jacksonian era. In: Fogel RW, Engerman S (eds) Re-interpretation of American economic history. Harper & Row, New York, pp 448–458
Rockoff H (1974) The free banking era: a reexamination. J Money Credit Bank 6(2):141–167
Rockoff H (1975a) Varieties of banking and regional economic development in the United States, 1840–1860. J Econ Hist 35(1):160–181
Rockoff H (1975b) The free banking era: a re-examination. Arno Press, New York
Rockoff H (1985) New evidence on free banking in the United States. Am Econ Rev 75(4):886–889
Rockoff H (1991) Lessons from the American experience with free banking. In: Capie F, Wood G (eds) Unregulated banking: Chaos or order? Macmillan Academic and Professional Ltd, London
Rockoff H (2009) Prodigals and projectors: an economic history of usury Laws in the United States from colonial times to 1900. In: Eltis D, Lewis FD, Sokoloff KL (eds) Human capital and institutions: a long-run view. Cambridge University Press, Cambridge, pp 285–323
Rockoff H (2018) It is always the shadow banks: the regulatory status of the banks that failed and ignited America's greatest financial panics. In: Rockoff H, Suto I (eds) Coping with financial crises: some lessons from economic history. Springer Nature, Singapore, pp 77–106
Rolnick A, Weber W (1982) Free banking, wildcat banking, and shinplasters. Fed Reserve Bank Minneap Q Rev 6:10–19
Rolnick A, Weber W (1983) New evidence on the free banking era. Am Econ Rev 73(5):1080–1091
Rolnick A, Weber W (1984) The causes of free bank failures: a detailed examination. J Monet Econ 14(3):267–291
Rolnick AJ, Smith BD, Weber WE (1998) Lessons from a laissez-faire payments system: the Suffolk banking system (1825–58). Fed Reserve Bank Minneap Q Rev 22(3):11–21
Rousseau P (2002) Jacksonian monetary policy, specie flows, and the panic of 1837. J Econ Hist 62(2):457–488
Rousseau P, Sylla R (2005) Emerging financial markets and early US growth. Explor Econ Hist 42(1):1–26
Schwartz AJ (1947) The beginning of competitive banking in Philadelphia, 1782–1809. J Polit Econ 55(5):417–431
Selgin G, White L (1994) How would the invisible hand handle money? J Econ Lit 32(4):1718–1749

Shambaugh J (2006) An experiment with multiple currencies: the American monetary system from 1838–60. Explor Econ Hist 43(4):609–645
Smith A (1966) George Smith's money: a Scottish investor in America. State Historical Society of Wisconsin, Madison
Smith B, Weber W (1999) Private money creation and the Suffolk banking system. J Money Credit Bank 31(3):624–659
Stevens E (1971) Composition of the money stock prior to the civil war. J Money Credit Bank 3(1):84–101
Sylla R (1976) Forgotten men of money: private bankers in early U.S. history. J Econ Hist 36(1):173–188
Sylla R, Legler J, Wallis J (1987) Banks and state public finance in the new republic: the United States, 1790–1860. J Econ Hist 47(2):391–403
Temin P (1968) The economic consequences of the bank war. J Polit Econ 76(2):257–274
Temin P (1969) The Jacksonian economy. W. W. Norton & Company, New York
Temin P (1974) The Anglo-American business cycle, 1820–60. Econ Hist Rev 27(2):207–221
Tobin J (1967) Major issues in the regulation of financial institutions: comment. J Polit Econ 75(4):508–509
Wallis J (2001) What caused the crisis of 1839? NBER working paper series on historical factors in long run growth. NBER historical working paper No. 133, Cambridge, MA
Weiman D, James J (2007) The political economy of the US monetary union: the civil war era as a watershed. Am Econ Rev 97(2):271–275

Financial Markets and Cliometrics

Larry Neal

Contents

Abstract

The study of financial markets is a growing part of cliometrics for at least three reasons. First, appreciation of the role financial markets played in the rise and spread of capitalism has grown, along with concerns about financial crises. Second, accessibility to the immense amount of data generated by financial markets keeps improving thanks to continued advances in digital communications technology. Third, analytical techniques for determining the behavioral patterns of time series have advanced. While typically only price data for financial assets are available without the corresponding volume of the assets being traded, the consequences of sharp, or sustained, changes in the price of financial assets can be detected in other economic data. Interesting insights on fundamental historical issues are also possible by applying economic and political theory to cliometric studies of financial markets.

Keywords

Bills of exchange · Sovereign bonds · Credible commitment · Threshold auto regression · Cointegrated time series · Adverse selection · Asymmetric information · Financial crises

L. Neal (✉)
Department of Economics, University of Illinois at Urbana-Champaign, Urbana, IL, USA
e-mail: lneal@illinois.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_14

Introduction

The repeated occurrence of financial crises, especially the unexpected length of recovery from the global crisis that began in 2007, continues to generate interest in historical studies of financial markets. Each crisis seems to elicit the reaction of what went wrong this time? Then, why didn't we learn the right lesson from the last one? Trying to extract lessons from the history of past crises drives financial historians (as well as policymakers, speculators, and journalists) in their research on financial markets. Beyond the narrow concerns raised by financial crises, however, financial history can also shed new light on fundamental historical issues. Examples include how long-distance trade was sustained among ancient and medieval societies, how fiscal states arose in early modern times, and, ultimately, how societies move from economic relationships underlying personal exchanges to institutions that allow impersonal exchanges to be sustained. Once scholars recognized the importance of finance for enabling these important transitions in human history to occur, the opportunities for meaningful research into financial markets by cliometricians kept expanding. Further, data generated by financial markets in the past can serve as useful measures of the success or failure of previous economic efforts, provided, of course, that they are interpreted correctly by modern cliometricians. To illustrate just a few of the possibilities for getting illuminating insights as well as for making mistaken inferences, this essay surveys two different literatures that have arisen over the past half-century, first on financial markets for sovereign government bonds and then on financial markets for bills of exchange. Bringing the two strands of analysis together for a better appreciation of the interplay between short-term finance and long-term assets is the next step in a research agenda that keeps expanding.

Sovereign Government Bonds

An extensive and growing literature has arisen from the realization that financial markets for sovereign government debts can be analyzed from a variety of perspectives and that governments issuing these debts kept records that are increasingly accessible to modern researchers equipped with digital cameras and laptop computers. The classic study by P. G. M. Dickson (1967) introduced the term "financial revolution" to the profession and also provided a useful finder guide to the wealth of material readily available in British archives. That material, combined with the daily price data on British funds from January 1698 on available in John Castaing's *The Course of the Exchange, & c.* (1698–1907), enabled Neal (1990) to demonstrate weak form efficiency[1] of the securities market for sovereign bonds issued by the British government in London. Combining these data with pricing of British funds in

[1]Weak form efficiency of efficient financial markets: all past prices of a stock are reflected in today's stock price, which typically follows a random walk.

Amsterdam, Neal also showed that these two preeminent financial markets were closely integrated, especially after the bubble year of 1720. Later work by Koudijs (2011) expanded these data to determine more precisely whether "news" affecting the securities widely traded in both Amsterdam and London arrived first in London or Amsterdam, depending on the arrival of the mail packet boats that sailed regularly between the two cities. The combined results from the London and Amsterdam markets suggest semi-strong efficiency[2] for these early stock markets, with news affecting the prices of English government securities typically reaching Amsterdam first. Beach et al. (2013) further examined the Amsterdam prices of British securities to argue that they were spot, not time, prices as Neal had inferred in his original work.

Beyond such technical issues concerning the efficiency and integration of the eighteenth-century financial markets through analysis of the prices of widely held and traded securities, the enthusiasm of Dickson for finding an early "financial revolution" corresponding to the Glorious Revolution of 1688/1689 in England became the basis for new ideas for economic policy generally. North and Weingast (1989) took Dickson's finding of a sharp, sudden fall in the interest rates offered on new debt issues after 1688 as strong evidence that the constitutional arrangements between the new monarchs of Great Britain, William III and Mary, and the Parliament had created a "credible commitment" that the British government would no longer interfere with private property rights. This constitutional arrangement, according to North and Weingast, laid the basis for the eventual industrial revolution in England and the initiation of the current era of modern economic growth. The appeal of this argument has spawned a growing literature on its own, both pro and con.[3] Assessing the price evidence from private banking accounts before and after the Glorious Revolution, Quinn (2001) found that interest rates on short-term bankers' loans actually rose after 1688. Sussman and Yafeh (2006) argued that the bulk of government debt issued to finance the two subsequent wars over the next 25 years had to pay higher interest rates than during peacetime. This, they argued, showed the importance of war finance over constitutional commitments, an argument they extended to later periods and other cases (Mauro et al. 2006). Wells and Wills (2000) tested for robustness of the later fall in long-term yields and found that the "credible commitment" of William III and the Whigs in 1688 was subject to severe shocks for at least 50 years after 1688 due to the persistence of the Jacobite threat to restore the Stuart dynasty.

Because these analyses that cast doubt on the North and Weingast interpretation of Dickson's findings relied simply on price data, the question does arise whether the quantity data, which were of most interest to Dickson, might give different results. Sussman and Yafeh disputed whether interest rates fell for British sovereign debt after 1688 as the demands of war finance forced the government to sell fresh issues of both short-term and long-term debt at increasing discounts, forcing up the actual

[2]Semi-strong efficiency: all public information available is incorporated into a stock's price.

[3]See (Coffman and Neal 2013) for an extended analysis and review.

market yield above the nominal interest rate. Nevertheless, they could not overturn Dickson's evidence on the huge sustained increase in the volume of sovereign debt issued and the eventual rise in its market price as the government continued to pay the promised interest. Even Quinn in his examination of private finance before and after the 1688 revolution found that the size of banking business expanded sharply and permanently. The increased volume of sovereign debt that continued to be serviced by whichever party was in power laid the basis for a remarkable expansion of banking business in London and later throughout the kingdom.

MacDonald (2013) argued in fact that it was the 1710 election of Tory party to power in Parliament that confirmed the commitment of the Stuart monarch (now Queen Anne) and Parliament to continue service of the outstanding debt, both short- and long-term. Stasavage (2003) used the price data on sovereign debt for Britain to show that interest rates fell when the Whig party was in power and rose whenever the Tory party replaced it. Moreover, yields on British sovereign debt fell after each war without defaults, unlike the case for French sovereign debt (Luckett and Lachaier 1996; Velde and Weir 1992).

Using game-theoretic constructs to find useful political variables in addition to the standard "fundamentals" used by economists as determinants of yields on sovereign bonds, Stasavage searched for evidence on bond yields from other political entities in Europe before the financial revolution in England. Stasavage (2011) concluded that in early modern Europe smaller cities, governed by more cohesive merchant elites, generally paid less interest on their sovereign debts. This helped explain Epstein's earlier finding (Epstein 2000), that Italian city-states paid much lower rates on their public debts for centuries before the constitutional commitment in England that had fascinated first Dickson and then North and Weingast. Epstein argued that the Italian success was due to solving coordination problems over a larger range of market activities, exemplified by the success of Milan in recovering from the effects of the Black Death.

City-states that maintained their own mints, tax systems, and financial records proliferated in Western Europe from the eleventh century on and their records become increasingly available after 1400. The Italian city-states of Venice, Florence, and Genoa in particular kept detailed records that have been the subject of studies by quantitative historians, economists, and sociologists. Luciano Pezzolo (2003, 2013, 2014) has compared the market interest rates paid by those three leading Italian city-states with those paid by the papacy in Rome in an effort to determine which political structure conveyed the most confidence for its creditors through the vicissitudes of state-building to 1700. Republics did best, until they fell under the rule of a prince (Florence) or of a closed oligarchy (Venice). David Stasavage expands the sample of sovereign city-states beyond Italy to include others in Spain, Germany, and the Low Countries (Stasavage 2011) to find that smaller city-states with more cohesive merchant groups did best of all.

Tomz (2007) enhanced the game theory underlying government commitment mechanisms for servicing their debt by adding the possibility of learning and political change to standard models for building and sustaining reputations. These modifications to cooperative game theory allowed him to predict uncertainty

premiums on issues by new governments, seasoning effects on prices of bonds that continue to be serviced by established governments, exclusion from existing markets for defaulters, and reentry of governments when compensation is offered to previous lenders, all with specific historical examples. His qualitative search for evidence of the factors that determine a government's reputation at any given time helped him explain apparent anomalies in the historical pricing of various sovereign bonds. Exploring the sources of reputation building opens up further avenues of research for cliometricians.

One of the most interesting episodes for testing various economic and political theories for explaining the attractiveness of sovereign bonds is the first so-called Latin American debt crisis, which occurred in London from 1822 to 1830. The new Latin American states that emerged from the collapse of the imperial authority exercised from Spain and Portugal at the end of the Napoleonic Wars all attempted to finance their new governments by issuing bonds on the Paris, Amsterdam, and London markets. All of them offered 6% interest on their bonds, and London investors willingly bought them at discounts up to 20% to get yields between 7% and 7.5% – until they learned more about the inability of the new governments to cover their current expenses with taxes, much less pay interest due on their outstanding debts. This was a classic early example of the "lemons" problem being solved by lenders assigning an arbitrary risk premium to the loans sought by new, untried borrowers. When news arrived of the shortfalls suffered by the various governments, the prices plummeted, implying sharp rises in yields as shown in Fig. 1.

Tomz (Chap. 2) uses these data to illustrate both the lemons problem (solving adverse selection with risk premium from 1822 to late 1825) and seasoning (maintaining prices for French bonds as well as a new price level for Brazil bonds

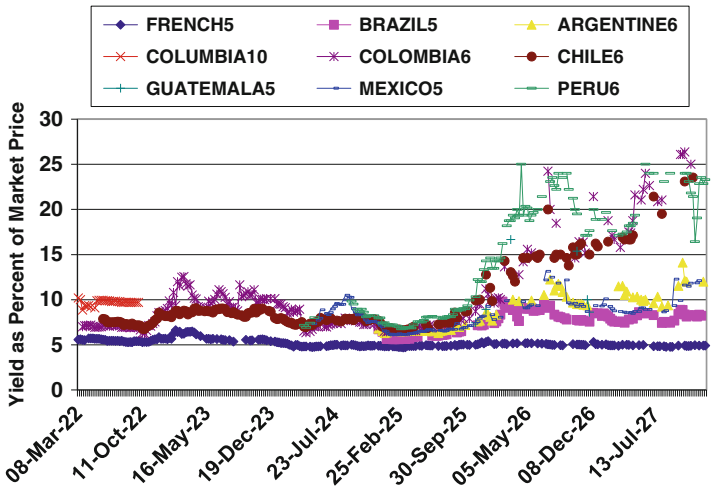


Fig. 1 French and Latin American Bond Yields in London, 1822–1827

while varying risk premiums for other Latin American bonds). Later work by Flandreau and Flores (2009) examined the role of the respective investment banks that took the lead in marketing each of the bonds to find out why yields stabilized for the Brazil and Argentine bonds. The answer, they found, was in the way Rothschilds (who handled the Brazil bonds) and Barings (who dealt with Argentina's bonds) imposed conditions upon those governments before lending their reputations to the issue. Indeed, Rothschilds issued bonds for a variety of new European governments – Austria, Belgium, Naples, Prussia, and Russia – in the decades following 1815 without any defaults, even during the revolutions of 1848. One may speculate on how the Rothschilds accomplished this, perhaps through monitoring the respective country mints and public banks for each country and imposing effective conditionality. Whatever was the secret to the success of the bonds underwritten by the Rothschilds, Flandreau and Flores argue that their effective "branding" of a country's debt enabled each government to save on future interest payments and cover the underwriting premium charged by the House of Rothschild.

Led by the merchant bankers Rothschilds and Barings, London and Paris became the centers for international sovereign bonds for the remainder of the nineteenth century. The faithfully recorded prices of the numerous government bonds have gradually been encoded and analyzed by an increasing number of cliometricians, for a wide variety of purposes. One study focused on the case of Peru, whose bonds stabilized marvelously in mid-century despite being one of the "lemons" in the 1825 crisis and despite enduring a series of unstable governments. Vizcarra (2009) explained this was due to the role of the British merchant banker Gibbs and Sons, who also managed the sale of guano in London and made sure that their clients who had purchased the bonds were given first claim on the guano revenue. A similar arrangement turned out to be the case for the kingdom of Denmark when it borrowed from Dutch merchant bankers at the end of the eighteenth century. In that case, the Dutch banking houses collected the tolls in advance from shippers leaving Amsterdam for Baltic ports. In that way they could assure these payments were applied to the interest due on the Danish bonds (Van Bochove 2014). An even earlier example of using tax collection authority to maintain reliable payment of interest on long-term sovereign debt was the bonds issued by the city of Paris to provide their occasional tributes to help finance the wars of François Ier (Vam Malle Sabouret 2008).

Over the course of the nineteenth century, however, states gradually took control of their own revenues to free themselves from external domination. Dincecco (2009) analyzes the bond yields for 11 European governments for varying periods from 1750 to 1913. He then tests for the relative importance for creditworthiness of each government of (1) centralized control over tax revenues versus (2) limitations on executive power. He considers these to be the two essential elements of the commitment mechanisms used so adroitly by the British throughout the eighteenth century to create a permanent market for their sovereign bonds. He finds that both effects are important individually, but most effective when they are combined, as was the case for Britain after 1688. Ending with this anodyne conclusion, Dincecco managed to avoid confronting directly the contentious issues raised by earlier scholars who had used the evidence of bond yields to validate their preconceptions.

Probably the most stimulating work was Bordo and Rockoff (1996), who argued it was the adoption of the gold standard as "a Good Housekeeping seal of approval" that allowed countries to increase their creditworthiness. This would be further confirmation of earlier work by Bordo and Kydland (1995) that the gold standard as such was a powerful commitment device to restrain governments from excessive, much less unwarranted, issues of debt or money. Ferguson and Schularick (2006), however, took evidence on bond yields from a larger group of countries during the classical gold standard period, 1880–1913, to argue that it was the rule of law, specifically British law, that allowed countries under the sway of Britain to assure creditors, regardless of their formal commitment to a gold standard. Then, Accominotti et al. (2011) showed that it made a lot of difference within the British Empire whether the colony was settled by British emigrants or simply ruled by British civil servants. The British guaranteed payment on bonds issued by the settlement colonies but not on bonds issued by non-settlement colonies with local rulers, which created yield spreads favoring the white settlement colonial government bonds. Similar effects of third party guarantees were also found for debt issued by the Ottoman Empire during the Crimean War, which enjoyed low rates of interest when jointly guaranteed by the British and French governments.

Later issues of Ottoman debt without such guarantees, however, suffered badly until ad hoc international financial commissions finally took control of the Ottoman revenues dedicated to service of the bonds in the 1890s (Tuncer 2011; Pamuk and Karaman 2010). Formal sanctions against defaulting governments, enforced by the so-called gunboat diplomacy during the nineteenth century, seem not to have been very effective and were seldom used. Private initiatives by European stock exchanges to refuse formal listing of new bonds by previous defaulters were coordinated by non-governmental institutions such as the Council of Foreign Bondholders (Esteves 2013). The Roosevelt Corollary of 1904 (to the Monroe Doctrine of 1823 that defended decolonization in the Western Hemisphere) reinforced the reluctance of the British government to undertake military measures against defaulters, especially in Latin America. Nevertheless, the Roosevelt Corollary had a noticeably positive effect on the markets for Latin American government bonds given the willingness of that American administration to use force (Mitchener and Weidenmier 2005, 2010).

Thanks to the ongoing revolution in information and communications technology, there is an overwhelming quantity of historical data on financial markets available for continued research and controversy. For example, Global Finance Data (https://www.globalfinancialdata.com/index.html) is a for-profit provider of the data sets created by various academics (including Neal (1996), at http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/1008) as well as by governments and other commercial firms. Individuals can subscribe for a limited free trial, or use their academic affiliation to access many of the 20,806 data series available as of February 14, 2014. The provenance of those data, however, has to be taken on faith, whereas for sovereign bonds, the European State Finance Database (http://www.esfdb.org) provides the academic sources for each of the data sets available there. These data were originally collected under the auspices of a project on "The Origins of the

Modern State in Europe, 13th to 18th centuries," directed by the Rev. Professor Richard Bonney with the assistance of Dr. Margaret Bonney from 1989 to 1992. It is now maintained by D'Maris Coffman at the Centre for Financial History housed in Newnham College, Cambridge University, and new data sets are added regularly from various academic studies. As most of these data are in nominal prices, investigators wishing to make current comparisons can access the database at Measuring Worth (http://www.measuringworth.com), which has price conversions as well as additional long-run data. Many scholars are making their data available as well at EH.net Data bases (http://eh.net/databases/), which is constantly adding new series underlying published, and sometimes unpublished, work.

Short-Term Commercial Finance

The regular publication of discount rates on commercial bills of exchange for the major cities in Europe and the Americas starting in the nineteenth century allows this kind of statistical analysis to complement analysis of movements in yields of government bonds, especially during financial crises or wars or changes in political regimes. For earlier periods, some cliometricians have taken advantage of the regular publication of exchange rates among major financial centers to extract the implicit interest rate from the difference between spot and forward rates. The pioneering study by Eagly and Smith (1976) focused on just the London quotes for bills on Amsterdam. Nevertheless, they were able to show a high level of financial integration between the two dominant money markets of eighteenth-century Europe. Further, if the gap between the time price of foreign exchange and the spot price widens, the intensity of the crisis can be measured as well. The first mark of a scramble for liquidity in London, which was the sudden, but short-lived, spike in the price of the pound sterling, I labeled "the Ashton effect" (Neal 1990, p. 67). Sometimes a reaction followed quickly to produce an offsetting spike in the price of *schellingen banco* as merchants in Amsterdam scrambled for liquidity in response to the difficulties in London. This movement I termed "the Kindleberger effect," as it was a clear marker of contagion in Kindleberger's view (Kindleberger 2000), but interdependence as Forbes and Rigobon (2002) would see it. Schubert (1989) demonstrated the initial integration of the exchange markets in the eighteenth century and their increasing disruption from the Seven Years' War on as both Ashton and Kindleberger effects increased in magnitude and frequency. Later work by Quinn (1996) highlighted the disruption caused by the pressures of war finance on the Amsterdam-London markets for bills of exchange after the currency reform in England in 1696. While restoring the previous value of the pound sterling in terms of gold, the reform also set the pound at a value in silver that made gold more valuable than silver in England relative to the Netherlands or France. Through the use of bills of exchange, Quinn showed how exports of silver from London to Amsterdam were often covered by imports of gold from Amsterdam to London, both financed by issuing bills of exchange.

More extensive work on these markets for commercial finance by Flandreau et al. (2009b) covered a wider range of exchange rate markets but Flandreau et al. (2009a) focused on the comparative interest rates for merchants in the three major mercantilist countries. Their findings also showed the effects of wars and occasional financial crises on private interest rates, but while London rates became lower than Amsterdam rates, which were also lower than Paris rates generally, none of the three were constrained by the usury laws that limited rates to 5% annually. They did find generally rising rates in the last quarter of the eighteenth century for all three cities, and more variance among them as did Schubert.

The intensive study of the European market for Mexican silver in the seventeenth and eighteenth centuries by Nogués-Marco (2013) reflects the incentives both for encoding more financial data from previously underexploited resources and for extracting more analytic insights from applying more sophisticated statistical techniques. While Nogués-Marco only confirmed that Great Britain was on a de facto gold standard even while maintaining it had a de jure bimetallic standard throughout the eighteenth century, she also managed to demonstrate why the Netherlands could be so closely connected financially with Britain while maintaining both a de facto and de jure bimetallic standard at a different ratio between gold and silver (14.65 for Amsterdam and 15.21 for London). She was building in part on the path-breaking work done by her thesis advisor, Marc Flandreau, on the sustainability of the bimetallic system in the nineteenth century (Flandreau 1996, 2004), as well as implementing the theoretical analysis of Velde and Weber (2000). It was Flandreau's intensive empirical work on the public and private actions in France over the tumultuous period of 1840–1878 that demonstrated how the stock of both monetary metals could be maintained at sufficient levels to warrant continuation of the bimetallic standard despite the huge increases in gold supplies entering world markets after 1849.

The theoretical work of Velde and Weber demonstrated that bimetallism could have been maintained indeed well into the twentieth century. The elimination of bimetallism in 1873 was probably due to the French decision not to subsidize Germany's plans to convert its diverse silver standard areas into a unified gold standard for the empire by buying up the excess silver released from German mints while decreasing French gold supplies. While general deflation followed globally, with much of the pain suffered by a defeated and diminished France, Velde and Weber's theoretical analysis suggested that both France and Germany were better off by adopting a single metal standard, whether it would have been gold or silver. US legislation in 1873 also made the fastest growing economy in the world committed to a gold standard thereafter, replacing its original de jure bimetallic and de facto silver standard before the Civil War and its fiat money inflation during the War Between the States.

The gold standard then became the dominant monetary standard for the world economy after 1873, the combined result of US legislation and French policy action. Lawrence Officer (1996) set the standard for empirical work on the operation of private bankers dealing with the exchange of dollars and sterling over the period 1791–1931. Officer's extensive studies were stimulated in turn by the seminal work

by Davis and Hughes (1960), two of the co-founders of the cliometrics meetings while they were both assistant professors at Purdue University. Most recently, Canjels et al. (2004) took the cliometric study of exchange rates yet another step forward, by increasing the quantity of data to obtain higher-frequency exchange rate quotes from the printed sources of the late nineteenth century and then applying more sophisticated econometrics, threshold autoregressive (TAR) time series analysis, than was available for earlier researchers.

TAR was developed to determine the price bands within which prices of a given good could vary without affecting the prices for the good in an adjacent or distant market. Used successfully to determine the degree of market integration for commodities over space and time, they could also be used to examine what were the effective gold points for arbitrage between London and New York when both countries were committed to a gold standard. Canjels et al. further extended their analysis to see if actual gold movements did occur when exchange rates hit or exceeded their estimated gold points, and found enough confirmations to reassure them. But the most reassuring aspect of their findings was that their estimates of the gold points were very much the same as those determined by the more tedious efforts of Officer (1996) and Flandreau (2004), namely, to find historical evidence of the actual costs of shipping, insurance, and interest payments incurred by operators in the foreign exchange markets of the time.

The demonstrated usefulness of TAR econometrics overall has stimulated other work on exchange rates in historical settings of financial markets for short-term commercial credit. Volckart and Wolf (2006), for example, use TAR to derive the implications for the extent of market integration and the speed of adjustment to changes in mint ratios for fourteenth- and fifteenth-century Flanders, Lübeck, and Prussia. They find that it took about 8 months for deviations between Flanders and Lübeck to fall back within bullion points but twice as long for adjustments to occur between Flanders and Prussia, showing the importance of seaborne trade for northern and central Europe. Following up on this study, Chilosi and Volckart (2011) apply TAR analysis to the 13,092 exchange rates that Volckart (1996) collected mainly from account books of guilds, merchants, ecclesiastical organizations, and city authorities in central Europe for the period 1400–1520. They used these data to determine which cities were integrated with each other and how integration changed over time. The results show that long-term trends toward improved financial integration dominated the cycles of debasements that occurred regularly, but also that integration seemed driven more by rising trade than by political unification. Work on the dominant role of Genoese bankers in the sixteenth and seventeenth centuries by Pezzolo and Tattara (2008) uses cointegration analysis of interest rates on rechange bills marketed in the Bisenzone fairs dominated by the Genoese. They find that the Genoese money market was directly affected by news from Spain about the war expenses or arrival of silver from America, but these shocks also affected the money markets in Florence and Milan, while Venice remained unaffected. The risks of dealing with short-term Spanish *asientos* by the Genoese therefore explain why short-term interest rates in Genoa were consistently higher than the yields on long-term Genoese government debt.

While Volckart's data are accessible on his website, Volckart (2014) the fourteenth- to sixteenth-century exchange rates, perhaps the most extensive data set is maintained at Rutgers University as *The Medieval and Early Modern Data Bank* (Bell and Howell 1998), http://www2.scc.rutgers.edu/memdb/. Researchers at the University of Reading (Bell et al. 2013a) have applied times series analysis to the exchange rates recorded there as well as higher-frequency rates recorded by the Tuscan merchant, Francesco Datini (downloadable from the Datini Archive 2014). They find evidence of seasonality, occasional trend breaks associated with debasements and military conflicts, and overall an inverted term structure of interest rates for early modern Europe. Long-term sovereign bonds appear to have been guaranteed against debasements in monetary regimes based on precious metals while commercial credit was subject to higher idiosyncratic risks for specific trades (Bell et al. 2013b).

Next Steps

As useful and insightful as these studies have proven to be so for, the next step – in addition to continuing to extract and encode ever more data from historical financial markets and continuing to apply ever more sophisticated statistical techniques to the data – should be to see how the markets for long-term sovereign debt interact with the markets for short-term commercial credit. Practitioners in finance have long acknowledged the importance of the existence of long-term government debt for facilitating the short-term finance of commercial activities but it took the work by Gelderblom and Jonker (2004) to document precisely how this occurred at the start of financial capitalism. Using the accounts of the Amsterdam merchant Hans Thijs in the period 1595–1611, they found that the interest rates he paid to investors in his various ventures dropped permanently once he invested in the permanent shares of the Dutch East India Company, founded in 1602. This was because he could now pledge the shares as collateral for loans from a much wider range of potential investors than before. Marketable financial assets backed by the commitment of revenue by the issuing government or corporation provide merchants with reusable collateral that can be posted repeatedly against short-term loans from any potential investor. This insight provides the logical link between the markets for sovereign bonds and short-term commercial credit, but one still has to determine how to test for the reciprocal effects between the two sets of financial markets.

One approach, taken by Neal and Weidenmier (2003), was to take financial crises as given and then to see whether contagion occurred during the subsequent crisis, all to indicate what kind of learning process might have been going on among policymakers in major industrial countries during the classical gold standard era. As with all the studies covered in this survey, this approach required the two steps of acquiring new data (high-frequency short-term interest rates) and applying new statistical techniques (adjusting standard deviations for heteroskedasticity). Given the interest in financial crises, which always start with a shock to the supply of short-term commercial credit somewhere in the payments system, and the possibility of

contagion to other parts of the financial system, whether domestic or international, short-term interest rates are equally interesting for cliometricians. Mishkin (1991) reviewed the financial crises in the USA from 1857 through 1987 to show that rises in short-term interest rates preceded each crisis. Increased spreads between the yields of lower rated commercial or corporate securities and government securities for both short- and long-term assets then accompanied each financial crisis.

Neal and Weidenmier (2003) found similar results for international financial crises dating from 1825 to 1907, although in their methodology contagion was limited to only the 1907 crisis. That corresponded to the similar rejection of the contagion thesis found by Forbes and Rigobon (2002) for the Asian financial crises in the late 1990s as well as for the Mexican crisis of 1994 and the US stock market crash of 1987. Their argument, which hasn't yet entered conventional wisdom, is that increased volatility of asset prices during a financial crisis increases standard measures of correlation across markets, which may or may not have been interdependent before the crisis. Adjusting for heteroskedasticity in standard deviations that accompanies financial crises allows one to determine if correlations among markets really did increase and therefore to differentiate between interdependence and true contagion. For Neal and Weidenmier, the anomaly of 1907 came when previous interdependence of the New York and London markets for short-term finance was disrupted by the decision of the Bank of England to prohibit dealing with American commercial paper earlier that year. Odell and Weidenmier (2004) traced this decision back to the gold outflows needed to cover the payouts by British insurance companies for the losses from the 1906 earthquake in San Francisco.

A complementary approach to combining analysis of markets for short-term and long-term financial products is taken by Schularick and Taylor (2012). They also take crises as given from a consensus of the profession, but they then extend the number of crises from 1870 through 2008. Further illustrating the theme of this essay, they both collect new data and apply new statistical techniques. Their new data are measures of bank credit (loans by financial institutions of all kinds and their total balance sheet assets) for 14 major countries, which they can compare with previously developed measures of money supply. Next, they apply new statistical techniques (and a new acronym, AUROC – area under receiver operating characteristic) to test which of their measures of financial markets, credit or money, do the best job of predicting financial crises, country-by-country and overall. The interesting finding is that both the credit and money measures perform equally well in predicting financial crises from 1870 up to World War II, but thereafter the credit measures become increasingly more powerful as predictors of crises. While the importance of credit booms for generating following crises even in the nineteenth century is not surprising for financial historians,[4] the failure of money measures to correlate closely with credit changes after 1948 is disappointing for economists trained in the monetarist school. Both the increased willingness of public authorities to inject high-powered money into the economy during a crisis and the increased

[4]See, for example, Davis and Gallman (2001) and Kindleberger and Aliber (2011).

reliance of private banking firms upon repo borrowing in place of deposits suggest that this is a permanent change.

The classic work of Friedman and Schwartz (1963) created a master framework for studies ever since by measuring the supply of money, broken into various components, over the near-century that included financial crises before the Great Depression and the financing of the US role in the two world wars of the twentieth century. But the authors saw the role of financial markets as at best secondary to the driving role of the public's demand for money interacting with the government's control over the supply of money, even during the Great Depression. Nevertheless, the "monetarist" movement among economists that they fostered was a necessary step away from the economics profession's focus on the "real economy," measured by adjusting for adventitious movements in monetary prices. Later work on the balance sheets faced by banks led to emphasis on the problem of debt deflation (Bernanke 1995, 2000; Calomiris 1993). By pointing out "just the facts," cliometricians not only force other historians to re-evaluate their interpretations of economic development in the past but also encourage economists to re-think their theories and policy prescriptions.

Concluding Remarks

Financial markets and cliometrics have a checkered history despite their obvious complementarity. Financial markets have always generated and publicized masses of data and quantitative historians supposedly desire lots of data to process and analyze. Why, then, have there not been more studies to draw on to date? The problem seems to be two-fold: first, secondary markets for securities often produce far too much data for the lone investigator to process readily and, second, the analysis will always be challenged for its usefulness, even theoretically much less practically. The first problem is well on the way to being overcome thanks to the continued technological progress in digitizing and encoding data from printed sources onto electronic formats, which in turn can be used to carry out any number of statistical analyses. The second problem has only gradually succumbed to acceptance that price data alone, even without measures of trading volume for the underlying securities, can yield interesting insights into historical issues of consequence. Do the movements in prices of financial assets in organized markets reflect simply the "madness of crowds" or the workings of efficient markets? Even if financial markets are efficient, what "real" fundamental factors determine the prices?

The efforts of financial historians, as well as historians in general, tend to separate them into those who hope that patterns can be found and those who resign themselves to human folly. In financial history, these alternative narratives are between those who hope that market participants jointly can learn to devise time-consistent rules for self-governance and those who are convinced that financial markets need prudential regulators and lenders of last resort. Cliometricians enter these ideologically and politically driven disputes with trepidation, but find that their focus on the past is no refuge from the conflicts of the present. Indeed, issues raised in each new

crisis help pose new questions for analysis of previous episodes, whatever the personal predilection of the historian may be. The meltdown of global financial markets with the unexpected bankruptcy of Lehman Brothers investment bank in September 2008, for example, brought renewed attention to the way banks finance their long-term loans with short-term debts as well as with demand and time deposits. The investigative report by the US Senate (2011) pinpoints the causes of the crisis as high-risk behavior by mortgage lenders, regulatory failure, inflated credit ratings, and investment bank abuses. The individual case studies provide the justification for many elements of the Dodd-Frank bill that were specifically designed to remedy the practices that led to the financial crisis of 2008. But as Gorton (2010) notes, every financial crisis has a unique pattern of events leading to a crisis that first shows up in the market for short-term credit but the unwinding of the crisis takes a particular course depending on historical circumstances and the responses made by governments, banks, and capital markets.

Gorton's unique analysis of the panic of 2007 remains a standard for cliometricians to emulate, eschewing the temptation to generalize found in works like Reinhart and Rogoff (2009) or Kindleberger and Aliber (2011). But using the evidence from the 2007 panic, cliometricians have found new interpretations for earlier crises that are enlightening, starting with the Mississippi Bubble of 1719–20 (Neal 1990, Chap. 4; Velde 2009, 2012) and the South Sea Bubble of 1720 (Neal 1990, Chap. 5; Carlos et al. 2002; Carlos and Neal 2006; Shea 2007, 2009; Frehen et al. 2013; Kleer 2013). Even the little known financial crisis of 1763 in Amsterdam and affecting all of northern Europe has had fresh interpretations in light of modern analysis of financial markets. Schnabel and Shin (2004) show how the chain of short-term credit based on acceptances covered by various commodity contracts broke down with a sudden price shock at the end of the Seven Years' War. Quinn and Roberds (2012) go further by showing how the Bank of Amsterdam acted as an early lender of last resort in response to the crisis, letting one merchant bank fail while supporting the others with repo finance based on silver coins and bullion as collateral. Their earlier work (Quinn and Roberds 2009) explained how the Bank of Amsterdam in the seventeenth century had created "inside" or "high-powered" money so it could play the role of a central bank in the future. Carlos and Neal (2011) argue, nevertheless, that the 1763 crisis marked the eclipse of Amsterdam by London as the center of European finance thereafter. Flandreau et al. (2009a) found that the combined effect of the Seven Years' War and the crisis of 1763 raised all short-term interest rates throughout commercial Europe leading up to the French Revolution. After the end of the French Revolutionary and Napoleonic Wars, the defining moment for British finance in the nineteenth century was the government's regulatory response to the crisis of 1825 (Neal 1998) with ripple effects in the USA (Hilt 2009).

One of the most intensively studied episodes that is still generating new scholarly findings by cliometricians is the international crisis of 1907, which started in the USA with the failure of the Knickerbocker Trust Company, much as the crisis of 2008 started with the bankruptcy of Lehman Brothers. Neal (1971) lauded the benefits of trust companies, while later work by Moen and Tallman (1992, 2012)

pieced together the way the crisis propagated after the initial collapse of a leading trust company and then how the private organization of the New York Clearing House intervened to limit the possibility of contagion. More recent analysis of balance sheet detail from the existing banks and trust companies in New York by Frydman et al. (2012) goes even further to show how information about the specific trust companies affected the pressures placed on them, much in the spirit of Gorton's plea for attention to information flows and content before, during, and after a crisis in financial markets.

Of course, the Great Depression has generated most of the work by cliometricians including dealing with the international aspects, often overlooked by American economists. The key role of bank borrowings to finance their foreign loans caught the original attention of economists at the time, as reported in the Hoover Commission reports at the time (President' Conference on Unemployment 1929, 2 vols), but has recently been reevaluated by cliometricians, first for the USA (White 1984, 1990; Calomiris 1993; Wheelock 1991), Germany (Schnabel 2009), and then for Great Britain (Accominotti 2012). The initiating role of France in undermining the newly established gold exchange standard by resuming its prewar demand for monetary gold was also appreciated at the time, but subsequent work focused on the futile efforts of the USA to cooperate with the UK as financial hegemons (Kindleberger 2000). While Eichengreen (1992) basically blamed the obsession with gold for the general dysfunction of the international financial markets, Irwin (2011) reprised the UK-American argument at the time that France's obsession with gold brought on the Great Depression. Only by reducing the prices of export goods could the rest of the world meet the excess demand for gold created by French policy. Worldwide deflation, as in the 1870s–1880s, again undercut the ability of emerging countries to service the sovereign bonds they had issued. Work by Wandschneider (2008, 2009) shows how differing central bank policies in central Europe created different responses to the challenges of servicing sovereign bonds while maintaining domestic output.

Work continues by cliometricians to pursue Gorton's plea for more in-depth studies of particular crises to see how information flows are disseminated among the various players in each case. Atack and Neal (2009) is one effort to collect deep historical case studies united by a common theme. Atack motivated the introductory chapter by the September 2007 run on Northern Rock's branches throughout the UK and Neal's concluding chapter assessed the evolving subprime crisis in the USA. Echoing the theme of Reinhart and Rogoff, we argued that this crisis had its historical antecedents, going back at least to seventeenth-century Amsterdam. But each author developed his own interpretation of a particular episode. At the heart of each story was the severing of the personal ties that had always been the basis of banking and then substituting reliance upon government agency oversight of the impersonal financial markets that allowed effective securitization for assets in secondary capital markets. The individual authors argued that credit institutions, capital markets, and governments always had difficulty in learning how to coordinate effectively the role of all three sets of organizations in the financial sector whenever financial innovations occurred, usually from the pressures of war finance

upon governments. The concluding chapter, *mea culpa*, argued pessimistically that policymakers usually misread the supposed lessons derived from previous crises. But it also noted optimistically that the current set of policymakers both in the USA and Europe had studied a number of past crises, some quite recent, and perhaps they could learn more quickly from their mistakes than was often the case in the past.

References

Accominotti O (2012) London Merchant Banks, the Central European panic and the sterling crisis of 1931. J Econ Hist 72(1):1–43

Accominotti O, Flandreau M, Rezzik R (2011) The spread of empire: Clio and the measurement of colonial borrowing costs. Econ Hist Rev 64(2):385–407

Archive D (2014) http://www.istitutodatini.it/schede/archivio/home_e.htm

Atack J, Neal L (eds) (2009) The development of financial institutions and markets from the seventeenth century to twenty-first century. Cambridge University Press, Cambridge/New York

Beach B, Norman S, Wills D (2013) Time or spot? A revaluation of Amsterdam market data prior to 1747. Cliometrica 7(1):61–85

Bell AR, Brooks C, Moore TK (2013a) Medieval foreign exchange: a time series analysis. In: Casson M, Hashimzade M (eds) Large data-bases in economic history: research methods and applications. Routledge, Aldershot

Bell AR, Brooks C, Moore TK (2013b) The 'buying and selling of money for time': exchange and interest rates in medieval Europe. Unpublished working paper, University of Reading

Bell RM, Howell M (eds) (1998) The medieval and early modern data bank. http://www2.scc.rutgers.edu/memdb/. Accessed 23 Mar 2014

Bernanke BS (1995) The macroeconomics of the great depression: a comparative approach. J Money Credit Bank 27(1):1–28

Bernanke BS (2000) Essays on the great depression. Princeton University Press, Princeton

Bordo MD, Kydland FE (1995) The gold standard as a rule: an essay in exploration. Explor Econ Hist 32(4):423–464

Bordo MD, Rockoff H (1996) The gold standard as a good housekeeping seal of approval. J Econ Hist 56(2):389–428

Calomiris CW (1993) Financial factors in the great depression. J Econ Perspect 7(2):61–85

Canjels E, Prakash-Canjels G, Taylor AM (2004) Measuring market integration: foreign exchange arbitrage and the gold standard, 1879–1913. Rev Econ Stat 86(4):868–882

Carlos A, Moyen N, Hill J (2002) Royal African company share prices during the South sea bubble. Explor Econ Hist 39(1):61–87

Carlos A, Neal L (2006) The microstructure of the early London capital market: bank of England shareholders during and after the south sea bubble, 1720–1725. Econ Hist Rev 59(3):498–538

Carlos A, Neal L (2011) Amsterdam and London as financial centers in the eighteenth century. Financ Hist Rev 18(1):21–46

Castaing J (1698–1907) The course of the exchange, & c. James Wetenhall, London

Chilosi D, Volckart O (2011) Money, states, and empire: financial integration and institutional change in Central Europe, 1400–1520. J Econ Hist 71(3):762–791

Coffman D'M, Neal L (2013) Introduction. In: Coffman D'M, Leonard A, Neal L (eds) Questioning credible commitment: new perspectives on the glorious revolution and the rise of financial capitalism. Cambridge University Press, Cambridge

Davis LE, Gallman RE (2001) Evolving financial markets and international capital flows. Britain, the Americas, and Australia, 1865–1914. Cambridge University Press, Cambridge/New York

Davis LE, Hughes JRT (1960) A dollar-sterling exchange, 1803–1895. Econ Hist Rev 13(3):58–64

Dickson PGM (1967) The financial revolution in England, a study in the development of public credit, 1688–1756. Macmillan, New York

Dincecco M (2009) Political regimes and sovereign credit risk in Europe, 1750–1913. Eur Rev Econ Hist 13(1):31–63
Eagly R, Kerry Smith V (1976) Domestic and international integration of the London money market, 1731–1789. J Econ Hist 36(2):198–212
EH.net Data bases. http://eh.net/databases/
Eichengreen B (1992) Golden fetters: the gold standard and the great depression, 1919–1939. Oxford University Press, New York
Epstein SR (2000) Freedom and growth: the rise of states and markets in Europe, 1300–1750. Routledge, London/New York
Esteves R (2013) The Bondholder, the sovereign, and the banker: sovereign debt and bondholders' protection before 1914. Eur Rev Econ Hist 17(4):389–407
European State Finance Data Base. http://www.esfdb.org
Ferguson N, Schularick M (2006) The empire effect: the determinants of country risk in the first age of globalization, 1880–1913. J Econ Hist 66(2):283–312
Flandreau M (1996) Adjusting the gold rush: endogenous bullion points and the French balance of payments, 1846–1870. Explor Econ Explor Econ Hist 33(4):417–439
Flandreau M (2004) The glitter of gold: France, bimetallism, and the emergence of the international gold standard, 1848–1873. Oxford University Press, Oxford
Flandreau M, Flores J (2009) Bonds and brands: foundations of sovereign debt markets, 1820–1830. J Econ Hist 69(3):646–684
Flandreau M, Galimard C, Jobst C, Nogués-Marco P (2009a) The bell jar: commercial interest rates between two revolutions, 1688-1789. In: Atack J, Neal L (eds) The origins and development of financial markets and institutions: from the seventeenth century to the present. Cambridge University Press: Cambridge/New York, 161–208
Flandreau M, Galimard C, Jobst C, Nogués-Marco P (2009b) Monetary geography before the industrial revolution. Camb J Reg Econ Soc 2(2):149–171
Forbes KJ, Rigobon R (2002) No contagion, only interdependence: measuring stock market comovements. J Financ 57(5):2223–2261
Frehen RGP, Goetzmann WN, Geert Rouwenhorst K (2013) New evidence on the first financial bubble. J Financ Econ 108:585–607
Friedman M, Schwartz AJ (1963) A monetary history of the United States, 1867–1960. Princeton University Press, Princeton
Frydman C, Hilt E, Zhou LY (2012) Economic effects of runs on early 'Shadow Banks': trust companies and the impact of the panic of 1907. NBER working paper 18624. National Bureau of Economic Research, Cambridge, MA
Gelderblom O, Jonker J (2004) Completing a financial revolution: the finance of the Dutch East India trade and the rise of the Amsterdam capital market, 1595–1612. J Econ Hist 64(3):641–672
Global Finance Data. https://www.globalfinancialdata.com/index.html
Gorton G (2010) Slapped by the invisible hand: the subprime panic of 2007. Oxford University Press, New York
Hilt E (2009) Wall street's first corporate governance crisis: the panic of 1826. NBER working paper 14892. National Bureau of Economic Research, Cambridge, MA
Irwin D (2011) La France a-t-elle cause la Grande Depression? Revue Française d'Economie 25(4):3–10
Kindleberger CP (2000) Manias, panics, and crashes, a history of financial crises, 4th edn. Wiley, New York
Kindleberger CP, Aliber RZ (2011) Manias, panics, and crashes, a history of financial crises, 6th edn. Palgrave Macmillan, New York
Kleer R (2013) Riding the wave: the company's role in the south Sea bubble. Unpublished working paper. Department of Economics, University of Regina, Regina, Saskatchewan
Koudijs P (2011) Trading and financial market efficiency in eighteenth-century Holland. Unpublished PhD Dissertation. Department of Economics, University of Pompeu Fabra
Luckett TM, Lachaier P (1996) Crises financières dans la France du XVIIIe siècle (1954-). Revue d'histoire modern et contemporaine 43(2):266–292

MacDonald J (2013) The importance of not defaulting: The significance of the election of 1710. In: Coffman D'M, Leonard A, Neal L (eds) Questioning credible commitment: new perspectives on the glorious revolution and the rise of financial capitalism. Cambridge University Press, Cambridge
Mauro P, Sussman N, Yafeh Y (2006) Emerging markets and financial globalization: sovereign bond spreads in 1870–1913 and today. Oxford University Press, New York/Oxford
Measuring Worth (2014) http://www.measuringworth.com
Mishkin F (1991) Asymmetric information and financial crises: a historical perspective. In: Glenn Hubbard R (ed) Financial markets and financial crises. University of Chicago Press, Chicago/London
Mitchener KJ, Weidenmier MD (2005) Empire, public goods, and the Roosevelt Corollary. J Econ Hist 65(3):658–692
Mitchener KJ, Weidenmier MD (2010) Supersanctions and sovereign debt repayment. J Int Money Financ 29(1):19–36
Moen J, Tallman EW (1992) The bank panic of 1907: the role of trust companies. J Econ Hist 52(3):611–630
Moen J, Tallman EW (2012) Liquidity creation without a central bank: clearing house loan certificates in the banking panic of 1907. J Financ Stab 8(4):277–291
Neal L (1971) Trust companies and financial innovation. Bus Hist Rev 45(1):35–51
Neal L (1990) The rise of financial capitalism: international capital markets in the age of reason. Cambridge University Press, Cambridge/New York
Neal L (1996) Course of the exchange, London, 1698–1823 and Amsterdamsche Beurs, Amsterdam, 1723–1794. ICPSR01008-v1. Inter-university Consortium for Political and Social Research, Ann Arbor
Neal L (1998) The Bank of England's first return to gold and the stock market crash of 1825. Fed Reserve Bank of St Louis Rev 80:77–82
Neal L, Weidenmier M (2003) Crises in the global economy from tulips to today: contagion and consequences. In: Bordo MD, Taylor AM, Williamson JG (eds) Globalization in historical perspective. University of Chicago Press, Chicago/London
Nogués-Marco P (2013) Competing bimetallic ratios: Amsterdam, London, and bullion arbitrage in mid-eighteenth century. J Econ Hist 73(2):445–476
North DC, Weingast B (1989) Constitutions and commitments: evolution of institutions governing public choice in seventeenth century England. J Econ Hist 49(4):803–822
Odell KA, Weidenmier MD (2004) Real shock, monetary aftershock: the 1906 earthquake and the panic of 1907. J Econ Hist 64(4):1002–1027
Officer LH (1996) Between the dollar-sterling gold points: exchange rates, parity, and market behaviour. Cambridge University Press, Cambridge
Pamuk S, Kivanc Karaman K (2010) Ottoman state finances in European perspective, 1500–1914. J Econ Hist 70(3):593–629
Pezzolo L (2003) Il fisco dei veneziani; Finance pubblica ed economia tra XV e XVII secolo. Cierre Edizioni, Verona
Pezzolo L (2013) Sovereign debts, political structure, and institutional commitments. In: Coffman D'M, Leonard A, Neal L (eds) Questioning credible commitment: new perspectives on the glorious revolution and the rise of financial capitalism. Cambridge University Press, Cambridge
Pezzolo L (2014) The *via italiana* to capitalism, ch. 10. In: Neal L, Williamson JG (eds) The Cambridge history of capitalism, vol 1. The rise of capitalism: from ancient origins to 1848. Cambridge University Press, Cambridge, pp 267–313
Pezzolo L, Tattara G (2008) 'Una fiera senza luogo': was Bisenzone an international capital market in sixteenth-century Italy?". J Econ Hist 68(4):1098–1122
President's Conference on Unemployment (1929) Recent economic changes in the United States, 2 vols. McGraw-Hill, New York
Quinn S (1996) Gold, silver, and the glorious revolution: arbitrage between bills of exchange and bullion. Econ Hist Rev 49(3):473–490

Quinn S (2001) The glorious revolution's effect of English private finance: a microhistory, 1680–1705. J Econ Hist 61(3):593–614
Quinn S, Roberds W (2009) An economic explanation of the early Amsterdam bank, debasement, bills of exchange, and the emergence of the first central bank. In: Atack J, Neal L (eds) The origins and development of financial markets and institutions: from the seventeenth century to the present. Cambridge University Press: Cambridge/New York, 32–70
Quinn S, Roberds W (2012) Responding to a shadow banking crisis: the lessons of 1763. Unpublished working paper. Department of Economics, Texas Christian University
Reinhart CM, Rogoff KS (2009) This time is different: eight centuries of financial folly. Princeton University Press, Princeton
Schnabel I (2009) The role of liquidity and implicit guarantees in the German twin crisis of 1931. J Int Money Financ 28(1):1–25
Schnabel I, Shin HS (2004) Liquidity and contagion: the crisis of 1763. J Eur Econ Assoc 2(6):929–968
Schubert ES (1989) Arbitrage in the foreign exchange markets of London and Amsterdam during the 18th century. Explor Econ Hist 26(1):1–20
Schularick M, Taylor AM (2012) Credit booms gone bust: monetary policy, leverage cycles and financial crises, 1870–2008. Am Econ Rev 102(2):1029–1061
Shea GS (2007) Financial market analysis can go mad (in the search for irrational behavior during the South Sea Bubble). Econ Hist Rev 60(4):742–765
Shea GS (2009) Sir George Caswall vs. the Duke of Portland: financial contracts and litigation in the wake of the South Sea Bubble. In: Atack J, Neal L (eds) The origins and development of financial markets and institutions: from the seventeenth century to the present. Cambridge University Press: Cambridge/New York, 121–160
Stasavage D (2003) Public debt and the birth of the democratic state: France and Great Britain, 1688–1789. Cambridge University Press, New York
Stasavage D (2011) States of credit: size, power, and the development of European polities. Princeton University Press, Princeton
Sussman N, Yafeh Y (2006) Institutional reforms, financial development, and sovereign debt: Britain, 1690–1790. J Econ Hist 66(4):906–935
Tomz M (2007) Reputation and international cooperation. Sovereign debt across three centuries. Princeton University Press, Princeton
Tuncer AC (2011) Fiscal autonomy, monetary regime and sovereign risk: foreign borrowing and international financial control in the Ottoman Empire, Greece and Egypt during the classical gold standard era. Unpublished PhD thesis. London School of Economics and Political Science
United States Senate. Permanent Subcommittee on Investigations (2011) Wall street and the financial crisis: anatomy of a financial collapse, US Government Printing Office, Washington, DC
Vam Malle Sabouret C (2008) De la naissance de la dette publique au plafond souverain: Rôle des gouvernements régionaux dans l'évolution de la dette publique. Unpublished doctoral thesis. Finances Internationales, Institut d'Etudes Politiques de Paris
Van Bochove C (2014) External debt and commitment mechanisms: Danish borrowing in Holland, 1763–1825. Econ Hist Rev (forthcoming)
Velde F (2009) John Law's system and its aftermath, 1718-1725. In: Atack J, Neal L (eds) The origins and development of financial markets and institutions: from the seventeenth century to the present. Cambridge University Press: Cambridge/New York, 99–120
Velde F (2012) John Law and his experiment with France, 1715–1726. In: Caprio G (ed) The handbook of key global financial markets, institutions, and infrastructure, vol I. Elsevier, Oxford, pp 169–174
Velde FR, Weber WE (2000) A model of bimetallism. J Polit Econ 108(6):1210–1234
Velde FR, Weir D (1992) The financial market and government debt policy in France, 1746–1793. J Econ Hist 52(1):1–39

Vizcarra C (2009) Guano, credible commitments, and sovereign debt repayment in the nineteenth century. J Econ Hist 69(2):358–387
Volckart O (1996) Die Münzpolitik im Deutschordensland und Herzogtum Preussen von 1370 bis 1550. Harssowitz, Wiesbaden
Volckart O (2014) 14th-16th century exchange rates. http://www.lse.ac.uk/economicHistory/Research/Late Medieval Financial Market/datasheets/datasheetindex.aspx. Accessed 23 Mar 2014
Volckart O, Wolf N (2006) Estimating financial integration in the middle ages: what can we learn from a TAR model? J Econ Hist 66(1):122–139
Wandschneider K (2008) The stability of the interwar gold exchange standard: did politics matter?". J Econ Hist 68(1):151–181
Wandschneider K (2009) Central bank reaction function during the inter-war gold standard: a view from the periphery. In: Atack J, Neal L (eds) The development of financial markets and institutions: from the seventeenth century to the present. Cambridge University Press: Cambridge/New York, 388–415
Wells J, Willis D (2000) Revolution, restoration, and debt repudiation: the Jacobite threat to England's institutions and economic growth. J Econ Hist 60(2):418–441
Wheelock D (1991) The strategy and consistency of federal reserve monetary policy, 1924–1933. Cambridge University Press, Cambridge/New York
White EN (1984) A reinterpretation of the banking crisis of 1930. J Econ Hist 44(1):119–138
White EN (1990) The stock market boom and crash of 1929 revisited. J Econ Perspect 4:67–84

Financial Systems

Caroline Fohlin

Contents

Abstract

This chapter elucidates the key debates surrounding the optimal design of financial systems and institutions: bank-based versus market-based; universal versus specialized banking; relationship versus arms-length banking. The chapter also examines the historical pattern of financial system development – explaining the economic, legal, and political factors that influenced the shape of these systems as well as the long-run growth outcomes observed among the group of economies that underwent industrialization prior to World War I. The extensive evidence and analyses available indicate that financial systems historically took on a wide and complex range of forms that are difficult to categorize narrowly, yet provided

C. Fohlin (✉)
Johns Hopkins University, Baltimore, MD, USA

Emory University, Atlanta, GA, USA
e-mail: fohlin@jhu.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_7

similar functions; thus arguing for a functional, rather than institutional, approach to financial system design and regulation. Moreover, the research to date strongly supports the idea of persistence and path dependency in financial system design, that economic conditions at the time of industrialization help set the initial conditions that shape financial system and banking institution design, and historical political conditions, such as centralization of power, plays an ancillary role via the extent of regulation on banks and the development of free capital markets. In other words, history matters.

Keywords
Financial systems · Law and finance · Finance and growth

What Does a Financial System Do?

The financial system is the set of institutions and markets that gathers excess funds from savers – whether households or businesses – and allocates financial capital to entrepreneurs and others in need of credit. In the process, the financial system produces information and distributes risk throughout the economy and among its participants. Merton (1993) summarizes even more succinctly the primary function of any financial system: "to facilitate the allocation and deployment of economic resources, both spatially and temporally, in an uncertain environment."

Well-functioning financial systems must provide several core functions (Merton 1993; Merton and Bodie 1995):

- Clearing and settling payments
- Pooling or mobilizing resources
- Transferring economic resources, inter-temporally or geographically
- Managing risk
- Pricing information
- Dealing with information and incentive problems

Financial systems may provide these services via a wide range of institutions and markets. Financial institutions include, among others, commercial banks, savings institutions and thrifts, credit cooperatives, investment banks, insurance companies, trust companies, pension funds, mutual funds, hedge funds, and private equity. Institutions come in a wide range of sizes and ownership structures – from private partnerships to enormous multinational conglomerates to government-owned enterprises. Financial markets offer centralized, liquid trading in essentially any financial claim, from debt to equities, commodities to foreign exchange, and a wide array of derivatives.

The core components of modern financial systems grew out of small, rudimentary, and entrepreneurial initiatives at the earliest stages of economic activity: the merchants of the medieval era, the goldsmiths of seventeenth-century London, and the fairs and early commodity markets that dotted Europe throughout the medieval

and modern periods.[1] In their own way, each of these organizations participated in payments clearing and settling, capital pooling and mobilization, risk management, information aggregation, asset pricing, incentive matching, and agent supervision.

Financial systems grew and diversified as industrialization took hold in England and then the European continent. New forms of financial contracting, institutions, and markets evolved to handle more extensive and complex needs of funding the larger-scale and scope of industrial enterprises. Thus, financial and industrial revolutions progressed largely in parallel, with entrepreneurial financiers innovating to serve the incipient demands from all sectors of the economy – industry, agriculture, transportation, and trade. Political boundaries and legal institutions also continued to shift repeatedly throughout this early stage of financial and industrial development, and monetary systems developed and changed as well. Some countries with stronger central government control instituted central banks and fiat currency, though the degree varied among countries and over a wide time span.

The greatest leap toward modernized financial systems came in rapidly industrializing areas of the early to mid-nineteenth centuries and spread with industrialization to most of the rest of the world over the remainder of that century. Significant shifts and redesigns of financial systems came with the crisis of the Great Depression, the post-WWII reconstruction, the wave of liberalization of the 1980s–1990s, and most recently in response to the global financial crisis of 2008 and the ensuing "great recession." For the most part, these episodes caused some reshaping of institutions and markets and their regulation by the government, but they did not set off fundamental change in the functions of the financial system or the existence of institutions and markets that provide these functions.

Academic study of financial systems dates back to the beginning of financial systems and continues unabated. The literature covers a wide array of topics, some of which provoke significant debates. The changing regulation and organization of financial institutions and markets in the late 1980s through the 1990s, along with several areas of transformation in political and economic systems, set off an active academic literature on financial system design that became particularly active in the late 1990s and early 2000s and continues today.

Three of the key areas of research and debate revolve around the following topics:

1. The design of financial institutions and systems: functional versus institutional approaches.
2. Why do financial systems differ across countries: legal origins versus political and economic explanations?
3. Does financial system design affect an economy's long-run economic growth rate?

[1] On the London goldsmith bankers and the British financial revolution, see Temin and Voth (2013).

The next three sections take up these topics in turn, providing a survey of the current thinking and remaining issues for further research. The discussion focuses on corporate finance systems and related areas of corporate governance.[2]

Designing Financial Systems: Functions Versus Institutions

Financing modern industry hinges on a system that allows those with surplus resources to convert their excess into financial capital and channel those funds into productive investment opportunities. This process often means connecting entrepreneurs with capital owners outside the entrepreneurs' circles of friends and families, creating a need for contracting and enforcement devices as well as a means for coping with asymmetric information and incentive problems. Virtually all developed economies employ limited-liability, joint-stock corporations to facilitate external financing. Most of these countries formalized, standardized, and liberalized incorporation and legal liability systems during the nineteenth century – many during the wave of heavy industrialization of the 1850s–1870s. Within a decade or two thereafter, businesses and entrepreneurs in these countries turned to corporations in order to grow and diversify, financing an unprecedented scale of operations. The acceleration of incorporation in most places during the last years of the nineteenth century and into the twentieth spurred rapid advancement in the corporate financial sector and of the securities markets. Despite their considerable differences in culture, society, legal systems, and political processes, the world's most advanced economies all created well-functioning systems for corporate finance by the late nineteenth century.[3]

For businesses in this period, banks often served as one of the most important sources of outside capital, whether for short-term trade credit or longer-term investment finance. Thus, industrial development usually proceeded hand in hand with the growth of commercial banking. As economies industrialized, financial intermediaries changed, and industrial organization of banking changed as well. The largest banks grew larger, and densely networked, nationwide banks emerged nearly worldwide.[4] Commercial banks took on a varying array of functions, sometimes quite narrowly focused on short-term credit, other times offering investment banking, brokerage, and even strategic advising.

[2]Given space and time constraints, the chapter leaves out monetary systems and central banking.

[3]Fohlin (2012) and Allen et al. (2010) provide detailed historical comparisons of the corporate finance systems of the United Kingdom, the United States, Germany, Japan, and (in Fohlin 2012) Italy. Fohlin (2012) also compares more schematically the financial systems of a larger set of industrialized economies of the prewar period. Morck's (2005) edited volume contains historical studies of the corporate governance systems of several different countries.

[4]Regulatory restrictions prevented the natural progression of banking in the United States. Even there, a few banks grew very large, and banks developed a correspondent system to replicate national branching.

Commercial banks also differed in their responses to changing needs in industrial finance and their engagement in corporate governance. The corporate firms that emerged over the last half of the nineteenth century began to loosen the ties between families and the firms they started. As corporate management began to separate from ownership, investors required new modes of corporate governance. Trading corporate securities on secondary markets often dispersed the ownership of firms and demanded oversight mechanisms to protect smaller shareholders. Thus, industrialized economies developed corporate governance institutions, and banks played varying roles in those arrangements as well.

All of these dimensions of the financial system – the organization of banks, the extent of securities markets, the relationship among banks and markets, and corporate governance – differ to some extent over time and across countries. Thus, financial systems can be characterized along these various dimensions, most notably by the functions they serve or the organizational forms they take.

Post-WWII economic historians took up this topic most actively with the publication of Gerschenkron's *Economic Backwardness in Historical Perspective* and Goldsmith (1969) *Financial Structure and Development*, among others. Gerschenkron, in particular, influenced a generation of financial historians to differentiate among the types or organizational forms that financial institutions could take, positing a relationship between the level of economic development of a country and the type of banking institutions they created. By the 1980s, when Germany and Japan were growing rapidly and the United States saw itself lagging, attention turned to the design of financial systems to explain why. Those cross-country comparisons led to the deregulation of US banking and the Big Bang in the United Kingdom – among other efforts to stimulate the development of German-style universal banking and relationship banking that seemingly helped produce the postwar economic miracle. These events led to a resurgence in interest and ultimately to a reevaluation of Gerschenkron's and Goldsmith's ideas on financial institution and system types and their importance for economic growth.

The Standard Paradigm of Financial System "Types"

The study of financial system types subsumes a number of issues: the organizational design of institutions and markets, the activities and functions of different institutions, and the relative use of financial institutions versus markets. The literature on financial systems focuses on the distinction between bank-based and market-based financial systems, between universal and specialized organizational forms of banking, and between relational versus arm's-length approaches to banking.

These distinctions, however, fit empirical observation only in a rough manner: most financial systems are better characterized using a functional approach that can mix the individual components of one or the other system "type." Still, the notion of type animates a long line of research on both historical and contemporary financial systems, and some kernel of truth remains in the notion of types of systems and of institutions. In this literature, systems and their respective

institutions are divided along three chief dichotomies: universal versus specialized banking, relationship versus arm's-length banking, and more generally bank-based versus market-based systems. The following considers the three issues in turn. The subsequent section examines what we know about historical cases.[5]

Universal Versus Specialized Banking

Banking institutions provide a range of functions, from very short-term credits to longer-term debt to underwriting of securities. The combination of services that an institution provides dictates how it is categorized. Institutions are commonly divided into two main types: universal or specialized, with the former offering a broad scope of services and the latter naturally providing a more limited range. A true universal bank is allowed to provide almost any financial product or service. However, the fundamental distinguishing feature of universal banking historically is the combination of commercial banking functions (short-term credit, deposit taking, payments clearing, bill discounting) with investment banking services (underwriting and trading in securities). Modern universal banks also sell insurance, mortgages, and investment funds, and they create and trade more complex financial products, usually through affiliates. The counterpoint to universal banking – so-called "specialized" banking – separates investment and commercial banking into separate sets of institutions.

Relationship Versus Arm's-Length Banking

The constructs of "relationship" and "arm's-length" banking classify institutions by their involvement in corporate governance. Compared to universality, there is less agreement over what precisely constitutes "relationship banking" in a formal, measureable sense. The term is sometimes used loosely to refer to banks that work closely with customers, but most research considers some combination of the following three types of more formal relationships: proxy voting of deposited equity shares taken by banks, equity shares held directly by banks, and corporate board positions filled by bank directors.[6]

The three methods of engaging in relationships bring different levels of ownership and control rights. The strongest relationship, direct ownership of equity, gives banks both ownership (cash flow) rights and control (voting power) rights. Equity stakes theoretically align banks' incentives with those of other firm shareholders and promote efficient provision of financing. In some cases the banks employed an indirect method of gaining control rights over corporations: proxy voting rights signed over by shareholders. In the proxy voting system, shareholders grant the bank power of attorney over their shares, resulting in additional voting power for the banks. Before the subsequent unraveling of the system in 1990, German banks held on average approximately 24.3% of effective voting rights due to direct equity holdings and 29.5% on average due to proxy voting rights at

[5]This section is based on Fohlin (2012).

[6]See Fohlin (2012, Chap. 3) and on Germany specifically see Fohlin (2005, 2007a, b).

general meetings of their current clients.[7] From this example, it is clear that the proxy voting system can provide banks with significant power over firm management even without ownership rights.

Using their voting power, whether direct or indirect, banks can theoretically help elect their chosen representatives to a company's board of directors and can vote or appoint their representatives into various positions within the corporate boards. These positions then allow the bank to influence the selection of management and other key corporate decisions.[8]

Relationship banking may prove even more important among firms that are organized without publicly traded equity. In these cases, relationship banking necessarily takes an informal shape. While these relationships consist of weaker legal connections, they may actually prove stronger, if firms have limited access to capital market alternatives. Presumably, relationship banking ought to also imply that banks provide helpful advice to young firms, but that sort of criterion is difficult to formalize or measure.[9]

In the dichotomy of financial systems, the natural opposite of relationship banking is "arm's-length" banking. In arm's-length systems, banks simply provide financing, perhaps in a one-shot deal, and take no enduring corporate governance role in nonfinancial firms. In "arm's-length" systems, profit motive theoretically drives information gathering that supersedes the need for closer monitoring by bankers. No system would fit this extreme characterization, and a few even match a weaker form of it.

Market-Based Versus Bank-Based Financial Systems

The third financial system dichotomy distinguishes between market-based and bank-based systems. Systems supporting large, active securities markets, and in which

[7] 1990 data taken from a survey of 144 large German firms' general meeting minutes, quoted in Elsas and Krahnen 2003.

[8] For historical country studies of corporate governance practices, see the volume edited by Morck (2005). A recent volume edited by David and Westerhuis (2014) provides long-run country studies more specifically on corporate networks. The *Oxford Handbook of Banking* edited by Berger et al. (2014) contains several relevant chapters on banking generally. In more recent times, the proxy voting system has come to incorporate a range of financial institutions, such as mutual funds and investment advisors. See Ferreira and Matos (2012) on the impact that proxy voting by banks has on corporate lending globally.

[9] Many studies of young firms focus on venture capital financing and the role of venture capitalists, as in Hochberg et al. (2007). Most young firms do not find financing from venture capital organizations but rather from banks. See Hellmann et al. (2007) on venture capital activities of banks. Ivashina and Kovner (2011) use proportion of lending to measure relationship strength in their study of the impact on lending costs of relationships between LBO firms and banks. Santikian (2014) emphasizes the role of noninterest revenue generation and added connections with new borrowers as measures of relationship strength. The Kauffman Foundation (2013) sponsored a large longitudinal survey of US firms founded in 2004, and they report results on their website: http://www.kauffman.org/~/media/kauffman_org/research%20reports%20and%20covers/2013/06/kauffmanfirmsurvey2013.pdf.

corporate firms use market-based financing, are often referred to as "market oriented." Systems in which banks provide the majority of corporate finance are known as "bank based."

Connections Among the Three System Dichotomies

The literature usually associates bank-based financial systems with universal banking and market-oriented systems with specialized banking. Bank dominance has become nearly synonymous with universality while market orientation has become linked to specialization.[10] The past literature also typically assumes that relationship banking is part and parcel of universal banking, perhaps because of Gerschenkron's focus on the German financial system of the late nineteenth century and similar systems.[11] Putting it all together, we arrive at the three-part financial system paradigm that aligns universal banking, relationship banking, and bank-oriented financing, on the one hand, and specialized banking, arm's-length lending, and market orientation, on the other.[12]

There is some justification for the view: banks and markets may compete in both the initial placement and the ongoing trading of securities. If universal banks internalize market functions, they may impinge on the liquidity of stock exchanges, implying a lower level of market development.[13] For example, universal banks that provided brokerage services may have traded securities among their customers and taken only the net transaction to the market. In contrast, market-based systems by definition support large, liquid equity markets. While such internalization could be plausible in a rudimentary financial system, or in thinly traded securities, universal banking generally works with, not against, active securities markets. A bank cannot become "universal" without investment banking operations – underwriting and brokerage services – to perform. And investment banking requires the use and intermediation of securitized financial instruments. The existence of markets in which to trade securities facilitates the use of these instruments and therefore promotes the investment side of the universal banking business.

Setting up banks and markets as opposites misses the fundamental complementarities between them and ignores their complexity and heterogeneity. The bank versus market dichotomy therefore provides a false sense of clarity in comparing national financial systems, as an examination of historical financial systems demonstrates.

---

[10]See Levine and Zervos (1998) on the 1990s and Fohlin (2012) for historical and long-term patterns.

[11]Gerschenkron's seminal work is his 1962 *Economic Backwardness in Historical Perspective*. He had also written on Italy in 1955 and later on Russia (1970). See also Gerschenkron (1968). Sylla and Toniolo's (1991) edited volume contains several essays relating to and analyzing Gerschenkron's work. See, in particular, Sylla's chapter on banking.

[12]The stylized view is most succinctly laid out by Dietl (1998).

[13]See Bhide (1993) and Levine (2002).

Classifying Historical Systems

The idea of financial system types arose mainly from observation of a relatively small range of countries and time period. Thus, to understand how well the typology fits the historical evidence more broadly, Fohlin (2012) went about classifying historical financial systems based on examination of 26 national financial systems starting in the mid-nineteenth century and extending to the late twentieth century.[14] The study included all countries for which reliable information was available, including a sampling from Europe (e.g., France, Germany, the United Kingdom, Denmark), North America (the United States, Canada, and Mexico), South America (Argentina and Brazil), and East Asia (India, Japan). The classification scheme included the three primary dichotomies of financial system structure and also examined the extent of bank branching:

- Universality versus specialization (whether or not commercial banks also perform investment services)
- Relationship versus arm's-length banking (the existence of any equity stakes, proxy voting, or interlocking directorates between banks and nonfinancial firms)
- Bank-based versus market-oriented system (heavy use of bank funding versus securities markets)

In addition to the broad-based survey evidence, the study included in-depth analysis of five classic cases: Germany, Italy, and Japan in the "universal relationship bank" category and the United States and United Kingdom in the "specialized arm's-length market" category. After pulling together a large array of qualitative and quantitative evidence, Fohlin (2012) argues that financial systems have no fit within clear, unchanging categories; however certain financial system characteristics do allow a rough classification (Table 1).[15]

Universality Versus Specialization

The necessity for investment banking services naturally grew with the onset of free incorporation and securitized debt, as investment bankers provide the intermediation between investors and issuers. The spread of publicly traded stocks and bonds propelled the development of secondary markets in which to trade these securities,

[14]For most of the countries listed, the determination of banking characteristics stemmed from exhaustive searches of secondary literature as well as discussions with several scholars who have studied these systems. Gaps remain where information is too sparse to support a certain categorization. Further studies have appeared since, including Musacchio's (2009) extensive study of Brazil and Colvin et al.'s (2014) analysis of a large sample of Dutch banks in the 1920s crisis there.

[15]One may also consider national laws and regulations regarding banking scope, corporate governance relationships, bank branching, and operations of securities markets. Because regulations constraining banking operations vary in their intensity and enforcement, as well, systems have historically differed even in the absence of regulatory restraints; the "de facto" approach may better capture actual rather than hypothetical differences among systems.

Table 1 Banking system characteristics in the nineteenth and twentieth centuries

Country	Time period	Universal	Bank seats on company boards	Equity shareholdings by banks	Proxy voting by banks[a]	Extensive branch networks[b]
Argentina	Esp. after 1890	Mixed	Some	Few	?	1
	1990s	Restricted	Restricted	Restricted	Restricted	1
Australia	Before 1890s	1	?	Some	?	1
	1895–1950s	0	?	Few	?	1
	1990s	Unrestricted	Some	Some	Some	1
Austria-Hungary	Pre-WWII	1	1	1	1	1
	1990s (Austria)	1	1	1	1	1
Belgium	1830s–1934	Mixed	?	1	?	1
	1934–1970s[c]	0	?	0	?	1
	1990s	Mixed	Restricted	Restricted	Restricted	1
Brazil	1850–1900	Mixed	0	Some	?	1
	Post-1900	1	Some	0	0	1
	1990s	Mixed	Restricted	Restricted	Restricted	1
Canada	1900–1913	Mixed	Some	Some	?	1
	Esp. after WWI	0	Some	Few	?	1
	1990s	Mixed	Restricted	Restricted	Restricted	1
Denmark	1870–1913	Mixed	Some	Some	?	0
	1990s	Unrestricted	Unrestricted	Unrestricted	Unrestricted	1

England	Esp. after 1850s	0	Few	Few	?	1
	1990s (UK)	Unrestricted	Unrestricted	Unrestricted	Unrestricted	1
Finland	Pre-WWI	0	Some	Few	1	1
	1920s–1980s	1	1	Some	1	1
	1990s	1	Some	Some	Some	1
France	1800–1880	1	Few	Few	?	0
	1880–1913	Mixed[d]	1	Some	1	1
	1941–1984	0	?		?	1
	1990s	Mixed	1	1	1	1
Germany	Pre-1880	1	Few	Few	?	0
	Esp. after 1890s	1	1	Some	1	1
	1990s	1	1	1	1	1
Greece	Pre-WWI	Mixed	Some	Some	?	1
	1928–1962	0	1	1	?	1
	1990s	Mixed	Unrestricted	Unrestricted	Unrestricted	1
India	Esp. after 1850s	0	?	Few	?	1
	1990s	Mixed	Restricted	Restricted	Restricted	1
Ireland	Esp. after 1850s	0	?	Few	?	1
	1990s	Unrestricted	Unrestricted	Unrestricted	Unrestricted	1
Italy	1890s–1920s	1	Top banks	1	?	1
	1930s–1980s	0	?	0	?	1
	1990s	1	1	1	1	1

(*continued*)

Table 1 (continued)

Country	Time period	Universal	Bank seats on company boards	Equity shareholdings by banks	Proxy voting by banks[a]	Extensive branch networks[b]
Japan	Pre-WWII	1[e]	Few	Few	?	1
	Post-WWII	0	1	1	?	1
	1990s	Restricted	Restricted	Restricted	Restricted	1
Mexico	1897–1913	Few	Some	Some	?	1
	1990s	Mixed	0	0	0	1
Netherlands	1860–1920s[f]	Mixed	1	1	?	1
	1990s	1	1	1	1	1
New Zealand	1870–1895	Mixed	?	Some	?	1
	1895	0	?	Few	?	1
	1990s	Mixed	Unrestricted	Unrestricted	Unrestricted	1
Norway	Pre-WWII	0	0	0	?	0
	1990s	Mixed	Some	Some	Some	1
Portugal	1890s–WWII	1	1	Some	?	few
	Post-WWII	1	1	1	?	1
	1990s	1	Some	Some	Some	1
Russia	1890s–WWII	1	1	1	?	1
	1990s	Mixed				
Spain	Esp. after 1890s	Mixed	1	1	?	1
	1990s	1	1	1	1	1

Sweden	Esp. after 1850s	Mixed	1	Some[g]	Some	1
	1990s	1	Restricted	Restricted	Restricted	1
Switzerland	Esp. post-1890s	Mixed	1	Some	?	1
	1990s	1	1	1	1	1
United States	Before 1914	1[h]	1	1	?	0
	1914–1933	1	Some	Few	?	Some
	After 1933	0	Some	0	?	Some
	1990s	Restricted	Restricted	Restricted	Restricted	Some

Source: Fohlin (2012, Table 6.1)

[a]In many cases, the extent of proxy voting by banks is difficult to measure accurately

[b]In most cases, branching proceeded slowly until after the second half of the nineteenth century or even later

[c]After 1934, mixed banks were required to split into deposit banks and holding companies and the banks could not hold shares

[d]Some universal banks, some specialized. French universal banks moved more toward straight deposit banking after 1880

[e]Japanese banks combined commercial and investment banking but underwrote little corporate equity; they were prohibited from acting as dealers in secondary markets

[f]Some universal, some primarily commercial. (Jonker argues that Dutch banks were universal only between 1910 and 1920. After about 1924, through WWII, the Dutch banks reverted to primarily commercial banking, with some low-risk company flotations)

[g]Intentional acquisition of shares was illegal until 1909. Shareholdings could result from collateral held on bad loans

[h]Bank structure varied considerably. Services were combined through commercial bank subsidiaries of investment banks. Compliance to (or interpretation of) the new laws also varied

especially toward the end of the nineteenth century. Thus, banks that provided underwriting and brokerage services evolved in a variety of functional and legal forms over the course of the nineteenth century, with the most rapid development in many countries in the mid-nineteenth to late nineteenth century – typically in conjunction with related developments in corporate and securities laws and institutions.

Germany, with its dozen or more large-scale universal banks, offers the classic example of universal banking (Fohlin 2007a), but most of continental Europe followed a similar pattern. Universal banks had emerged in Belgium even earlier and in France almost simultaneously. Universal banking spread to several other European countries in the 1890s: Finland, Italy, Spain, Sweden, Ireland, and Switzerland. In Italy, the financial system remained compartmentalized until the early 1890s, when it suffered a severe crisis and the failure of many banks. The crisis prompted the establishment of a central banking system and the importation of German-style universal banking.

Universal-type banks spread over many parts of the industrialized world in the nineteenth century. Even where universal banking institutions grew up and dominated the corporate banking scene, other types of institutions often thrived. For example, in Belgium, a small number of large-scale, typically limited-liability universal banks operated along with smaller, specialized banks focusing on a narrower range of services. To varying degrees, this mixture of institutions emerged in all parts of continental Europe (Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden, and Switzerland), parts of Latin America (e.g., Argentina, Brazil, and Mexico), and, in a limited way, even in Australia, New Zealand, and the United States.

Specialized banking grew out of the more advanced economic context of England and its long history of commercial and merchant operations around the globe. The investment banks and merchant banking houses evolved separately from the commercial banks in part as a natural consequence of the extent of the markets for those services and the fact that the early investment banking services revolved heavily around government finance with little possibility to gain from economies of scope between investment and commercial banking.[16] Commercial and investment banking remained mostly separated in the British financial system throughout the nineteenth and much of the twentieth centuries. Most countries with similar financial systems imported their legal and financial structures through colonization or other close ties with England.[17] American banks retained significant legal and organizational separation even while combining functions in some institutions and creating close operational ties between investment and commercial banks. Thus, Fohlin (2012) refers to the US banking system as quasi-universal in the pre-WWI era.

Some countries, such as Australia, France, the Netherlands, Belgium, Italy, Russia, and the United States, developed universal banking practices in the

[16]See Collins and Baker (2004) on commercial banking in England and Wales from 1860 to WWI.

[17]See Fohlin (2012) for a list of countries and further discussion.

nineteenth century, but then restricted or abandoned it at various points later on.[18] Notably, the United States began the twentieth century with (quasi-) universal banking but sharply restricted it with the passage of the Glass-Steagall Act in 1933 and the Bank Holding Company Act in 1956, both as responses to the Depression Era bank failures. Even into the 1990s, the United States did not develop unrestricted universal banking. The Glass-Steagall Act persisted until its repeal in 1998, after much debate and as financial and political reality overtook the antiquated law.[19] Yet another group of countries developed mixed or partially restricted systems: Argentina, Belgium, Brazil, Canada, Greece, India, Mexico, New Zealand, Norway, and Russia.

Germany, Austria-Hungary, and Portugal were the only countries to maintain universal banking institutions continuously from the late nineteenth century into the late twentieth century. Germany is the archetype of the universal system, having developed joint-stock universal banks in the mid-nineteenth century and then using these institutions to mobilize extensive capital to finance a growing population of corporations and large private enterprises.[20]

Relationship Versus Arm's-Length Banking

The historical evidence on prevalence of relationship banking remains incomplete, and there is no precise way of determining whether a particular set of banking institutions constitutes a relationship banking system. Recent efforts toward categorization have turned up new evidence and have established some classification parameters regarding bank engagement in some mixture of the three primary attributes: bank representatives on firm boards, direct equity shares held by banks, and proxy voting. The crucial point is that banks' activities gain them significant formal control over the management decisions of nonfinancial firms; ownership, or rights to the companies' cash flows, takes a lesser priority.

Prior to WWI, formalized banking relationships developed gradually and unevenly in different places. Until the 1860s–1870s, when many countries liberalized incorporation laws and instituted corporate governance requirements, such as boards of directors, the opportunities for formal bank connections remained constrained. Few studies have attempted to quantify the extent of these practices, but the qualitative descriptions available suggest that most banks played a small role in nonfinancial corporate governance for most of the nineteenth century.

The first industrial banks of the 1850s in Germany, Belgium, France, the Netherlands, and elsewhere often took over the capital of a few firms for which the banks

[18]See Amidei and Giordano (2010).

[19]The merger between Travelers Insurance Group and Citibank in early 1998 was a direct challenge to the early twentieth century banking acts.

[20]Gerschenkron (1962) provided the seminal postwar exposition of the German system; however, Riesser (1910 German original; translated by the US National Monetary Commission in 1911) and Jeidels (1905) offered detailed contemporary accounts of the German banking system, and Whale (1930) added further analysis – all of which seem to have influenced Gerschenkron's thinking on the German banking system. See the discussion in Fohlin (2007a).

were managing a new issue. The downturn in the markets of the mid- to late 1850s left the banks holding major stakes in a few firms, and a significant number of banks failed. The losses taught the surviving banks and newcomers to avoid such costly mistakes in the future (prominent examples include the French Crédit Mobilier and the German Disconto-Gesellschaft and Darmstädter Bank). Equity participations were largely accidental, in this case a result of the market declines, and were not pursued as a means of corporate control. In fact, historical studies highlight the dismay of bank shareholders when bank funds became tied up in long-term equity holdings.[21]

Fohlin (2007a) argues that interlocking directorates arose in Germany most extensively toward the end of the nineteenth century, and from the viewpoint of the early decades of the twentieth century, Germany not only had one of the largest and most complete universal banking systems but had also developed relationship banking practices of various sorts. The banks could vote their representatives onto corporate boards using proxy voting rights gained by taking equity shares placed on deposit by customers. The larger the bank and the more widely held the corporation, the more likely the bank would receive proxy votes with which to vote its representatives onto the company board. Banks in a number of countries took to relationship practices much more actively around the turn of the twentieth century, but relationship banking practices varied quite a bit in their origins and importance. In some systems, what looked like equity stakes in fact arose out of underwriting activities of the investment banking arms of universal banks. Most banks, as in Germany, engaged via proxy voting and board positions, rather than long-term, direct equity stakes. Moreover, banks took board positions in a minority of firms.

Fohlin (2012) also evaluated relationship banking practices in the sample of 26 countries, demonstrating that not all universal banks perform the complete range of relationship banking functions and not all financial institutions that provide some of these functions are universal banks. The study showed that the strength and prevalence of relationship banking practices varies across countries and across time periods. In the late nineteenth century, Austria-Hungary was the only country (for which there is data) that engaged in the full range of relationship banking activities in a widespread fashion: seats on company boards, equity share holdings, and proxy voting.

Proxy voting data is difficult to collect, so we cannot say for sure how widespread the practice was. In Italy, the Netherlands, Russia, Spain, and the United States, banks in the late nineteenth century also took seats on company boards and held equity share holdings. In none of these cases is there comprehensive data on proxy voting. Anecdotal evidence from well-known bankers – such as J. P. Morgan – suggests that some version of proxy voting did provide bankers with a measure of corporate control rights. Certainly German, Austrian, Belgian, and Italian universal banks took positions on a significant number of firms' boards, but they did so primarily in the largest firms with publicly traded equity.[22] Most of the large banks

[21] See Paulet (2002) on the Credit Mobilier and Fohlin (2007a, b) on the German case.

[22] See Fohlin (1997, 1999, 2007b) on Germany and Italy. See Van Overfelt et al. (2009) on Belgium.

geared toward industrial finance held board positions and possibly proxy votes, but few held long-term equity stakes. Thus, we can surmise that most industrializing economies practiced a relatively high degree of relationship banking by the early twentieth century.

Notably, Fohlin (2012) finds that universal banking existed without widespread and comprehensive relationship banking (at least 9 of the 26 historical cases of universal banking examined), suggesting that universal banks do not require formal banking relationships to remain viable. This institutional independence is important, because some have hypothesized that formal institutions help enforce repeated interaction between individual firms and a single bank – the German "house-banking" idea – that in turn yields informational economies of scope.[23] In many cases, firms developed relationships with multiple banks, particularly if the firm was large enough to require substantial securities issues, and therefore underwriting or lending from a consortium of banks. Thus, historical evidence also suggests that firms do not always engage in exclusive, long-term banking relationships.

Moreover, banks in "specialized" systems also formalize and maintain relationships through some combination of equity stakes, proxy voting, or sitting on the board of the client firm. Of the primarily specialized systems identified in Fohlin (2012), bankers took up board positions in Canada, Finland, Greece, Japan, the United States, and also in financial systems that had become specialized (Belgium, France, and Italy) during the regulatory initiatives of the interwar years. England was home to apparently the least engaged bankers. However, even there, a new study estimates half of the members of the parliament held seats on corporate boards.[24]

Among the hybrid banking systems (neither truly universal nor specialized), the United States stands out. J. P. Morgan and George F. Baker (respectively, the preeminent investment banker and the chairman of the board of First National Bank of New York) and other investment and commercial bankers played such a high-profile role in US industrial firms in the pre-WWI era that Congress undertook an investigation into the so-called Money Trust through extensive hearings in 1912 and 1913 and passed the Clayton Antitrust Act in 1914.[25] For the majority of the twentieth century, legal restrictions, such as stipulations on equity stakeholding or board memberships, have hindered, but not eliminated, the development of close and formal relationships between US banks and their clients. In a study of more recent times, US bankers sat on the boards of one third of large firms.[26]

It is also worth noting that the United States pioneered the development of intensive "relationship banking" for new firms in the form of post-WWII venture

[23]See Calomiris (1995) for a review of these and related arguments.

[24]Braggion and Moore (2013).

[25]See American Bar Association (1984) on the Clayton Act provisions regarding interlocking directorates.

[26]Kroszner and Strahan (1999). G. William Domhoff, a sociologist at UC Santa Cruz, maintains a website that provides extensive information on interlocking directorates in the United States: http://www2.ucsc.edu/whorulesamerica/power/corporate_community.html.

capital organizations. Venture capitalists fund predominantly untested projects for which the market has yet to enter the picture, and therefore asymmetric information problems may stand in the way of financing externally. Indeed, venture capital financing is most viable for firms with a high chance of ultimately going public and accessing market-based finance. In other words, financing needs vary by stages of individual firm development and may necessitate varying levels of relationship banking over time.

Bank Versus Market Orientation

While it is exceedingly difficult to gather accurate and comprehensive historical measures of securities market activity, the data that are available for a few countries along with qualitative evidence from historical studies indicate that virtually all industrializing economies supported thriving secondary markets for securities before WWI. Later developing countries supported markets as well: stock markets appeared in Istanbul, Madrid, Belgrade, Athens, and elsewhere. Even some of the poorest economies, such as India, Russia, and Brazil, had one or more relatively active financial markets.[27] Only a few countries – Finland, New Zealand, and Norway, for example – lacked significant capital markets. Thus, the evidence so far available indicates that financial markets emerged regardless of banking design. The list of true bank-based systems might dwindle down to nothing. Even Japan is not viewed as an entirely bank-based system, but a hybrid of bank- and market-based systems plus the addition of the zaibatsu (before WWII) as an extra complexity.[28]

In some cases, governments intervened in markets, usually in response to crises. In the archetypal universal banking system of Germany, the government intervened in financial markets and institutions, including requirements on stock market listing, levying of taxes on issues and trades, and imposition and removal of a ban on futures, trading on nearly all industrial shares. The government also created among the most advanced accounting, reporting, and corporate governance standards. One tax law did seem to temporarily shift trading activity from markets to large banks: a tax loophole that failed to impose trading taxes on all orders, even those executed through banks, allowed Berlin-based universal banks to offer savings to their customers who traded through them instead of through smaller intermediaries or brokers. The more trades the banks could gather and net out within their own client networks, the further the eventual net trading fees were spread. This loophole was closed by 1900, but even before that, it did not prevent the expansion of the Berlin exchange. This example, however, may say more

[27]On Brazil, see Mussachio (2009). For a general examination of stock market development, see Michie (2006). See Battilossi and Morys (2011) for a brief survey of markets in Madrid, Vienna, Belgrade, Bucharest, Sofia, Athens, and Istanbul.

[28]Dietl (1998) and Hoshi and Kashyap (2004). See Morck and Nakamura (2005) for an exhaustive treatment; they explain the (substantial) differences between the modern (post-WWII) keiretsu and the prewar zaibatsu.

about the idiosyncratic influences of the government than the innate substitutability of financial markets and universal banks.[29]

The German experience suggests that universal banking became useful and successful because financial markets existed in which to trade securities. Germany was home to several active securities markets, with thousands of share companies listed.[30] In 1905, approximately 30% of the 5,500 German *Aktiengesellschaften* (joint-stock companies) maintained listings on one or more German exchanges – with the majority of these listings in Berlin. Listings grew rapidly after WWI into the 1920s.

It is worth noting the element of path dependency and idiosyncratic development in market development. The first countries to develop liquid securities markets could draw foreign firms to list securities with them, reducing the role of national securities markets in other European or North American countries. Countries that led the prewar international monetary system, such as Great Britain, France, the United States, and Germany, also took the leading role in international financial markets of the late nineteenth and early twentieth centuries. So, London, Paris, New York, and Berlin topped the list of financial markets around the turn of the twentieth century, regardless of differences among their banking organizations.

Bank Branching Versus Unit Banking

One additional characteristic of banking systems that falls somewhat outside of the three dichotomies of financial system design is the question of bank branching and whether it relates to the size and structure of banks. The survey of banking systems conducted in Fohlin (2012) indicates that extensive, national branch networks emerged in most industrialized economies around the world by the early twentieth century. Only Portugal, Denmark, Norway, and the United States failed to develop widespread branching before WWI. The study also finds that the reasons for a lack of branching are not entirely clear: while the United States imposed a variety of restrictions on branching, even in states with no anti-branching law (notably, California), branching developed gradually over the 1910s and after. Likewise, Portugal, Denmark, and Norway did not prohibit branching. Their lack of branching might be attributed to lack of economic development, except that many far poorer countries, such as India, Brazil, Mexico, and Japan, did maintain branch networks.[31] Moreover, although these three non-branching countries were on the European periphery, so were several branching countries: Spain, Russia, Finland, and Sweden, for example. Finally, even though these three countries were small and had small industrial sectors, so were New Zealand, Finland, Ireland, and Greece. In any case, by the early post-WWII years, only the United States perpetuated the unit banking system

[29]See Fohlin (2000).

[30]Fohlin (2007a, b).

[31]Apparently, Brazil imposed restrictions on interstate branching by domestic banks but permitted branching within states. Foreign banks could branch as they pleased.

in many parts of the country – but even then branching within states was taking hold in several states, to the degree it was permitted.[32]

In other words, the available literature indicates that branching appears in all types of financial systems and is neither necessary nor sufficient for universal banking to arise. As the previous discussion explains, universality arose in most places in the middle of the nineteenth century, and branching followed in most places decades later, when the level of development encouraged larger-scale banking. Fohlin (2012) points to two cases that illustrate the point: on the one hand, Germany developed joint-stock universal banking by 1848 but, like most other countries, created widespread branch networks only in the 1890s; England, on the other hand, maintained specialized deposit and investment banking even throughout most of the twentieth century, but developed an extensive nationwide branching system even earlier than the universal banking countries. The literature suggests that despite some modern theoretical arguments, universality of banking services historically required a very modest minimum scale of operations.[33] Thus, while bank branching surely affects market structure in banking and may impinge on the stability of the commercial banking sector, it does not link intimately with overall financial system design – such as the activity of financial markets or the structure of banking institutions.

Financial System Evolution Over the Twentieth Century

The tendency to identify universal-style banking with bank domination and specialized banking with market domination stems from the focus on the post-WWII era, as well as from the narrow range of cases examined. The typology is usually based on comparisons of the United States, Great Britain, Germany, and sometimes Japan in the 1950s through 1980s. The first two countries, having hosted the most important international financial markets for much of the twentieth century and having eschewed both universal banking and formalized bank relationships for most of that time (particularly in the United States postwar), head up the market-based, specialized, arm's-length group. Germany and Japan, with their enormous banks and widely discussed networks of clients and house-bank relationships, lead the bank-dominated, universal, relational group.

After WWII, Austria, Germany, Greece (to some extent – there is no data for proxy voting), Japan (also no data on proxy voting), the Netherlands, Portugal, Spain, and Switzerland all maintained some degree of relationship banking practices.

[32]See Calomiris (2000) for a collection of his previous articles dealing largely with branching and relevant political and regulatory debates. See Kroszner and Strahan (2014) for a study of US banking regulation mostly since the 1930s.

[33]See Benston (1994) for a survey of some literature on banking economies of scale and scope in postwar times and Fohlin (2006) for a historical comparison of banking scale in the United States, the United Kingdom, and Germany. It is important to keep in mind the times in which authors analyze banking scale and scope, because they are influenced by the contemporary macro-financial context (post-WWII boom versus later stagnation) and policy debates (e.g., regulatory tightening since the Great Recession versus deregulation in the 1990s and early 2000s).

In the late twentieth century, Italy, France, and Finland also developed relationship banking. At the same time, these practices became restricted in Japan. Most countries whose banks held seats on company boards allowed them to have equity share holdings in nonfinancial firms. On the whole, these two characteristics of relationship banking did appear to go together, but the extent of long-term stakeholding varied a great deal. When equity stakes coincided with board representation, the motivation was simple to understand: through board seats and equity stakes, banks could provide corporate oversight and simultaneously manage their investments.

The data on proxy voting is sufficiently patchy to make observations of broad patterns virtually impossible. In Germany, however, the data and qualitative evidence on proxy voting (testimony from contemporary observers) suggest that throughout most of the twentieth century, banks held significant control over corporate governance via proxy voting.[34] It is worth noting that US regulation prevented banks from holding equity in companies to which they provided financing – an arm's-length relationship, as discussed earlier.

Even these cases, however, defy rigid classification, since closer scrutiny has revealed a number of contrary facts: for example, a lack of widespread, exclusive house-bank relations in Germany, the unraveling of interlocking directorates and unwinding of equity stakes in Germany at the end of the twentieth century, the frequent appearance of bankers on American boards of directors (approximately one third of large US firms have at least one bank representative on their boards), the lack of universality in post-WWII Japan, and the large size and high level of activity of the securities market in Japan.

Moreover, many systems underwent significant upheaval in the aftermath of the two world wars, so that some systems changed significantly during the interwar and early postwar years. Banking institutions in a number of countries suffered both political and economic consequences of war and depression. Many countries enacted legislation in response to political pressure in the 1920s–1930s, and countries such as Belgium, Greece, Italy, Japan, and the United States went so far as to legally prohibit full-scale universal banking. At the same time, economic and political crises hit financial markets, particularly in the early 1930s and during and after WWII. Rajan and Zingales (1999) suggest that governments, because they could exert less control over markets than over firms, and because of the growing discontent of their constituents, found ways to effectively hinder or even shut down markets of all sorts. These authors argue further that the extent of the anti-market backlash varied most significantly with

[34]Fohlin (2005) surveys long-run patterns of corporate governance in Germany, and Fohlin (2007a, b) proposes the hypothesis that proxy voting by banks related closely to the listing of corporate equity on stock exchanges and the depositing of these shares by shareholders. For an early analysis, see Passow (1922). Franks et al. (2006) attempted to measure proxy voting based on shareholder lists from new issue offerings that were required to publish a register of all shareholders present at the preceding general meeting. By this measure, proxy votes cast by banks increased from 13.3% to 41.8%.

the legal-political system, civil law countries being more susceptible to centralizing command and control than common law countries.[35]

Germany presents, again, one of the most striking examples. The fallout after WWII included the cession of vast portions of eastern German industry and resources, along with the very site of the primary stock exchange (and important provincial exchanges), and the near obliteration of the vibrant Berlin market of the pre- and early post-WWI era. The weight of foreign occupying powers, the urgent bailouts of industrial firms by financial institutions, the strengthening of the social-welfare state, the imposition of hefty capital gains taxes on sales of shares, and other exigencies of postwar reconstruction conspired to produce a financial system in which banks were extremely large, industry partly subordinated its ownership and governance to financial institutions and the government, and markets failed to flourish. Yet, given the country's unique position in the events of the 1930s–1940s, Germany's path differs from the experiences in most other countries – even those with universal banks. Germany's experience therefore does not work as a paradigm case of a universal banking system. Particularly salient is the observation of a reunified Germany at the start of the twenty-first century that has moved away from the archetypal house-banking form, demonstrating that its existence stemmed from the particular needs of postwar Germany.

Elsewhere, the move away from universality varied in its implementation and lasted only a few decades even where it was enforced. By the 1990s, most systems had deregulated and reverted to something resembling their pre-WWI state (see Table 2 and 3). Using the traditional meaning of universal banking – the combination of investment and commercial banking by one institution – banking structure since the 1990s became highly correlated with structure in 1913. For those countries that had begun to industrialize by the mid-nineteenth century, the correlation persists back to at least 1850. Of the 26 cases surveyed, no system clearly and permanently switched from one category to the other over this period of 100–150 years. This evidence of path dependency is all the more impressive in light of government interventions specifically intending to alter institutional design.

Despite much continuity, of course, bank structures, activities, and instruments have evolved over time. Most banking systems, whether universal or "specialized" in the prewar era, underwent a conglomeration movement starting in the 1970s. This development created quasi-universal banking in nearly all industrialized countries, in the sense that financial institutions of several types began operating under the umbrella of bank-holding companies. Thus, even the steadfastly specialized system of England is home to financial services conglomerates. Likewise, the traditionally universal systems of Germany, Belgium, and many other continental European countries have outgrown the centralized universal banking form, so that the commercial and underwriting arms of banks are less closely integrated.

From the research to date, it is clear that attempting to fit particular countries into a few narrowly defined, overarching categories of financial system – for example,

[35]Sylla (2006) offers a critical appraisal of the Rajan and Zingales "great reversals" thesis.

Table 2 Persistence of banking system characteristics over the twentieth century

Country	Universal in 1913? 0–2 (subjective)	Universal in 1990s? 0–2 (subjective)	Universal in 1913? 0–1 (subjective)	Universal in 1990s? 0–1 (subjective)	Bank based in 1990s? 1 = yes	Structure index for 1990s	Development of equity markets in 1913? 0–2 (subjective)
Argentina	1	0	0	0	1	−0.18	1
Australia	0	2	0	1	0	0.80	1
Austria-Hungary	2	2	1	1	1	−1.27	1
Belgium	1	1	1	1	1	−0.17	1
Brazil	2	1	1	1	0	1.01	1
Canada	1	1	0	0	0	0.82	1
Denmark	1	2	1	1	0	0.17	1
England	0	1	0	0	0	1.24	2
Finland	1	1	1	1	1	−0.76	0
France	1	1	1	1	1	−0.17	2
Germany	2	2	1	1	0	0.17	2
Greece	1	1	1	1	1	−0.66	.
India	1	1	0	0	1	0.14	1
Ireland	0	2	0	1	0	0.33	.
Italy	2	2	1	1	1	−0.55	1
Japan	1	0	1	0	0	0.86	1
Mexico	1	1	0	0	0	0.90	1
Netherlands	1	1	1	1	0	0.33	1
New Zealand	0	1	0	1	0	0.49	0
Norway	1	0	0	0	1	−0.23	0
Portugal	2	1	1	1	1	−1.43	1

(*continued*)

Table 2 (continued)

Country	Universal in 1913? 0–2 (subjective)	Universal in 1990s? 0–2 (subjective)	Universal in 1913? 0–1 (subjective)	Universal in 1990s? 0–1 (subjective)	Bank based in 1990s? 1 = yes	Structure index for 1990s	Development of equity markets in 1913? 0–2 (subjective)
Russia	2		1			.	1
Spain	2	2	1	1	1	−0.31	1
Sweden	1	2	1	1	0	0.80	1
Switzerland	1	2	1	1	0	1.58	1
United States	1	0	0	0	0	1.34	2

Sources: Fohlin (2012, Table 6.2). The structure index for the 1990s comes from Levine and Zervos (1998)

Table 3 International comparisons of financial system structure, circa 1990

Country	Securities	Insurance	Real estate	Nonfinancial firms	Stock market cap	Structure index	Market
Argentina	3	2	2	3	0.05	−0.15	0
Australia	1	2	3	2	0.43	0.09	1
Austria	1	2	1	1	0.07	−0.23	0
Belgium	2	2	3	3	0.26	−0.13	0
Brazil	2	2	3	3	0.12	0.03	1
Canada	2	2	2	3	0.46	0.12	1
Switzerland	1	2	1	1	0.71	0.12	1
Germany	1	3	2	1	0.19	−0.14	0
Denmark	1	2	2	2	0.22	−0.08	0
Spain	1	2	3	1	0.18	−0.17	0
Finland	1	3	2	1	0.18	−0.16	0
France	1	2	2	1	0.20	−0.17	0
United Kingdom	1	2	1	1	0.76	0.21	1
Greece	2	3	3	1	0.08	−0.18	0
India	2	4	3	3	0.13	−0.07	0
Ireland	1	4	1	1	0.27	0.15	1
Italy	1	2	3	3	0.12	−0.19	0
Japan	3	4	3	3	0.73	0.06	1
Mexico	2	2	3	4	0.15	0.13	1
Netherlands	1	2	2	1	0.41	−0.04	0
Norway	2	2	2	2	0.15	−0.15	0

(*continued*)

Table 3 (continued)

Country	Securities	Insurance	Real estate	Nonfinancial firms	Stock market cap	Structure index	Market
New Zealand	2	2	2	1	0.40	0.07	1
Portugal	1	2	3	2	0.08	−0.23	0
Sweden	1	2	3	3	0.38	0.07	1
United States	3	3	3	3	0.58	0.17	1

Source: Levine and Zervos (1998)
Note: The variables securities, insurance, real estate, and nonfinancial firms may take values 1–4 as follows:
1. Unrestricted: banks can engage in the full range of the activity directly in the bank
2. Permitted: the full range of those activities can be conducted, but all or some of the activity must be conducted in subsidiaries
3. Restricted: banks can engage in less than full range of those activities, either in the bank or subsidiaries
4. Prohibited: the activity may not be conducted by the bank or subsidiaries
Stock market capitalization is given as a share of GDP. Market equals one if the structure index is positive and zero otherwise. All variables come from Levine and Zervos (1998)

the United States as a specialized banking system or Germany as a universal banking system – can be misleading.[36] Most financial systems have a mixture of characteristics and do not fit neatly into narrow classifications. Many economies undergoing industrialization in the mid- to late nineteenth century supported a small number of large-scale universal banks but simultaneously maintained many more specialized banks. Nationwide branching appeared in most countries between the 1890s and WWI; only the United States persisted with widespread unit banking after WWII, and this is related to regulatory factors. Relationship banking was more common in universal systems but the two institutional features also existed separately from each other. In addition, there has been no link between branching and the design of financial institutions.

The distant history of banking systems reveals that the relationship between universal banking and limited securities markets, to the extent that it exists, is a post-WWII phenomenon. The loss of highly active securities markets is much more persistent than changes in banking design. Among the countries surveyed, no system permanently switched from universal to specialized; banking structure exhibits path dependency, or path reversion, over the past 100–150 years. At the same time, financial conglomerates with fairly distinct functional units have emerged in most industrialized countries. This relatively recent phenomenon appears to be driving the partial convergence of financial system design: formerly "specialized" banks are becoming more universal, while traditional universal banks have become more compartmentalized. Over the past 150 years, banking systems in industrialized countries have become remarkably similar, regardless of their initial development, and many systems have evolved back to their pre-regulation configuration. Almost all countries today have extensive branch networks. And in most economically advanced countries, there are at least some universal banks and some of the attributes typically associated with relationship or house banking, even in systems that would not typically be associated with either institutional form.

What Causes Financial System Differences Historically?

The question of national financial system origins has stimulated much research and debate over the past decade or so. The literature is dense enough to have spawned extended literature reviews of its own. Thus, this section serves to provide a cursory overview and point interested readers to sources for further study.[37]

[36]See Levine and Zervos (1998) and the recent update in Beck et al. (2010) for the World Bank's effort in categorizing financial systems based on legal restraints on financial services. See also Rajan and Zingales (2003) and further discussion later in this chapter.

[37]For much more detail, see Fohlin (2012), Chap. 7, on which this section is based.

Theories: Economics, Law, and Politics

Gerschenkron (1962) offered probably the best known general hypothesis about the genesis of financial institutions, at least concerning industrial banking on the European continent in the nineteenth century.[38] In essence, he argued that banks played a more important role in industrialization for "moderately backward" economies than they had played for the earliest industrializer, Great Britain. Follower economies needed institutions capable of mobilizing a high volume of capital from disparate sources and also that were able to compensate for a shortage of entrepreneurship. In Gerschenkron's view, the German universal banks were just such an institution.

In situations of extreme underdevelopment, as in Russia, however, financial institutions were insufficient to support the transition to modernized industrial activity; such cases demanded centralized institutional intervention, mostly from the government.

In the past 20 years, the so-called "law and finance" literature has turned its attention to legal and regulatory factors that create variation in financial system structure. Government intervention may hamper all development or might promote certain institutions at the cost of others.[39]

Regulation of nonbank institutions – such as securities markets, corporate chartering, limited liability, and bankruptcy – may have further altered the shape of financial systems. For example, laws that protect investors, contracts, and property rights might be argued to encourage the development of all kinds of financial institutions and particularly atomistic market arrangements.[40]

Certain legal systems produce more enabling legislation than do others. Some have argued for the importance of legal traditions in determining the development of financial markets.[41] The modern evidence suggests that countries adhering to a French civil law system have both the weakest investor protection, through both legal rules and law enforcement, and the least developed capital markets. Common law countries fall at the other end of the spectrum, so that American and British

[38]Gerschenkron (1962, 1968, 1970). Sylla (1991) reviews Gerschenkron's theories and related work. Knick Harley (1991) addresses Gerschenkron's idea of "substitution for prerequisites" of industrialization.

[39]The historical literature (such as that spearheaded by Gerschenkron 1962) had always paid due attention to political and regulatory factors. The contemporary study by La Porta (1998), spawned an enormous literature, much of which attempts to reject their fairly simplistic framework, notably Rajan and Zingales (2003).

[40]On Germany, see the edited volume by Horn and Kocka (1979) especially those by Horn, Friedrich, and Reich.

[41]See the series of papers, La Porta et al. (1997, 1998, 1999). In Besley and Persson's (2009) model, if the cost of protecting property rights is lower under common law than under civil law, then common law would allow for more credit as a share of GDP. Pagano and Volpin (2005) make related arguments, discussed subsequently under "Political Factors." Of course, by now, many others have used a similar legal tradition indicator to help explain a number of financial and economic phenomena.

economies or societies have led to market-oriented financial systems. Similarly, Dietl (1998) lays out the poles, admittedly highly stylized, of neoclassical versus relational regulation. These extremes map directly to common law and civil law legal systems, respectively.

La Porta et al. (1998) conclude that countries that provide weak laws for creditor or shareholder protection or weak enforcement of those laws develop substitute mechanisms, such as concentration of ownership, to safeguard owners' rights. Acemoglu and Johnson (2005) argue similarly that individuals adapt their financial intermediation approaches to fit the constraints placed by contracting institutions.

Rajan and Zingales (2003) propose a related theory for the determinants of overall financial system development and specifically contrast legal and political influences. Directed primarily at the La Porta et al. series (1997, 1998), Rajan and Zingales point out that, except for the outlier, Britain, the most developed countries in 1913 maintained similar levels of financial development, regardless of legal system.[42] These authors argue that not legal systems, but political contexts – the support of financial institution growth by the government and interest groups – determine the course of development.

Verdier (1997, 2002) hits on similar themes, but lays out a political-economic view of the development of financial systems. In doing so, he takes direct aim at Gerschenkron's hypothesis about the relationship between the extent of economic backwardness and the role of financial institutions. In this view, political structure, not relative backwardness, determines the shape of financial systems. In particular, universal banking arose in the coincident presence of two conditions: first, a segmented deposit market dominated by nonprofit and provincial banks and, second, a reliable lender of last resort facility insuring liquidity in the banking system. Furthermore, Verdier argues these two preconditions for universality emerged simultaneously only when state centralization was sufficient to provide a strong central bank (with credible lender-of-last-resort status) but limited enough to permit coexistence of provincial and, in his parlance, "center" banks. The issue of legal system does not appear in Verdier's analysis, but the other work reviewed here suggests a possible connection. As Verdier concedes, however, political centralization was neither solitary nor decisive in determining financial structure in most cases. Thus, whether or not Verdier correctly characterizes the relationship between political and financial development, he does not clearly subvert Gerschenkron's hypothesis.

Neither political nor legal structure is clearly independent of economic development, and the three factors may be mutually enhancing, rather than mutually exclusive. For example, Pagano and Volpin (2005) find that proportional voting systems yield less shareholder protection (and greater worker protection) than majoritarian systems and vice versa. These arguments resonate with those in Besley and Persson (2009), who relate similar financial development with legal origins.

[42]On the advanced level of financial development in Britain, Schultz and Weingast (2003) argue that the emergence of liberal democratic political institutions in the seventeenth century prompted a financial revolution that expanded credit availability (government debt at that stage).

Thus, the existing literature leaves room for all three types of factors – economic, political, and legal – in determining the shape of financial development. The formal theoretical models have yet to rationalize endogenous development of distinct financial system designs. Given the variety of theories proposed, assembling a wider range of evidence may shed more light on the issue.

Empirical Evidence

While Gerschenkron's view of financial system development prevailed for several decades, it was rarely put to a rigorous, general test. The first such attempt, by David Good (1973), set out to test that (1) the level of banking development at the end of the so-called great spurt of industrialization or (2) the growth rate of the banking sector during the "great spurt" relates positively to the extent of backwardness at the time of initiation of industrialization. Good's effort underscored the difficulty of clearly specifying Gerschenkron's theory in a testable manner, but he succeeded in raising questions about its generality.

Fohlin (2012) took up the empirical challenge, evaluating economic, legal, and political origins of financial development. Fohlin finds that the economic factors show the greatest power in explaining financial system types and size. In particular, the stage of economic development helps predict the type of banking system that subsequently developed among the pre-WWI industrial nations and also factors into the strength of financial system development. The analysis starts by posing the following test of Gerschenkron: for Europe around 1880, the most and least developed economies should have the lowest rates of financial system growth, while the moderately advanced economies should have the highest rates. Based on the theoretical framework, the level of financial development may be high in the most industrialized economies, but it should certainly be high in the moderately advanced economies and low in the least advanced. Rates of economic growth, in contrast to levels, should yield an essentially linear relationship between economic and financial development: the fastest growing economies should have the most rapid financial development. In the traditional view, slow growers include both those that have passed their earliest phases of industrialization and those that have so far failed to industrialize. Notably, these tests get at financial development generally, as opposed to financial system type.

For the analysis of economic factors, Fohlin (2012) computes GDP per capita growth rates for various subperiods and also constructs a ratio of industrial to agricultural employment and its percentage growth rate from 1880 to 1913. Lastly, Fohlin measures industrial development as the product of GDP per capita and the industrial/agricultural employment ratio, in order to capture the combined effects of wealth and industrial development. The results confirm the hypothesized inverted U-shaped relationship between GDP per capita in 1880 and the level of financial system assets in both 1880 and 1900 (using a robust estimator to mitigate outlier bias). The results for financial development circa 1900 prove much more statistically significant than those for 1880. At the same time, the growth rate of financial assets

relates negatively with the level of GDP per capita in 1880, both from 1880 to 1900 and from 1900 to 1913. The level of GDP per capita in 1900 is also negatively related to financial system asset growth over the succeeding 13 years. Notably, the rate of growth of GDP per capita from 1880 to 1900 relates very strongly and positively to subsequent growth of financial system assets (1900–1913). The reverse relationship – from financial system asset growth to GDP per capita growth – does not appear.

Fohlin also tests the hypothesis that financial structure (both market orientation and universal banking) is related to the level of development and finds that a U-shaped relationship emerges between the structure index reported in Beck et al. (2000) and GDP per capita in both 1880 and 1900. For the most part, in these early industrial economies, market orientation is increasing in the level of development. Similarly, the ratio of industrial to agricultural employment also relates positively to market orientation. At the same time, universal banking was more likely in countries with lower levels of GDP per capita in 1880 and with higher rates of growth of GDP per capita between 1880 and 1900.

On the issue of political factors and financial system type, Fohlin's test analyzes the link between political centralization (a fiscal measure) and both the extent of universal banking at the time of development **as well as** the market orientation index from the late twentieth century. As predicted, state centralization as of 1880 relates negatively and very significantly to market orientation – even 100 years later. In contrast, state centralization cannot be linked statistically to the extent of universal banking. In a related but distinct vein, Fohlin also tests the legal origins theory that the growth (and, implicitly, the design) of financial systems is correlated with legal tradition. In general, markets supersede banks in common law countries. The evidence indicates only weakly that pre-WWI financial development proceeded faster in common law countries, though as expected, full-fledged universal banking only appeared in civil law countries. As Fohlin (2012) points out, the historical pattern may stem from the fact that common law countries are virtually all related to England and adopted English institutions and norms in banking and finance.

Financial Systems and Economic Growth

The principle reason that economists study financial system design is to understand whether the shape of institutions or systems influences the real economy and the welfare of the population. Most studies have focused on the role of finance generally in promoting economic growth, while a smaller literature centers on the varying effects of different systems.

Literature on the Finance-Growth Nexus

Empirical studies on the relationship between long-run growth and financial intermediation show that increased intermediation, or financial development more

broadly, significantly increases growth. Intermediaries presumably lower costs of investment by diversifying idiosyncratic risk and by exploiting economies of scale in information processing and monitoring; they also provide insurance for entrepreneurs, who cannot diversify their risk on their own.[43] Large fixed initial investment costs, R&D for example, can force entrepreneurs to seek external financing; without financial intermediaries, agency problems could make the cost of finance too high, discouraging innovation (and therefore growth). Joseph Schumpeter argued in 1912 that financial intermediaries promote innovative activities, decrease transaction costs, and improve allocative efficiency; in this manner, the financial sector becomes the "engine of growth." Without intermediaries, the cost of R&D projects would be prohibitively high. Financial intermediation also lowers the required rate of return on innovation by lowering fixed costs, thereby spurring growth through investment in R&D. The financial crisis of 2008 prompted a new look at the connection between financial development and growth, as in Beck (2012), and a greater concern for the impact of financial fragility – episodic crises – on economic activity.

In a range of cross-country empirical studies of the postwar era, financial development appears to help predict growth rates.[44] Historical studies, though a bit sparse, show a strong positive effect of financial intermediation in the pre-depression period as well.[45] In one such study, however, finance loses much of its explanatory power for growth when legal origin appears in the regression.[46] Yet none of the legal-origin factors are statistically significant, suggesting that if legal origin matters for growth, it does so through financial development. Moreover, political variables (proportional representation election systems, frequent elections, infrequent revolutions) correlate with larger financial sectors and higher conditional rates of economic growth. Caveats do apply: for example, small countries may import capital, so that for them, domestic financial intermediation sectors may not serve the same purpose as they do in large, diverse countries. Moreover, the link between finance and growth seems to differ depending on a country's level of development, appearing most significant in modern periods for countries at earlier stages in economic development. Countries that had already attained moderately high levels of GDP per capita in 1900 – but not necessarily

[43]These propositions surely seem almost preposterous in light of the crisis of 1907–1909 (and the financial crisis of 2007–2009). The severe drop in economic growth following the loss of liquidity and the general malfunctioning in the financial sector actually underscores the key part that a properly functioning financial system plays in permitting economic growth. Gaytan and Ranciere (2006) develop an overlapping generations model that incorporates liquidity crises and demonstrates a variable relationship between financial development and growth.

[44]King and Levine (1993) and Levine and Zervos (1998), for example. The cross-country growth literature does struggle with identification and other econometric problems. See Manning (2003) for some discussion.

[45]Rousseau and Sylla (2003) do a similar exercise as King and Levine for 17 countries from 1850 to 1997.

[46]Bordo and Rousseau (2006).

the richest ones – grew fastest in the years leading up to WWI.[47] The wealthiest countries in 1880 produced among the slowest growth of financial institution assets between 1900 and 1913, relative to GNP, arguably because they were already well along the path to industrialization by that time.

Time series analyses offer an alternative approach to evaluating the growth impact of financial development. While these methods improve the causal inference possible, the range of studies so far provides mixed answers to the question. Again, the differences among countries stand out, and for contemporary developing economies, Demirgüç-Kunt (2012) emphasizes the key role of government policy.[48]

Financial System "Types" and Long-Run Growth Patterns

Economists and other observers have hypothesized that the distinction between bank-based and market-based financial systems relates systematically to patterns of national economic growth.

For most of the post-WWII era, economists studying financial system design generally argued that financial systems based on banks engaged in relationship banking promoted effective corporate control, long-run perspectives on investment, and sustained economic growth.[49] This assumption stems from the view that banks play a positive role as intermediaries in collecting and disseminating information, in managing risks of various dimensions, and in mobilizing large amounts of capital quickly. By playing this regulatory and information sorting role, banks arguably enhance investment efficiency and thereby economic growth (Allen and Gale 1999), improve capital allocation and corporate governance (Diamond 1984; Gerschenkron 1962), and mitigate the effects of moral hazard (Boot and Thakor 1997). In this view, the long-run relationships that banks form with their clients enable them to smooth the flow of investments and reduce transaction costs and asymmetric information distortions. More recent work, based on the US deregulation experience, attributes firm-level efficiency gains to universal banking (Neuhann and Saidi 2014).

Still, analyses of long-run patterns of development have argued that markets may also enhance growth because they increase incentives to acquire and profit from information about firm performance; under market-based systems, managerial compensation may be more easily tied to firm performance and markets may reduce inefficiencies associated with bank control.[50] In general, the relative strength of banks versus capital markets, however, seems not to affect the overall availability of external finance, though it does relate to the composition of financing between short

[47]Fohlin (2012).

[48]Beck (2013) also surveys the literature on financial development and growth, with a focus on government policy.

[49]See Levine (2002) for a summary.

[50]See Levine (2002).

and long maturities. In less economically advanced countries, it appears that bank finance is particularly important for economic growth.[51]

Historical analysis indicates that neither financial system types – bank based versus market based, branching versus unit, and universal versus specialized – nor legal traditions in themselves can explain the different experiences across countries over the last 100 years or more (Fohlin 2012). That study, encompassing all countries with pre-WWI data available, shows that the wealthier countries among those that began industrialization before WWI tended to deepen their financial base more than the less well-off. In other words, financial and real development went hand in hand in that period of rapid industrial growth. Overall, the set of relatively developed economies at the end of the nineteenth century experienced remarkably similar long-run growth rates, even though they displayed different financial system types, rates of financial development, and legal orientation for most of the twentieth century. The wide range of historical evidence leads to the conclusion that the specific type of financial system or institutions that develop is far less important for economic growth than the development of *some* well-functioning financial system.

Conclusion

The literature on financial system design and development, particularly historical studies of financial institutions and systems, provides a vast array of evidence on how and why institutions take shape and what impact they have on the real economy. The body of research shows the complexity of financial systems among the industrialized economies of the nineteenth and early twentieth century and the range of institutions and markets available to individuals, businesses, and governments. These studies also demonstrate the variety of organization and design of these systems, all focused on similar functions and ultimately on mobilizing enormous amounts of capital toward productive ends.

The research has also shown that the strict dichotomy between market-based and bank-dominated systems does not capture historical or contemporary reality. History offers interesting insights into the multiplicity of financial system designs and the lack of tight links among various banking characteristics, suggesting that going forward, researchers should consider financial systems as an amalgamation of a set of functions rather than as a fixed typology of institutions. The split between universal and specialized banking is most relevant and pronounced in the historical period, before the conglomeration movement of recent years.

Moreover, the research to date strongly supports the idea of persistence and path dependency in financial system design that economic conditions at the time of

[51]See Kpodar and Singh (2011) **as well as** the World Bank's (2013) report on Financing for Development Post-2015: http://www.worldbank.org/content/dam/Worldbank/document/Poverty%20documents/WB-PREM%20financing-for-development-pub-10-11-13web.pdf.

industrialization help set the initial conditions that shape financial system and banking institution design and that historical political conditions, such as centralization of power, plays an ancillary role via the extent of regulation on banks and the development of free capital markets. In other words, history matters.

References

Acemoglu D, Johnson S (2005) Unbundling institutions. J Polit Econ 113(5):949–995

Allen F, Gale D (1999) Comparing financial systems. MIT Press, Cambridge, MA

Allen F, Capie F, Fohlin C, Miyajima H, Sylla R, Yafeh Y, Wood G (2010) How important historically were financial systems for growth in the U.K., U.S., Germany, and Japan? http://ssrn.com/abstract=1701274. Accessed 25 Oct 2010

American Bar Association Antitrust Section (1984) Interlocking directorates under Section 8 of the Clayton Act. Monograph 10, vol 15, issue 5, ABA Press, Chicago

Amidei F, Giordano C (2010) Regulatory responses to the 'roots of all evil': the re-shaping of the bank-industry-financial market interlock in the U.S. Glass-Steagall and the Italian 1936 banking acts. In: Gormez Y, Pamuk S, Turan MI (eds) Monetary policy during economic crises: a comparative and historical perspective, Bank of Italy, Rome

Battilossi S, Morys M (2011) Emerging stock markets in historical perspective: a research agenda. CHERRY discussion paper series CHERRY DP 11/03

Beck T (2012) The role of finance in economic development: benefits, risks, and politics. In: Mueller DC (ed) The Oxford handbook of capitalism. Oxford University Press, New York

Beck T (2013) Finance, growth and fragility: the role of government. Int J Bank Account Finance 5(1):49–77

Beck T, Demirgüç-Kunt A, Levine R (2000) A new database on financial development and structure. World Bank Econ Rev 14:597–605

Beck T, Demirgüç-Kunt A, Levine R (2010) Financial institutions and markets across countries and over time: the updated financial development and structure database. World Bank Econ Rev 24 (1):77–92

Benston GJ (1994) Universal banking. J Econ Perspect 8:121–143

Berger AN, Molyneux P, Wilson JOS (eds) (2014) The Oxford handbook of banking, 2nd edn. Oxford University Press, New York

Besley T, Persson T (2009) Repression or civil war? Am Econ Rev 99(2):292–297

Bhide A (1993) The hidden costs of stock market liquidity. J Financ Econ 34:31–51

Boot AWA, Thakor AV (1997) Financial system architecture. Rev Financ Stud 10(3):693–733

Bordo M, Rousseau P (2006) Legal-political factors and the historical evolution of the finance-growth link. Eur Rev Econ Hist 10(3):421–444

Braggion F, Moore L (2013) The economic benefits of political connections in late Victorian Britain. J Econ Hist 73(1):142–176

Calomiris C (1995) The costs of rejecting universal banking: American finance in the German mirror, 1870–1914. In: Lamoreaux N, Ra D (eds) Coordination and information. University of Chicago Press, Chicago

Calomiris C (2000) U.S. bank deregulation in historical perspective. Cambridge University Press, New York

Collins M, Baker M (2004) Commercial banks and industrial finance in England and Wales, 1860–1913. Oxford University Press, London

Colvin CL, de Jong A, Fliers PT (2014) Predicting the past: understanding the causes of bank distress in the Netherlands in the 1920s. Working paper

David T, Westerhuis G (eds) (2014) The power of corporate networks: a comparative and historical perspective. Routledge, New York

Demirgüç-Kunt A (2012) Finance and economic development: the role of government. In: Berger Allen N, Molyneux P, Wilson JOS (eds) The Oxford handbook of banking. Oxford Press, New York
Diamond D (1984) Financial intermediation and delegated monitoring. Rev Econ Stud 51(3):393–414
Dietl H (1998) Capital markets and corporate governance in Japan, Germany and the United States: organizational response to market inefficiencies. Routledge, New York
Domhoff GW. http://www2.ucsc.edu/whorulesamerica/power/corporate_community.html, accessed December 2014.
Elsas R, Krahnen JP (2003) Universal banks and relationships with firms. In: Krahnen JP, Schmidt R (eds) The German financial system. Oxford University Press, New York, pp 197–232
Ferreira MA, Matos P (2012) Universal banks and corporate control: evidence from the global syndicated loan market. Rev Financ Stud 25(9):2703–2744
Fohlin C (1997) Universal banking networks in pre-war Germany: new evidence from company financial data. Res Econ 51(3):201–225
Fohlin C (1999) The rise of interlocking directorates in Imperial Germany. Econ Hist Rev LII (2):307–333
Fohlin C (2000) Economic, political, and legal factors in financial system development: international patterns in historical perspective. Social science working paper no 1089, California Institute of Technology
Fohlin C (2005) The history of corporate ownership and control in Germany. In: Morck R (ed) A history of corporate governance around the world: family business groups to professional managers, NBER series. University of Chicago Press, Chicago, pp 223–277
Fohlin C (2006) Banking industry structure, competition, and performance: does universality matter?" Social science working paper no 1078, California Institute of Technology
Fohlin C (2007a) Finance capitalism and Germany's rise to industrial power. Cambridge University Press, New York
Fohlin C (2007b) Does civil law tradition (or universal banking) crowd out securities markets? Pre-World War I Germany as counter-example. Enterp Soc 8(2007):602–641
Fohlin C (2012) Mobilizing money: how the world's richest nations financed industrial growth. Cambridge University Press, New York
Franks J, Mayer C, Wagner J (2006) The origins of German corporation – finance ownership and control. Rev Finance 10(4):537–585
Gaytan A,Ranciere R (2006) Banks, liquidity crises and economic growth. Unpublished working paper. http://www.romainranciere.com/research/banks.pdf. Accessed 16 Dec 2014
Gerschenkron A (1955) Notes on the rate of industrial growth in Italy 1861–1913. J Econ Hist XIV:473–499
Gerschenkron A (1962) Economic backwardness in historical perspective. Harvard University Press, Cambridge, MA
Gerschenkron A (1968) The modernisation of entrepreneurship. In: Continuity in history and other essays. Belknap Press of Harvard University Press, Cambridge
Gerschenkron A (1970) Europe in the Russian mirror: four lectures in economic history. Cambridge University Press, New York
Goldsmith R (1969) Financial structure and development. Yale University Press, New Haven
Good D (1973) Backwardness and the role of banking in nineteenth-century European industrialization. J Econ Hist 33:845–850
Harley CK (1991) Substitution for prerequisites: endogenous institutions and comparative economic history. In: Sylla R, Toniolo G (eds) Patterns of European industrialization. Routledge, London/New York, pp 29–44
Hellmann T, Lindsey L, Puri M (2007) Building relationships early: banks in venture capital. Rev Financ Stud 21(2):513–541 (2008)
Hochberg Y, Ljungqvist A, Lu Y (2007) Whom you know matters: venture capital networks and investment performance. J Finance LXII(1):251–301

Horn N, Kocka J (eds) (1979) Recht und Entwicklung der Großunternehmen im 19. und frühen 20. Jahrhundert. Vandenhoeck & Ruprecht, Göttingen
Hoshi T, Kashyap AK (2004) Japan's financial crisis and economic stagnation. J Econ Perspect 18:3–26
Ivashina V, Kovner A (2011) The private equity advantage: leveraged buyout firms and relationship banking. Rev Financ Stud 24(7):2462–2498
Jeidels O (1905) Das Verhältnis der Deutschen Großbanken zur Industrie. Duncker und Humblot, Leipzig
Kauffman Foundation (2013) An overview of the Kauffman firm survey. http://www.kauffman.org/~/media/kauffman_org/research%20reports%20and%20covers/2013/06/kauffmanfirmsurvey2013.pdf. Accessed 17 Dec 2014
King RG, Levine R (1993) Finance and growth: Schumpeter might be right. Q J Econ 108:717–737
Kpodar K, Singh RJ (2011) Does financial structure matter for poverty? Evidence from developing countries. World Bank policy research working paper series, World Bank, Washington, DC
Kroszner R, Strahan PE (1999) Bankers on boards: monitoring, conflicts of interest, and lender liability. NBER working paper, Cambridge
Kroszner RS, Strahan PE (2014) Regulation and deregulation of the U.S. banking industry: causes, consequences and implications for the future. pp 485–543
La Porta R, Lopez-De-Silanes F, Shleifer A, Vishny RW (1997) Legal determinants of external finance. J Finance 52:1131–1150
La Porta R, Lopez-De-Silanes F, Shleifer A, Vishny RW (1998) Law and finance. J Polit Econ 106:1113–1155
La Porta R, Lopez-De-Silanes F, Shleifer A (1999) Corporate ownership around the world. J Finance 54:471–517
Levine R (2002) Bank-based or market-based financial systems: which is better? J Financ Intermed 11:398–428
Levine R, Zervos S (1998) Stock markets, banks, and economic growth. Am Econ Rev 88:537–558
Manning M (2003) Finance causes growth: can we be so sure? Contrib Macroecon 3(1):1100
Merton RC (1993) Operations and regulation in financial intermediation, a functional perspective. In: Englund P (ed) Operation and regulation of financial markets. The Economic Council, Stockholm
Merton RC, Bodie Z (1995) A conceptual framework for analyzing the financial environment. In: Crane D et al (eds) The global financial system: a functional perspective. Harvard Business School Press, Boston
Michie R (2006) The global securities market: a history. Oxford University Press, New York
Morck R (ed) (2005) A history of corporate governance around the world: business groups to professional managers, NBER series. University of Chicago Press, Chicago
Morck R, Nakamura M (2005) A frog in a well knows nothing of the ocean: a history of corporate ownership in Japan. In: Morck R (ed) A history of corporate governance around the world: family business groups to professional managers, NBER series. University of Chicago Press, Chicago, pp 367–459
Musacchio A (2009) Experiments in financial democracy: corporate governance and financial development in Brazil, 1882–1950. Cambridge University Press, New York
Neuhann D, Saidi F (2014) The firm-level real effects of bank-scope deregulation: evidence from the rise of universal banking. Available at SSRN: http://ssrn.com/abstract=2468269 or http://dx.doi.org/10.2139/ssrn.2468269
Pagano M, Volpin P (2005) The political economy of corporate governance. Am Econ Rev 95:1005–1030
Passow R (1922) Die Aktiengesellschaft. EineWirtschaftswissenschaftliche Studie. G. Fischer, Jena
Paulet E (2002) The role of banks in monitoring firms: the case of the credit mobilier. Routledge, New York
Rajan RG, Zingales L (1999) The politics of financial development. Working paper, University of Chicago and NBER

Rajan RG, Zingales L (2003) The great reversals: the politics of financial development in the twentieth century. J Financ Econ 69:5–50

Riesser J (1910) Die Deutschen Großbanken und ihre Konzentration. Verlag von Gustav Fischer, Jena. English translation: The German Great Banks and their Concentration. Published by The National Monetary Commission. Government Printing Office, Washington, DC, 1911

Rousseau P, Sylla R (2003) Financial systems, economic growth, and globalization. In: Michael D. Bordo, Alan M. Taylor and Jeffrey G. Williamson (eds) Globalization in historical perspective. University of Chicago Press, Chicago, pp 373–416

Santikian L (2014) The ties that bind: bank relationships and small business lending. J Financ Intermed 23(2):177–213

Schultz KA, Weingast B (2003) The democratic advantage: institutional foundations. . . Int Organ 57(1):3–42

Sylla RE (1991) The role of banks. In: Sylla R, Toniolo G (eds) Patterns of European industrialization. Routledge, London/New York, pp 45–63

Sylla R (2006) Schumpeter redux: a review of Raghuram G. Rajan and Luigi Zingales's saving capitalism from the capitalists. J Econ Lit XLIV:391–404

Temin P, Voth H-J (2013) Prometheus shackled: Goldsmith banks and England's financial revolution after 1700. Oxford University Press, New York

Van Overfelt W, Annaert J, De Ceuster M, Deloof M (2009) Do universal banks create value? Universal bank affiliation and company performance in Belgium, 1905–1909. Explor Econ Hist 46(2):253–265

Verdier D (1997) The political origins of banking structures. Policy Hist Newsl 2:1–2

Verdier D (2002) Explaining cross-national variations in universal banking in 19th-century Europe, North America and Australasia. In: Forsyth D, Verdier D (eds) The origins of national financial systems: Alexander Gerschenkron reconsidered. Routledge, London, pp 23–42

Whale PB (1930) Joint stock banking in Germany. A study of the German credit banks before and after the war. Macmillan, London

World Bank (2013) Financing for development post-2015. http://www.worldbank.org/content/dam/Worldbank/document/Poverty%20documents/WB-PREM%20financing-for-development-pub-10-11-13web.pdf. Accessed 17 Dec 2014

The Cliometric Study of Financial Panics and Crashes

Matthew Jaremski

Contents

Abstract

Financial crises present an identification challenge. On one hand, declines in economic activity often lead to bank failures, while on the other, bank failures often lead to declines in economic activity. To understand the causes of crises and determine their influence subsequent growth, it is vital to untangle these various factors. Approaches require well-constructed empirical models as well as knowledge of existing data and institutions. Each section of this chapter highlights empirical approaches that have been successfully used to study specific aspects of financial crises. Starting with survival and hazard functions, the chapter goes on to cover data envelopment analysis, vector autoregressions, instrumental variables, and difference-in-difference models.

M. Jaremski (✉)
Colgate University and NBER, New York, USA
e-mail: mjaremski@colgate.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_12

Keywords
Financial panics · Bank failures · Bank regulation

What makes it so vital to understand financial crises also makes them so difficult to study. Imprudent regulations, speculation, agricultural shocks, and declines in economic activity cause financial crises and bank failures, while, at the same time, large-scale bank failures lead to new regulation, financial innovation, and decreased economic activity. Therefore, to understand the causes and effects of financial crises, it is vital to separate the various determinants. Approaches require well-constructed empirical models as well as knowledge of existing data and institutions. Each section of this chapter highlights empirical approaches that have been successfully used to study specific aspects of financial crises. The chapter does not contain an exhaustive description of historical financial panics but rather is intended to be used as a primer for future studies.

Survival Models and Hazard Functions

Bank failures form the heart of nearly all financial panics. While failures are not a sufficient condition, they are a necessary one. Therefore, the majority of financial panic studies address why banks failed in the first place. Authors use a variety of methods to identify the cause of bank failures, but since survival analysis forms the foundation of a large number of papers, it is helpful to start by addressing these models in detail.

Survival analysis attempts to understand what proportion of the population (i.e., banks) will survive past a certain time based on a set of characteristics. Each bank i) is observed for j periods (with $j = 1 \ldots J_i$). The failure time for each bank is then defined as t_{i,J_i}. The cumulative distribution function for the duration T is given by $F(t) = \text{Prob}(T < t)$, with the corresponding density function $f(t) = \frac{dF(t)}{dt}$. The survivor function gives the probability that a bank will survive the period and is defined as $S(t) = 1 - F(t)$. The framework can also be used to study the probability of failure. The hazard function gives the probability of failure within the interval $(t, t + h)$, conditional on the bank surviving to time t, and is defined for period T as

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h \mid T \geq t)}{h} = \frac{\frac{dF(t)}{dt}}{1 - F(t)} = \frac{d(\ln S(t))}{dt}$$

The models observe banks each period, and the effect of the explanatory variables is identified off variation across starting and failure dates.

Researchers usually take one of two approaches when using this framework. First, some model the survival function, examining how long a bank lasted until it failed. These types of models are often called survival or duration models

(e.g., Calomiris and Mason 2003b; Richardson and Troost 2009; Carlson and Mitchener 2009). Second, some studies model the hazard function, examining the instantaneous probability of failure (e.g., Wheelock and Wilson 1995; Jaremski 2010). These types of models are often called hazard models. Both duration and hazard models are discussed in greater detail below, along with examples of how they have been used to study aspects of financial panics.

Branch Banking and Duration Models

Duration models examine the time it takes before bank failure or suspension occurs.[1] In order to implement this approach, a researcher must specify a parametric distribution (e.g., log-logistic, Weibull, etc.). As described by Kiefer (1988), each underlying distribution brings with it a unique survival rate. For instance, the log-logistic distribution (i.e., the most commonly used distribution in the financial panic literature) implies that the probability of failure rises over time and then declines as time goes to infinity. Log-likelihood ratio tests are generally used to choose from the various distributions.

In order to understand the use and approach of duration models, it is helpful to examine a single study as an example. Carlson and Mitchener's (2009) study of the long history of branch banking in California exemplifies the incisive value of using duration models to study financial crises. Branching allows banks to geographically diversify their loan portfolios, potentially making them more likely to survive idiosyncratic economic shocks.[2] Authors such as Calomiris and Gorton (1991) argue that the United States' general lack of branching laws was a factor in many financial panics. Studies at the state and county level generally also support this theory by finding that areas with branching had lower failure rates during financial panics (Mitchener 2005). However, at the same time, studies of individual banks find branch banks are more likely to fail (Carlson 2004). By examining the rise of branch banking in California from its inception in 1909 through the Great Depression, Carlson and Mitchener test (1) whether parent banks acquired weak banks to serve as branches and thus should have been more likely to fail and (2) whether branching had an effect on surrounding banks. Both of these tests are conducted using duration models.

Because most of California's branch banking resulted from acquisitions, the authors start by analyzing where banks were acquired and the types of banks acquired. The location-specific equation is a logistic model that measures the probability of a city having at least one bank acquired between 1922 and 1929. The explanatory variables consist of measures of the city's banking system,

[1] As a result, they produce opposite coefficients from the traditional discrete choice models. A positive coefficient, therefore, implies a negative relationship between the covariate and failure.

[2] Hanes and Rhode (2013) show that financial panics often coincide with exogeneous in cotton harvests.

population, agriculture, and geographic location. As might be expected, acquisitions were more likely to take place in populated areas with greater growth in agricultural income. Acquisitions were also less likely in locations with only one bank or locations close to San Francisco and Los Angeles.

Next the authors address what types of banks were acquired. The bank-specific equations are duration models that take the number of days from June 30, 1922 until a merger as a dependent variable. Lacking income data for non-Federal Reserve members, the authors estimate two separate duration models. The first contains balance sheet information for all banks, while the second contains balance sheet, income, and cost information for Fed banks.[3] The second model adds the return on equity, net losses on assets, and the ratio of administrative costs to total assets but drops many of the other balance sheet variables. The duration models show that banks did not take over weak banks but also did not take over the strongest banks either. Similar to modern mergers and acquisitions, parent banks preferred banks with low net losses and low return on equity. This result is likely due to the fact that parent banks kept the management and structure of the acquired banks and, therefore, wanted to choose a reliable operation that was underperforming. At the same time, the choice of acquisition also depended on preferences of the parent bank. Bank of America (one of California's largest acquirers) sought banks with low net worth and high cash reserves, and other acquirers preferred banks with high net worth and high cash reserves.

Carlson and Mitchener next examine whether the establishment of a branch bank altered the composition of surrounding incumbent banks. They use an Ordinary Least-Squares model (OLS) where the ratio of loans to assets, securities to assets, demand deposits to total deposits, or the growth rate of income-earning assets is on the left-hand side. The sample consists of banks present in 1922 and 1929 to avoid attrition bias. The explanatory variables include the bank and location characteristics that were in the previous models, but the variable of interest is a dummy variable for whether the city gained at least one large branch bank. The data indicate that the entry of a branch bank caused incumbent banks to shift from securities to loans, decrease their administrative costs, and increase their returns on assets. Incumbent banks thus shifted to active portfolios and became more involved in their communities upon facing competition with a large branch bank.

To test whether branching made other banks more stable, the authors examine bank performance during the Great Depression. Returning to a log-logistic duration model, the dependent variable is the number of days from June 30, 1929 until failure and the explanatory variables include the previous bank and location characteristics. The results are quite clear. Having branches in a city increased incumbent banks' number of days before failure (i.e., made them less likely to fail). The duration analysis shows that branch banking pushed unit banks to become more stable.

[3]The authors take a typical approach by including the log of assets to control for size and using balance sheet ratios to control for how portfolio compositions varied across banks.

Putting the results in a broader context, the paper suggests that a further expansion of branching across the nation would have reduced the number of failures and financial panics.

Free Bank Failures and Cox Proportional Hazard Models

Unlike duration models, hazard models examine the probability of failure. The dependent variable is a dummy variable for whether the bank failed between that observation and the following one. These models are similar to binary choice models such as probit and logit models, but they gain further efficiency by taking into account the time before failure.[4] In order to measure the impact of explanatory variables upon the hazard function, researchers typically assume that the explanatory variables act as a scale function on the base hazard rather than adjust the hazard function itself.[5] This is often referred to as a proportional hazard assumption as it allows the probability of failure of bank i given survival to the period t to be written as

$$\lambda(t, X_i, \beta, \lambda_0) = \lim_{h \to 0} \frac{P(t \leq T < t + h \mid T \geq t)}{h} = \lambda_0 g(X_i(t), \beta)$$

where λ_0 is the baseline hazard function common to all banks and the g function captures the effects of the explanatory variables X_i.

The proportional hazard model with time-varying covariates proposed by Cox (1972, 1975) is the most commonly used approach. The approach estimates the effect of the parameters without any consideration of the hazard function. It is thus a semi-parametric "partial likelihood" approach, as it requires the specification of the scale g function (usually an exponential function) but not the baseline hazard function. The drawback, however, is that additional assumptions on the initial hazard function need to be made before calculating the marginal effect of each variable. Thus, while the model can capture the direction of the effect of each variable and compare it to the effect of other variables, it cannot demonstrate the marginal effect on the probability of failure.

As before, it is helpful to discuss the hazard analysis in the context of a single study. Jaremski (2010) uses a Cox model to identify the causes of bank failures during the free banking period (1837–1862). Banks during this period were chartered in two different ways: either through a unique act of the state legislature (called charter banks) or through a general enabling law (called free banks). While both charter and free banks were susceptible to financial panics, free banks exhibited

[4]The addition of dummy variables for the number of years in operation to binary choice models has been used to approximate the same type of relationship.

[5]Alternatively, models can assume that the explanatory variables affect the time directly (often called accelerated time models). Similar to the duration models, however, these models must assume a distribution.

a particularly high susceptibility. Almost a third of all free banks were unable to reimburse note holders for the full value of their bank notes upon closure, compared to under a fifth of charter banks. Prior to Jaremski's study, papers written on the topic focused on two explanations. Free banks either were subject to poorly designed regulation (Rolnick and Weber 1984) or did not sufficiently diversify their asset and liability portfolios (Rockoff 1972). The two hypotheses are straightforward, but testing them separately provided inconclusive results. A comparison of bank failures and collateral bond prices would identify the negative correlation between the two but not prove that properly diversified banks also failed. As most free banking laws were not passed until relatively late in the period, a simple comparison of newly created free banks and old charter banks might also lead to biased estimates.

To account for relationships between explanatory variables, Jaremski's hazard model contains both a bank's financial (cross-sectional) and environmental (time series) information to estimate the roles that nature (bank structure) and nurture (market fluctuations) had in bank failure. Market price fluctuations are captured using the total appreciation or depreciation of a bank's bond portfolio since the bank was in operation, and the undiversified portfolio hypothesis is tested using the corresponding balance sheet ratios. The model also includes a free bank dummy to capture the differential failure rate between the bank types (i.e., the intercept difference in terms of the hazard function itself) and the interaction between the dummy and each explanatory variable to capture how each characteristic affected the different types of banks (i.e., the approximate slope differences).

Free banking's connection between bank notes and state bond prices is the underlying cause of the system's high failure rate relative to the charter banking system. While bond price declines were significantly correlated with free bank failures, they were not correlated with the failure rate of charter banks. Therefore, those banks that did not have to back notes with bonds did not fail because of depressed bond prices. Solvent free banks also diversified their assets away from bonds and their liabilities away from note circulation. Although the addition of balance sheet variables to the hazard model does not reduce the statistical significance of the bond price effect, their combined effect would have been sufficient to at least partially shield banks from bond price declines. After controlling for the time before failure, free banks were not helpless and could have decreased their probability of failure. The results thus show that the financial panics of 1837, 1839, and 1857 are likely tied to the large declines in bond prices that occurred just prior to the bank failures.

Financial Panics and Archival Scraping

A clear example of the need for cliometrics is seen in the attempts to document when financial panics occurred. For instance, if one wishes to understand the causes and effects of financial crises, one must first know when they occurred. Economic downturns and fluctuations often lead to bank failures, but this does not mean that every downturn was caused by a financial panic. Large-scale crises such

as the Great Depression and the Panic of 1907 are easily identified, but smaller panics are much harder to pin down. Many financial panics were also regional in nature, making their identification difficult and even more important for studying output fluctuations.

Researchers have created over nine different US panic series, which vary substantially. Some series document panics occurring roughly once a year, whereas others have panics occurring every 10–20 years. These differences are likely the result of the data being examined and the underlying definition of a financial panic. Unfortunately, the way that each series is calculated is not always clear. For instance, Sprague (1910) highlights periods of monetary stringency during the National Banking Period yet does not describe how he arrived at the dates, whereas studies by Bordo and Wheelock (1998) and Reinhart and Rogoff (2009) use previous studies to define their panic series. It was not until Jalil (2010) that a more consistent framework for identifying panics emerged. Rather than reaching for all the various types of financial panics, he measures banking panics rather than stock market panics or currency panics. This is important as each different type of panic would lead to different outcomes and be measured in different ways. While banking panics could be measured by the number of bank failures, the approach has a number of drawbacks. First, the data simply do not exist for all banks and periods. Weber (2005) has failures before 1861 and the Comptroller of the Currency's Annual Report has failures for national banks after 1863; however, there are no reliable data on the failures of state banks, private banks, savings banks, or trust companies after 1861. As these institutions make up the majority of the banking sector and were most susceptible to panics, their exclusion would dramatically affect the results. Second, bank failures and suspensions do not always result in a banking panic. For instance, 50 banks failing in 50 different states would not generally constitute a panic, whereas 5 banks failing in a large city might signal one. In this way, a panic series needs to examine where bank failures were located, the context, and whether they were correlated.

Jalil identifies banking panics using three large financial newspapers: the *Niles Weekly Register*, *The Merchants' Magazine and Commercial Review*, and *The Commercial and Financial Chronicle*. By consulting the index pages of each newspaper, he searched for terms associated with financial panics (e.g., bank failure, bank suspension, bank run, bank crisis, bank panic, etc.). He then defines banking panics by counting clusters of articles on bank suspensions. Specifically, a cluster is three or more terms with resulting articles that contain a reference to other bank suspensions or reports of a general panic.

The approach avoids scattered, unconnected bank failures and better captures the depth of the panic that the bank failures caused. Moreover, it allows a specific date for when a panic occurred as well as whether it was minor (i.e., local) or major (i.e., national) in scope. Jalil defines a major banking panic as a cluster that (1) spans more than one geographic unit (i.e., a state and its immediately surrounding states) and (2) appears on the front page of the newspaper. All other clusters are labeled as minor panics. Before the Great Depression, he finds evidence of seven major banking panics (November 1833–April 1834, March–May 1837, October 1839,

August–October 1857, September 1873, May–August 1893, October–November 1907) and about 20 minor, geographically specific panics. Using these dates, subsequent studies can better measure the causes and effects of banking panics.

Deposit Insurance, Efficiency, and DEA Analysis

Deposit insurance has often been implemented by legislatures to prevent bank runs. By promising that deposits will be repaid even after a bank's closure, governments hope that individuals will have no incentive to run on the bank. However, because deposit insurance reduces the incentive for depositors to monitor, insured banks might not have a large incentive to act safely. In this way, the legislation might trade idiosyncratic bank runs for larger financial panics. These perverse incentives clearly parallel the recent "too big to fail" discussions.

Testing whether deposit insurance leads to bank failures seems straightforward, but several complications stand in the way. First, most deposit insurance laws apply to all banks in a country. For instance, the FDIC's establishment in 1933 placed deposit insurance on all US commercial banks. This means that there is no variation in which to compare insured banks versus uninsured banks. Second, the creation of deposit insurance might be a response to the risk-taking of banks.

Wheelock and Wilson (1995) get around these issues by making use of a unique law in Kansas and accounting for bank efficiency. In response to the Panic of 1907, Kansas introduced a deposit guaranty system in 1909. Unlike most other systems, membership was optional in response to complaints from conservative banks. As many banks chose not to join the system, the period allows for a comparison of the performances of insured versus uninsured banks under the same regulatory and economic environment. If anything, the odds should be biased in favor of finding a stabilizing effect as insured banks were required to maintain minimum capital ratios and hold reserves with the state banking commissioner. Nevertheless, 94 of the 122 state-chartered banks that failed between 1920 and 1926 were insured banks.

In order to understand whether unstable banks choose to join the insurance system, Wheelock and Wilson measure efficiency using a data envelopment analysis (DEA). The nonparametric approach is a relatively simple idea (i.e., determine how close a bank is operating to the minimum inputs for a given level of output or maximum output for a given level of inputs) but is complicated to implement. Using Shephard (1970), the input and output distance functions are computed by solving the linear programs:

$$\left(D_i^{\text{in}}\right)^{-1} = \min\left\{\theta | y_i \leq Yq_i, \theta x_i \geq Xq_i, Iq_i = 1, q_i \in \mathbb{R}_+^N\right\}$$

and

$$\left(D_i^{\text{out}}\right)^{-1} = \max\left\{\theta | x_i \geq Xq_i, \theta y_i \leq Yq_i, Iq_i = 1, q_i \in \mathbb{R}_+^N\right\}$$

where $Y = [y_1 \ldots y_n]$, $X = [x_1 \ldots x_n]$, with x_i and y_i denoting the $(n \times 1)$ and $(m \times 1)$ vectors of observed inputs and outputs for the ith bank $(i = 1,\ldots, N)$, $x_i \in \mathbb{R}_+^n$ and $y_i \in \mathbb{R}_+^n$ for all $j = 1,\ldots, N$, and I is a $(1 \times N)$ vector of ones and q is a $(N \times 1)$ vector of intensity variables which serve to form a piecewise linear approximation of the technology.[6] The values of D_i^{in} and D_i^{out} measure the radial distance from the bank's observed point (x_i, y_i) to the boundary of the convex hull of all observations. In other words, the value of D_i^{out} measures how much output can be increased by holding inputs fixed but moving to the frontier production set, whereas the value of D_i^{in} measures how much the inputs can be decreased by holding output constant but moving to the frontier. While the distance provided by the analysis is in terms of the inputs or outputs, it is often normalized to represent the proportional increase of outputs or decrease of inputs that can be achieved.

The main choice when using a DEA analysis is the selection of outputs and inputs.[7] The most common approach is to look at outputs such as the value of loans, demand deposits, and time deposits and inputs such as labor, capital, and purchased funds. Wheelock and Wilson chose two outputs (loans and bond holdings, and demand deposits) and four inputs (time and savings deposits, borrowed funds, value of bank premises, and number of bank officers). They then calculate the value of D_i^{out} for each bank in each year.

Using biannual data for nearly all of Kansas' state-chartered banks, Wheelock and Wilson implement a Cox proportional hazard function. The model includes the measure of efficiency, a dummy for whether the bank had deposit insurance, total assets, and balance sheet ratios. The results show that efficient banks were the most stable, but being an insured bank does not matter. The authors explain that the insignificance of the insurance dummy is likely due to the suspension of payments in 1925. Once the suspension occurred, depositors lost confidence in the system and insured banks had to adjust their portfolios or face bank runs. Indeed, when the insurance dummy is split into years before and after the suspension, the former is statistically significant and positive whereas the latter is insignificant but negative. The paper suggests that deposit insurance schemes may cause the very problems they were intended to fix even after controlling for efficiency.

Fed Intervention and Difference-in-Difference Models

Friedman and Schwartz (1963) posit that the inaction of the Federal Reserve System was to blame for the Great Depression's continued waves of bank panics and the depth of the economic decline. They argue that instead of pumping liquidity into the system to prop up illiquid but solvent banks, the Fed's tight monetary policy caused the system's series of banking panics. Much like other topics in this chapter, the

[6]This description is taken from Wheelock and Wilson (2000) which also used a DEA.

[7]It is important to note that the model is relatively sensitive to outliers that would push the frontier out too far.

problem with testing whether the Fed could have halted bank failures is that there were many other factors in play and the majority of Federal Reserve District Banks acted the same way.

In an effort to test Freidman and Schwartz's argument, Calomiris and Mason (2003b) examine whether failures reflected fundamental deterioration in bank health or sudden crises of systemic illiquidity. To the extent that the failures can be explained by fundamentals, they argue that any actions taken by Fed or Congress would have had little to no effect.

The authors use a log-logistic survival model to estimate the log days until failure after December 31, 1929. Bringing forth an impressive database on all Federal Reserve member banks, they use a variety of bank-level characteristics (observed biannually), county-level characteristics (observed only in 1930), and state- and national-level characteristics (observed monthly or quarterly).[8] The bank-level variables include the bank's type, size as well as measures of its asset quality, liability mix, and costs. The county-level characteristics include measures of the local economic conditions, whereas the state-level and national-level characteristics include broader economic measures. The model measures waves of illiquidity using dummy variables for the panics that affected all regions (December 1930–January 1931, May–June 1931, September–November 1931, January 1933, February 1993, and March 1933) and the few regional-specific panics identified by Wicker (1996).

Bank fundamentals are significantly correlated with failure risk, but only the panics in January and March of 1933 seem to have induced more failures. The evidence supports the theory that contagion, liquidity crises, and the inaction of the Fed have less of a role in explaining the early part of the Great Depression; however, at the same time, the data do not explicitly measure the effect of Fed intervention. It is also possible that the Fed's inaction allowed bank balance sheets to deteriorate over time.

Building on Calomiris and Mason, Richardson and Troost (2009) take the analysis a step further by targeting the Federal Reserve's actions during the Great Depression. They do this by focusing on Mississippi, one of the few states served by two different Federal Reserve Districts during the Great Depression. The northern section was located in the 8th district (St. Louis) and the southern section was located in the 6th district (Atlanta). Because the Atlanta Fed was one of the few district banks that provided liquidity to member banks, the authors are able to study the effect of bank intervention with a natural control group. This difference-in-difference approach makes use of the idea that all banks in the same state should be subject to the same economic conditions and regulation and should only have differed through the actions of the Fed district banks. The authors primarily restrict the sample to banks within one degree of latitude of the border to ensure this similarity, but they also show the results hold when looking within 50 miles of the border or only in border counties.

[8]Given the varying frequencies, each observation is a bank-month.

Richardson and Troost start by calculating the raw survival function using the Kaplan-Meier Method and the smoothed raw hazard functions separately for each Fed district.[9] The Kaplan-Meier Method is a nonparametric approach to survival analysis which plots the fraction of banks that survive each period after correcting for censoring. The function is

$$S(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

where n_i is the number of banks in business at the beginning of time period t, d is the number of banks experiencing an event at time t, and t_i indicates the ith time period. Alternatively, the smoothed hazard function for period t is

$$g(t_i) = \sum_{z=-u}^{u} K \frac{d_{i+z}}{n_{i+z}}$$

where u is the bandwidth (chosen to be 28 days) and

$$K = \frac{(u+1)^2 - z^2}{\sum_{z=-u}^{u} \left[(u+1)^2 - z^2\right]}$$

The survival functions of the Atlanta banks are much higher than those of St. Louis banks, and the hazard functions are much lower, suggesting that Fed intervention was important for bank stability during the Great Depression.

To the extent that banks in the two districts are exactly the same, the simple comparison of the difference in the failure rates should be a reliable measure of the causal effect of Fed intervention on bank failures. However, Richardson and Troost apply a log-logistic survival model to further control for other factors. The dependent variable is the logarithm of days until bank distress (i.e., liquidation, suspension, or consolidation) from 1929 to March 1933. In addition to the fundamental variables included by Calomiris and Mason, they include a dummy variable for whether the bank was located in the Atlanta District and the interactions between the Atlanta dummy variable and the panic dummies. The Atlanta dummy variable captures whether Atlanta District banks were fundamentally different from St. Louis District banks and the interaction captures the importance of being in the Atlanta District when liquidity was provided. The model shows that banks in the Atlanta District had fewer days until distress (i.e., were more likely to fail) across all periods, but during each of the specific banking panics, the actions of the Atlanta Fed increased the number of days until distress (i.e., were less likely to fail). The actions of the Atlanta

[9]The approach was first developed in Kaplan and Meier (1958).

Federal Reserve thus seem to have had an effect on bank stability, and a more concerted effort across all the districts might have mitigated bank losses during the Great Depression.

The Effect of Bank Failures and Accounting for Endogeneity

So far this chapter has focused on the causes of bank failures and financial panics, but their effect on the economy is often just as important. Based on the work of scholars such as Friedman and Schwartz (1963), Bernanke (1983), and Reinhart and Rogoff (2009), financial crises and bank failures have a dramatic effect on the economy in two ways. First, bank failures cause a sudden decline in the money stock due to the loss of deposits and shareholder equity. This reduces consumer demand and dampens any recovery. Second, bank failures destroy institutional knowledge and increase the cost of credit intermediation. Those institutions that survive and new institutions that arise might not provide the same number of loans and might focus on safer borrowers. Despite these potential effects, studies have cast doubt on their importance. Cole and Ohanian (2000) show that state-level income was already declining before bank failures during the Great Depression. Chari et al. (2002) show that simple neoclassical growth models without nonmonetary effects predict changes in investment quite well.

As with most finance-led growth papers, endogeneity is a major issue and could be driving the contradicting results. Economic downturns are often responsible for bank failures, and bank failures are often responsible for economic downturns. Therefore, in order to view the causal effect of bank failures on economic activity, one must account for this feedback effect. The rest of this section examines three ways that this has been done in the literature.

Vector Autoregression (VAR)

One way to account for the endogeneity is to explicitly model it. Following many macroeconomic studies, Anari et al. (2005) model the effect of bank liquidations on output using a vector autoregression (VAR).[10] The VAR methodology investigates the dynamic interactions between variables without imposing a priori structural restrictions. It involves estimating a separate regression equation for each variable on its own lags and those of the other variables in the system. For instance, a VAR with three variables would be modeled as

[10]Bordo and Landon-Lane (2010) is another good example of the use of VARs to model the causes and effects of financial panics using Great Depression data.

$$x_{1,t} = a_{1,0} + \sum_{i=1}^{k} a_{1,i}x_{1,t-i} + \sum_{i=1}^{k} b_{1,i}x_{2,t-i} + \sum_{i=1}^{k} c_{1,i}x_{3,t-i} + u_{1,t}$$

$$x_{2,t} = b_{2,0} + \sum_{i=1}^{k} a_{2,i}x_{1,t-i} + \sum_{i=1}^{k} b_{2,i}x_{2,t-i} + \sum_{i=1}^{k} c_{2,i}x_{3,t-i} + u_{2,t}$$

$$x_{3,t} = c_{3,0} + \sum_{i=1}^{k} a_{3,i}x_{1,t-i} + \sum_{i=1}^{k} b_{3,i}x_{2,t-i} + \sum_{i=1}^{k} c_{3,i}x_{3,t-i} + u_{3,t}$$

where $x_{i,t}$ is the ith variable in period t. Each variable is thus allowed to affect every other variable. Usually a series of nested likelihood ratio tests are used to optimally select the lag length.[11]

After jointly estimating the coefficients, there are three ways to proceed using a VAR model. First, one can study how an exogenous shock to one variable affects the other variables. This is done by estimating the model's impulse response function. The impulse response function graphs how the time series of each variable would change after a hypothetical but exogenous shock to a single variable. Since the estimated relationship explicitly accounts for endogeneity, the shock provides realistic results of what is expected to happen; it does not estimate, though, exactly what happened. The approach is sensitive to the variable order specified by the author.[12]

Second, one can decompose the variance of the forecast error of the model. The forecast error variance decomposition (FEV) provides the fraction of the squared prediction error explained by each variable. Similar to an impulse response function, the estimate is calculated for each time period so that the size of the effect of each variable can be seen over time.

Third, one can study the effect of a variable on another variable using a Granger causality test. The test examines whether the past values of a variable are significant predictors of another variable. While identifying something more than a correlation, it is not a direct test of causality as variables could Granger cause each other. The validity of the approach also suffers when the underlying time series are not stationary. In general, the distributions of these tests are nonstandard when a VAR contains variables with unit roots, and differencing is usually required to ensure stationarity. Sims et al. (1990), however, show that Granger tests conform to standard distributions in tri-variate VARs with unit roots so long as a single cointegrating relationship exists among the variables.[13]

The authors concentrate on the model's FEV and impulse response functions for a four variable VAR. The four variables are (1) industrial production, (2) wholesale price index, (3) M1, and (4) stock of failed national banks' deposits for credit availability. The sample is estimated monthly from January 1921 through December 1940 in order to provide a large quantity of observations, capture preexisting trends

[11]The authors use the asymptotic chi-square test developed by Sims (1980) for the determination of lag order.

[12]Often variables are ordered more exogenous to least exogenous but this in itself must be based on the authors' opinion.

[13]Cointegration (i.e., the existence of a long-run relationship between the variables) is often tested using the approach suggested by Johansen (1991).

before the Great Depression, and view the entire slow recovery. Before estimating the VAR, they show that most of the variables have unit roots but there are also two conintegrating vectors, allowing them to estimate the model in levels.[14]

The FEV is quite clear. The effect of the deposits of closed banks were small in the short run but grew much larger over time. The value of closed bank deposits explains about 9% of the forecast error by 12 months, 20% by 18 months, and 21% by 24 months. In fact, the pattern of results is nearly equal to the effect of the money supply. On the other hand, the wholesale price index explains a significant proportion of the decline in industrial production regardless of the time period. Looking at the impulse response functions, a shock to closed bank deposits and money supply takes half a year to have a significant effect on production, whereas prices have an immediate effect. The impulse response functions also show that the effect of closed bank deposits does not seem to be permanent, while prices and money supply continue to have an effect several years into the future.

Anari, Kolari, and Mason find that the occurrence of bank liquidations had large effects on output. However, as with any approach, there are a few drawbacks to VAR models. First, the VAR framework cannot control for a large number of other variables. Recently, structural and panel VAR models have emerged to help account for fixed effects and common shocks, but they still are sensitive to the choice of main variables.[15] Second, the models are relatively sensitive to the choice of variables, lags, and the order of variables in the model.

Instrumental Variables (IV)

In a companion piece to their study of bank failures, Calomiris and Mason (2003a) use a two-stage least-squares (2SLS) model to test the effect of deposit and loan decline on the amount of state-level income growth. The approach attempts to isolate the portion of the change in deposits or loans between 1930 and 1932 that is exogenous to income growth over the same period. To do this, the authors find variables that are only correlated with state income growth through the change in deposits or loans and thus can be excluded from the main estimating equation. While only one instrument is needed per endogenous variable, Calomiris and Mason overidentify the equation by using three instrumental variables measured in 1929. The log of bank assets captures the initial amount of banking in the state and state regulation. The ratio of real estate owned relative to loans captures the previous amount of loan foreclosures and the exposure to agricultural loans. The ratio of net worth to total assets captures the buffer that banks had going into the 1930s. Because

[14]In systems with cointegtation, a VAR model with first differences is misspecified. However, Engle and Granger (1987) show that a VAR in levels avoids the problem.

[15]For instance, Kupiec and Ramirez (2013) apply a panel VAR to study financial panics across the various states.

these variables are all observed prior to the Great Depression, the authors argue that they are exogenous and excludable.

The 2SLS approach begins by modeling the endogenous variables as a function of the other explanatory variables and the instruments and then models the change in income as a function of the other explanatory variables and the predicted value of the endogenous variable from the first stage.[16] While standard OLS models are unbiased, 2SLS models do not have the same characteristic, and estimates from small samples could deviate from their target parameters. This is even more so when instruments are "weak" (i.e., do not significantly predict the exogenous variable). The standard practice is to report the F-statistic and test for overidentification in the first stage model. As long as the sample and the F-statistic are relatively large, then the second-stage model's estimates are generally accepted.

In the first stage, the log of bank assets positively predicts deposit and loan growth, while the ratio of real estate to noncash assets negatively predicts loan growth and the ratio of net worth to assets positively predicts deposit growth. The second-stage model shows that the change in deposits and loans is significantly and positively related to state income growth even after controlling for the growth in building permits, production income, and the liabilities of failed businesses.

While the model produces stark results, the authors are concerned over the limited number of observations. They thus augment the analysis by looking at the effect of the change in loans on building permits for 131 major cities. The 2SLS model shows that the instrumented change in deposits over the early 1930s is positively and significantly related to permit growth. The authors thus conclude that banking distress was an important propagator of shocks.

Calomiris and Mason admit a couple limitations. First, the focus on mostly statewide data limits the number of observations and introduces potential measurement error. Second, the database only contains information on state income and city building permits but not information on manufacturing or production. Third, instrumenting with initial conditions could fail the exclusion restrictions when there is serial correlation.

Difference-in-Difference (DD)

In order to avoid the small sample and instrument problems with aggregated data, Ziebarth (2013) uses a difference-in-difference approach similar to Richardson and Troost (2009). Ziebarth collects a plant-level dataset from the Census of Manufactures taken in 1929, 1931, 1933, and 1935. The database contains information on revenue, output, price, number of workers, and hours per worker for manufacturing plants in Mississippi. By comparing plants in the St. Louis Federal Reserve District

[16]While these two equations could be estimated with OLS separately, the standard errors would not be estimated correctly because the second-stage estimates would take into account the fitted values instead of the original endogenous variable.

with those in the Atlanta District, the model compares plants in areas where bank failures occurred more frequently to those in areas where bank failures occurred less frequently. Because the district boundaries only affected banks, all plants should have been subject to the same economic conditions and regulations.

The approach controls for endogeneity, but the data present another potential problem: selection bias due to attrition. In addition to industry-specific time trends, Ziebarth uses a set of variety different specifications to measure and control for this problem. First, he uses an unbalanced panel of all plants and does not explicitly correct for selection. Second, he limits the sample to plants that survived the entire period, effectively eliminating any bias due to banks altering their behavior before exit. Third, he uses time-invariant plant-level fixed effects to at least partially control for the selection bias.[17] The main variable of interest is the interaction between a dummy for plants in the St. Louis District and a dummy variable for the bank panic in 1931, but the dummies are also separately included in the regressions to soak up any time-invariant differences.

The evidence shows that plants in the St. Louis District experienced a dramatic reduction in plant-level revenue, physical output, and hours per worker compared to Atlanta. At the same time, the Difference-in-Difference estimate does not have a consistent effect on price, average wage, or the number of workers. Plants thus responded to bank failures by reducing the number of hours worked and output but not by adjusting their price or firing workers. Extending his results across time, Ziebarth finds that most manufacturing plants bounced back after 1931. By 1933, there were no remaining differences between the two sections of Mississippi even though banks had not returned. In this way, the paper shows that bank failures can lead to large crashes in production, but recovery can still occur even before credit conditions improved.

Conclusion

Regardless of country or time period, financial crises are one of the principal causes of sudden economic change. Crises reduce production, income, and prices and also bring about broader regulatory and institutional changes. As such, it is critically important to understand their causes and effects. The study of historical financial panics, in particular, has become even more vital as authors (e.g., Sprague 1910; Friedman and Schwartz 1963; Wicker 2000; Reinhart and Rogoff 2009) continue to show that panics often occur for the same reasons. For instance, Carlson and Mitchener (2009) argue that had branching been allowed to spread beyond a few isolated states, the Great Depression would have been less severe, whereas Wheelock and Wilson (1995) caution that deposit insurance might allow banks to take too much risk even today. The conclusions of these studies are straightforward, but they

[17]The drawbacks are that the estimate of the effect of the Atlanta Fed's action would only be based on within plant variation and the limited number of observations available to study.

would not have been possible without first solving substantial identification problems. The use of cliometrics, therefore, is more than just a historical application of modern techniques. Rather, it is an attempt to better understand previous problems so that we might not repeat them.

References

Anari A, Kolari J, Mason J (2005) Bank asset liquidation and the propagation of the U.S. great depression. J Money Credit Bank 37:753–773

Bernanke BS (1983) Nonmonetary effects of the financial crisis in the propagation of the great depression. Am Econ Rev 73:257–276

Bordo M, Wheelock DC (1998) Price stability and financial stability: the historical record. Fed Reserve Bank St Louis Rev 80:41–62

Bordo M, Landon-Lane JS (2010) The lessons from the banking panics in the United States in the 1930s for the financial crisis of 2007–2008. NBER working paper no 16365

Calomiris C, Gorton G (1991) The origins of banking panics: models, facts, and bank regulation. In: Glenn Hubbard R (ed) Financial markets and financial crises. University of Chicago Press, Chicago, pp 109–174

Calomiris C, Mason JR (2003a) Consequences of bank distress during the great depression. Am Econ Rev 93:937–947

Calomiris C, Mason JR (2003b) Fundamentals, panics, and bank distress during the depression. Am Econ Rev 93:1615–1646

Carlson M (2004) Are branch banks better survivors? Evidence from the depression era. Econ Inq 42:111–126

Carlson M, Mitchener K (2009) Branch banking as a device for discipline: competition and bank survivorship during the great depression. J Polit Econ 117:165–210

Chari VV, Kehoe P, McGrattan E (2002) Accounting for the great depression. Am Econ Rev 92:22–27

Cole H, Ohanian L (2000) Re-examining the contributions of monetary and banking shocks to the U.S. great depression. In: Bernanke BS, Rogoff K (eds) NBER macroeconomics annual 2000, vol 15. MIT Press, Cambridge, MA, pp 183–227

Cox DR (1972) Regression models and life-tables. J R Stat Soc 34B:187–220

Cox DR (1975) Partial likelihood. Biometrika 62:269–276

Engle R, Granger C (1987) Cointegration and error-correction: representation, estimation, and testing. Econometrica 55:251–276

Friedman M, Schwartz AJ (1963) A monetary history of the United States: 1867–1960. Princeton University Press, Princeton

Hanes C, Rhode P (2013) Harvests and financial crises in gold standard america. J Econ Hist 73:201–246

Jalil A (2010) A new history of banking panics in the United States, 1825–1929: construction and implications. PhD dissertation, University of California-Berkley

Jaremski M (2010) Free bank failures: risky bonds vs. undiversified portfolios. J Money Credit Bank 42:1565–1587

Johansen S (1991) Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models. Econometrica 58:1551–1580

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Kiefer N (1988) Economics duration data and hazard functions. J Econ Lit 26:646–679

Kupiec P, Ramirez C (2013) Bank failures and the cost of systemic risk: evidence from 1900 to 1930. J Financ Intermed 22:285–307

Mitchener K (2005) Bank supervision, regulation, and instability during the great depression. J Econ Hist 65:152–185
Reinhart CM, Rogoff KS (2009) This time is different: eight centuries of financial folly. Princeton University Press, Princeton
Richardson G, Troost W (2009) Monetary intervention mitigated banking panics during the great depression: quasi-experimental evidence from a federal reserve district border, 1929–1933. J Polit Econ 117:1031–1073
Rockoff H (1972) The free banking era: a reexamination. Dissertations in American History, revised PhD dissertation, University of Chicago
Rolnick A, Weber WE (1984) The causes of free bank failures: a detailed examination. J Monet Econ 14:269–291
Shephard RW (1970) Theory of cost and production functions. Princeton University Press, Princeton
Sims CA (1980) Macroeconomics and reality. Econometrica 62:520–552
Sims CA, Stock JH, Watson MW (1990) Inference in time series models with some unit roots. Econometrica 58:113–144
Sprague OMW (1910) History of crises under the National Banking System. National Monetary Commission, S.Doc. 538, 61st Cong., 2d session
Weber WE (2005) Listing of all state banks with beginning and ending dates. Research Department, Federal Reserve Bank of Minneapolis, http://research.mpls.frb.fed.us/research/economists/wewproj.html
Wheelock D, Wilson P (1995) Explaining bank failures: deposit insurance, regulation, and efficiency. Rev Econ Stat 77:689–700
Wheelock D, Wilson P (2000) Why do banks disappear: the determinants of U.S. bank failures and acquisitions. Rev Econ Stat 82:127–138
Wicker E (1996) The banking panics of the great depression. Cambridge University Press, Cambridge
Wicker E (2000) Banking panics of the gilded age. Cambridge University Press, Cambridge
Ziebarth N (2013) Identifying the effect of bank failures from a natural experiment in Mississippi during the great depression. Am Econ J Macroecon 5:81–101

Payment Systems

John A. James

Contents

Abstract

The payments system is the complex of financial instruments and relationships that transfer value between buyers and sellers to complete their transactions. The character and reliability of the payments system, the rules, practices, and institutions by and through which value is transferred from payors to payees, is obviously a crucial underpinning to a market economy. Cash is the simplest means of payment. However, the vast majority of transactions, especially in developed economies, involve non-cash payments instruments. The use of non-cash instruments in the payments process can take time and involve some risk since they are promises to make future payments. Payments system improvements that reduce costs and/or risks should have a salutary effect on the operations of a market economy.

In the absence of the "double coincidence" of wants, some arrangement is necessary to facilitate the exchange of goods. The payment system is the complex of financial instruments and relationships that transfer value (or good funds) between buyers and

J. A. James (✉)
Department of Economics, University of Virginia, Charlottesville, VA, USA
e-mail: jaj8y@virginia.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_13

sellers to complete their transactions. The character and reliability of the payment system, rules, practices, and institutions by and through which value is transferred from payers to payees, is obviously a crucial underpinning to a market economy. By the same token, payment system improvements, which either reduce costs or risks involved in making payments, should have salutary effects. The payment system is in Kahn and Roberds's words the "plumbing of the economy," essential and pervasive, the "glue that binds together the gains from trade" (2009, pp. 1, 19).

Cash is the simplest means of payment. Cash transactions directly transfer good funds (a generally accepted means of exchange) from buyer to seller and so constitute final settlement ("the final and unconditional transfer of the value specified in a payment instruction" which legally or effectively discharges any financial obligation of the payer to the payee (Juncker et al. 1991, p. 847)) without mediation by third parties. The vast majority of transactions however, especially in developed economies, involve noncash payment instruments like checks or credit card receipts. Unlike cash, they represent payment orders directing the transfer of good funds between intermediaries and ultimately each party's transactions account.

The use of noncash instruments in the payment process typically involves several steps: the transmission of the payment order to the buyer's intermediary, its verification and approval, the reverse delivery of funds to the seller's intermediary, and the crediting and debiting of each party's account. Needless to say, noncash payments can take time – several days in the case of checks and at least a month for credit card transactions. Consequently, the payment system is built on flows of credit, or to use somewhat dated terminology, noncash payment instruments are also credit instruments (Kinley 1910). In any case in which payment and settlement do not occur simultaneously, some risk is involved. Payments are generally promises to deliver funds of a certain amount. The recipient of a payment faces some uncertainty about receiving value (in cash or good funds) until settlement has occurred even though a payment has been made to them. The historical development of payment systems is in many respects a story of the evolution of economic institutions. Cliometrics, as we shall see here, can make and has made important contributions to this literature – both quantitatively, in assembling financial data on note prices and exchange rates, for instance, and analytically, notably in applications of network analysis among other modeling.

Coinage, Money Changers, and Deposit Banking

With the demonetization following the collapse of the Roman Empire in the west, both the demand for and supply of coinage declined (by the mid-fifth century and for 200 years thereafter, coins ceased to be used as a medium of exchange in Britain, e.g.). Some Roman coins continued to circulate for centuries in various states of degradation supplemented by the output of some former Roman mints then operated by barbarians (Spufford 1988). Monetary, as well as political, unity came in much of western Europe at the beginning of the ninth century as Charlemagne became Emperor of the Romans and the Carolingian silver penny became the standard of

exchange. Over time the coinage system evolved into one with three tiers – gold coins, large-value silver coins, and smaller-value (initially silver) coins. These different kinds of money served different functions – gold coins used by merchants in nonlocal trade, large silver coins used in large-value local transactions, and smaller-value coins used in retail transactions. In Kuroda's phrase (2008; Fantacci 2008), these were complementary currencies functioning in effect in parallel. Fantacci (2005, 2008) in turn draws out the distinction between the unit of account and medium of exchange in this period and in a most interesting model shows how the relationship between them could have been altered ("mutation") – through changes in metal content (debasement/reinforcement) or through changes in nominal value (abatement/enhancement).

Now back to the Carolingian silver penny – although the standard of exchange, they were necessarily used all that often. With essentially self-sufficient manors becoming the basic unit of economic organization, there was still little need for money both internally within the manor and also externally because there was little trade (other than in a few items such as salt). In the relatively few towns extant the value of the single coin, the silver penny was too large to be used in small payments – the daily wage of twelfth-century English domestic servant was around a penny, for example. As a result of the problems in using cash, many transactions necessarily involved the extension of credit.[1] For example, book credit could have been advanced until the debt could be discharged by a cash transfer. Or the butcher and the baker might have accumulated offsetting mutual obligations which at some point could have been settled through bilateral netting.

The subsequent political fragmentation of Charlemagne's empire was accompanied by a monetary fragmentation as well as local princes and lords began to mint their own pennies and then later other coins. With the revival of distance trade, merchants consequently faced a befuddling array of coins in circulation with varying weights and degrees of fineness as well as differing degrees of clipping and abrasion, plus the fact that they were heavy to carry around (see Kohn 1999c). Money changers then, with their specialized skills and expertise in assessing coinage in circulation, became essential intermediaries in transactions between merchants. De Roover (1948, p. 186), for example, observed that "in the Middle Ages, the regulation of the currency practically rested upon the money-changers; hence they performed a very important and quasi-public function."

Fairs, which brought together agents from distant locations for brief periods, developed as centers of long-distance trade. The archetype and most famous were the thirteenth-century fairs of Champagne which functioned as meeting places of northern (from Flanders) and southern (from northern Italy) European merchants. Once a merchant's coin holdings had been presented, the value of which was

[1] In contrast to the "historical" school of political economy which saw three main stages of economic development (and payments): "the prehistorical and early medieval stage when goods were exchanged for other goods; the later medieval state of 'cash' (money) economy, when goods were bought for ready money; and the modern stage of credit economy when commercial exchange was based on credit" (Postan 1973, p. 2; also, pp. 21–27).

assessed based on condition and exchange rates, and safely locked away with the money changer (Italians at the Champagne fairs and more generally not necessarily a he – money-changing being one of the few medieval professions in which there was no discrimination against women (de Roover 1948, p.174)), they could readily be used in payments. Since this assessment process was time consuming, thenceforth payment could be made by transferring ownership of the coins rather than the coins themselves. Such balances were "assignable" by oral order. The payer and the payee would appear together before the money changer (or bank), at which time the payer would order funds to be transferred from his account to the payee's account, rather than settling directly in coin. "Deposit contracts" were redeemable in full (at par) on demand.

Final settlement was generally deferred however. At the Champagne fairs, for example, trading took place in two periods. In the first Flemish merchants typically sold cloth to the Italians, while in the second the Italians sold spices to the Flemish. Italian purchasers in period one would have been allowed to overdraft their accounts in view of accumulating credits in period two. Final settlement of balances would then come at the end of the fair, or perhaps not. Merchants with debit balances at the end of the fair commonly extended their overdrafts until the next fair or else borrowed from those who accumulated a final credit balance.

Deposit banking therefore had its origins in money changing. Moreover, McAndrews and Roberds (1999) argue that the activity of payment was central to the original function of banks of deposit. Commodity money could adequately serve the purpose of a medium of exchange when settlement was immediate. But if the ability to enforce intertemporal commitments is limited, promises to pay are not completely satisfactory substitutes for cash transfers. In such a situation, profitable exchanges could be facilitated by a deposit bank lending to merchants via overdrafts with merchant deposits serving as collateral for the temporary overdrafts. Other merchants are willing to accept bank funds since incoming funds would be used to pay off their own overdraft loans. Roughly offsetting payments among merchants then are facilitated by the bank's provision of liquid payment services. In turn, or in reverse, banks are liquid because their customers' payments tend to be mutually offsetting. And there are liquidity economies of scale potentially at work here – more customers mean more mutually offsetting transactions and more liquidity. McAndrews and Roberds view this linking of buyer and seller as the "initial and key role of banks," with the function of intermediary linking saver and investor a consequence of the original payment intermediation (1999, p. 32).

Bancherii were known in Genoa as early as the twelfth century, and deposit banking was well established there by 1200 (spreading to Venice). In the North, in Flanders (Bruges), it was practiced by the second quarter of the fourteenth century (de Roover 1948, p. 247). The course of deposit banking, however, had its ups and downs. One of the downs came in the late fifteenth century when a wave of bank failures swept across Europe, importantly in Venice and Bruges. The Burgundian authorities in the Low Countries as a result decided to ban it completely – orders in 1489 prohibited the taking of deposits. In Venice there was a more delayed response. With the failure of the last two Rialto banks, in the wake of numerous earlier failures,

private banking there came to an end. In 1584 a monopoly public bank, the Banco della Piazza di Rialto, was established, the purpose of which was again to facilitate payments rather than financial intermediation. It proved so successful that by the close of the seventeenth century, public banks handled most of the deposit banking on the continent (Kohn 1999b, p. 25).

Amsterdam was one city which founded a public bank, the Bank of Amsterdam or Wisselbank, in 1609. It was city owned, accepted deposits, and did not lend. Its purpose was akin to the traditional money changers, to assure the quality of coins in the face of a proliferation of issues and issuers and accompanying risks of debasement. In circa 1600 there were between 800 and 1,000 different coins in circulation there, issued by mints in each province and many cities as well as by private mints (and not to mention counterfeits). When coins were deposited, the depositor was told the bullion value of and a receipt for the particular coins presented. Withdrawals could be made at a small fee, but more importantly transfers between accounts were allowed with no fee. In the 1680s when the Wisselbank abolished the right of withdrawal without a receipt, there was no protest, which could be taken as an indication that depositors were no longer interested in making withdrawals. Settlement of transactions was accomplished by the transfer of exchange bank money denominated in banco florins, a unit of account tied to no particular coin. As a result, Quinn and Roberds (2006, 2007) characterize the Wisselbank as the first modern central bank, one in which large-value payments were settled through the transfer of balances held there, the value of which were maintained through open market operations.

Bills of Exchange

Although the bill of exchange probably dates to around 1200, it began to come into its own later in the century as the fairs of Champagne began to wane (to be sure, fairs in general did not die out; they just shifted to other venues). As merchants became less itinerant and more "sedentary," it became increasingly necessary to make payments at a distance. Specie, coins or bullion, of course, could have been shipped, but that was costly and risky. The bill of exchange, an order to pay a certain person a certain amount (in a different currency) at a distant place, developed instead as the primary instrument of remittance (Usher 1914, pp. 569–570).[2] The typical bill involved four parties. First there was the remitter who wanted to make a payment abroad and provided the funds in local currency to the taker who in turn made out the bill to his agent or correspondent in a distant city. Then there was the payer or drawee on whom the bill was drawn and who was expected to pay it at maturity in the local currency and the payee, in whose favor the bill was made out (de Roover 1948, p. 53). For security two or three copies of the bill were usually sent. The taker on a

[2]A more comprehensive history of the bill of exchange than that which follows here may be found in Denzel (2010, pp. xxii–xlvi).

bill of exchange was often a trading company or merchant banker with branches or correspondents in several cities. Thus, merchant banking developed with the diffusion of the bill of exchange or vice versa.

Medieval bills were not discountable because of usury restrictions. But the price of a bill did reflect foreign exchange charges in which interest might be concealed. Bills of exchange were initially used to finance trade, but increasingly over time they were employed for purely financial purposes. Such transfers of funds over space and time, called "dry exchanges," were divorced from remittance. Indeed, as early as the fourteenth century, most bills of exchange were drawn out of financial, rather than trade, transactions (de Roover 1948, pp. 66–67; Kohn 1999a, p. 9). Such bills, later known as "accommodation paper," were held in low esteem in some circles because they did not arise out of real commercial transactions.

Medieval bills also were not negotiable. Bills were assignable, that is, the collection of the debt could be assigned to a third party. Agent A owes money to agent B as evidenced in a bill of exchange. B in turn might assign the right to collect the debt to a third party, agent C, by endorsing the bill on its back (this practice began in the 1570s, before that a formal document, an assignment note, had been drawn up, usually before a notary). If in addition the instrument was transferable, C would receive the full rights of B. Therefore if A, the original debtor, did not pay, C could have legal recourse against A, but not against B. With a negotiable instrument the debt to C is discharged only when he/she receives final payment. If A reneges, C has recourse not only against A but also against B (or against any other endorser in the chain of assignment). Thus, every additional endorsement strengthens the credit of the negotiable bill.

In Antwerp, where deposit banking had been banned, the principle of transferability was recognized by the courts by 1507 and that of negotiability by an edict of Charles V in 1536. As a result, in lieu of the transfer of deposits, negotiable bills became a standard means of payment among merchants. As bills were transferred from hand to hand, and they were in fact, with each subsequent endorsement, they became more secure. During this time in Antwerp, the practice of modern discounting developed as well, increasing the liquidity of bills outstanding (van der Wee 1977, pp. 322–332; Kohn 1999a, pp. 23–28). Usher (1914, p. 576) then could write "there can be little doubt of the essential perfection of the bill [of exchange] by 1650." The use of bills of exchange as a local medium of exchange spread to other areas as well, perhaps most famously in late eighteenth-/early nineteenth-century Lancashire where they served as the principal means of payment (Gilbart 1836, p. 79).

Notes, Checks, and Clearing Houses

In England there was no medieval tradition of banks of deposit evolving from private money-changing operations, because money changing had been the exclusive province of the Royal Mint. Instead, goldsmiths became the safe-keepers of cash and

valuables. Following the relaxation of regulations under Cromwell's Protectorate, many goldsmiths moved into deposit banking. By the time of the Restoration (1660), a network of bankers had developed in London (32 by 1670). These goldsmith bankers issued paper banknotes, promissory notes that were payable to, or redeemable by, the bearer at the issuing bank on demand. Such bearer notes could circulate without endorsement by each holder. When a bill of exchange was transferred by endorsement, each signer assumed a share of the collective responsibility for the final settlement of the bill. In contrast, the value of a banknote depended solely on the reputation of the issuing banker – each holder could transfer it in exchange without assuming any liability. The policy toward rival banks' note issues that developed among London goldsmith banks was one of mutual acceptance at par or face value and bilateral clearing without a formal coordinating institution (a pair of banks each presented the notes of the other, which it had accumulated over time. Any pair-wise imbalances were settled by a transfer of good funds – specie or later Bank of England notes). Quinn (1997) argues that mutual acceptance developed endogenously as a dominant strategy Nash equilibrium. Mutual acceptance produced positive externalities – each banker benefitted from the overall increased demand for notes and bills resulting from the expansion of participating agents. The practice of taking other banknotes at par did subject the accepting bank to the risk that the issuing bank might default, so rapid and regular clearing in which the issuing bank was presented with its notes for redemption lowered the receiving bank's exposure to default risk. Moreover, rapid and regular clearing performed an important monitoring or disciplining function of competitors in case issuing banks might become too enthusiastic (Quinn 1997).

By the time the Bank of England was founded in 1694, the practice of creating liabilities that would serve as a medium of exchange was already established. The London public should have been quite familiar with paper banknotes. The Bank however enjoyed some advantages in the issue of bearer banknotes. For one thing, it was granted a monopoly on joint-stock (incorporated) banking (until 1826 vis-à-vis country banks and 1833 vis-à-vis London banks), while private banks with more than six partners were forbidden from issuing notes. However with only the central office available for redemption, Bank of England notes typically circulated in or near London. By the last third of the eighteenth century, London private banks had largely given up on issuing their own notes in competition with those of the Bank of England.

Instead, the liability of choice among London bankers became the check (or cheque). A check is a simpler way of transferring deposits than having both parties appear in person before the banker, as was required in many medieval banks. It is a written order to pay a certain sum from the depositor's account to the payee when that order is presented at the depositor's bank. In practice, it is the local equivalent of a bill of exchange. The first checks appeared in Europe around 1400 in areas with deposit banks, but not always to universal approval – a Venetian ordinance banned checks there in 1526. In London they appeared with goldsmith banking in the mid-seventeenth century. As with banknotes, a system of mutual acceptance developed with bilateral clearing (Joslin 1954; Quinn and Roberds 2008, pp. 3, 7).

Checks offered several obvious advantages over notes – they were less subject to theft, provided a record of transactions, and were convenient for large-value transactions. On the other hand, they were "double claims," based on both the deposit bank itself and a particular agent's account there. Therefore, the recipient of a check had to consider both whether the bank upon which it was drawn would pay specie at par and also whether the check writer had sufficient funds in his/her account to cover the check. Both banknotes and checks were examples of "demandable debt," bank liabilities which may be converted into cash (specie) on demand. This feature enhanced their status as payment instruments by reducing uncertainty about their true value.[3] As checks were sent out for collection between banks, they began to hold credit balances with each other, particularly if the bank was in a distant location. These "interbank balances" were held in order to economize on shipping cash for immediate settlement of check clearances. As a result, banks needed to monitor and manage these implicit interbank loans. Goodfriend (1991, p. 12) argues that the same skills necessary to evaluate, monitor, and enforce loan agreements to nonfinancial borrowers proved useful to the efficient provision of payments services as well. Thus, "institutions specializing in information-intensive lending, i.e., banks, have applied their expertise jointly to the production of payments services and nontraded loans" (also see Kashyap et al. 2002).

The increasing volume of interbank exchanges led to the formation of the London Clearing House in 1773 by City bankers. These were the descendants of goldsmith bankers who dealt primarily with commercial customers and were located primarily in the City of London in and around Lombard St. This systematized previously informal private exchanges among bank clerks. Thirty-one of thirty-six city banks participated; West End bankers, originally located around the Strand and Fleet St. and dealing primarily with "gentlemen," were excluded. This was still a system of bilateral clearing and settlement, but at a central location on Lombard St., thereby minimizing the comings and goings of clerks around the City. The clerks came twice a day and dropped into the drawer of each those bank bills and checks payable, which were then summed up and offset by two salaried inspectors. Settlement at the end of the day was accomplished by transfer of specie or Bank of England notes.

In 1841 a single net payment system was finally adopted. In net clearing each bank calculated its net position against all other clearing house members, the total of checks it received which had been drawn on other banks compared with checks presented by other member banks drawn on it. Differences were settled by only one transaction between the bank and the clearing house. If the balance was positive, the bank's clearing house account was written up; if the balance was negative, good funds were transferred to the clearing house (after 1854 settlement was accomplished through transfers of Bank of England accounts). Multilateral net clearing certainly

[3]One other thing about banknotes here. Quinn and Roberds (2003) draw parallels between the development of privately issued banknotes and online currencies more recently, both motivated by the need to conduct transactions with strangers to provide a form of finality, "of being able to extinguish other debts by virtue of their transfer from debtor to creditor."

represented an advance in efficiency by reducing the level of reserves needed to be maintained for final settlement, but it also involved risk. The costs of the failure of any member fell on the clearing house as a whole. Nevertheless, as Seyd (1872, p. 56) wrote, "It would be difficult indeed to imagine anything more logical in construction and closer to perfection than the London Bankers' Clearing System . . . Were it not for the very dry matters of fact connected with the whole proceeding, a poet might be found to sing its praises."

Checks were essentially restricted to local payments. Within England the domestic or inland bill of exchange was the principal instrument used in nonlocal commercial transactions. Their negotiability made them a highly liquid financial instrument. Purchasing or discounting them (before maturity) became the standard method by which commercial banks extended credit. Secondary markets in domestic bills developed. They were intermediated by bill brokers who received bills drawn from net borrowing areas such as manufacturing districts and arranged their sale to agents in London or surplus agricultural areas. Initially, those brokers were just that, never taking a position in the transaction. Eventually they began intermediating these transactions using their own accounts. Over time the domestic bill drawn on London became the standard internal payment instrument even when neither party was located there.

The growth of bill brokers coincided with the growth of country (non-London) banking, in the second half of the eighteenth century. Country banks primarily issued paper banknotes which served as a local means of payment. Checks were used less as a means of payment than as an order to the bank to pay out cash. Not until the second quarter of the nineteenth century did they come into more general use in the countryside. Nevertheless, with the growth of industry and trade later in the eighteenth century so also was there an increase in the need to make payments at a distance. Although there were no formal legal restraints on size or branching, the six-member partner limitation inhibited raising substantial amounts of capital, as well as monitoring of distant locations. While family linked some banks together, most country banks were unit banks. As a result, correspondent networks developed between interior country banks on the one hand, and London banks on the other, to facilitate interbank transactions (Quinn 2004).

Every country bank needed a London agent or correspondent bank. The agent would collect bills payable in London as they matured, serve as a source of London funds for local customers who needed to make payments there, provide investment assistance and advice, serve as a redemption agent for any country banknotes that wandered too far from their issuer, and provide funds when needed by rediscounting bills held by country clients, among other things. Later, as the use of checks outside London increased, London banks began to clear country checks. The name of its London correspondent was printed in the corner of the check form. Checks received were sent to the London correspondent for presentation to the Clearing House which in turn mailed them to the banks on which they were drawn. Customers within 1 day's post from London saw their account credited within 2 days. The threat by country banks of starting their own clearing house led to the establishment of a separate Country Clearing section in London in 1858. Except for provincial

operations in Manchester, Liverpool, Birmingham, and a few small local exchanges, every check drawn on an English bank was collected through London. The price for such services was typically maintaining a substantial non-interest-bearing deposit with the London correspondent, although sometimes fixed annual fees were paid (James 2012, pp. 135–138). Ties between country and London banks weakened after 1826. The development of large joint-stock banks with nationwide branches meant that much clearing and settlement previously routed through London could be done internally ("on us"). As the Bank of England opened provincial offices, maintaining a London agent became less important.

Correspondent Banking Networks in Nineteenth-Century America

As internal trade grew in the early United States, so did the problems of making payments at a distance (distances were greater for one thing). The shipment of specie was again an obvious way of settling accounts, but this was rarely used in nonlocal, non-retail transactions (Colwell 1859, pp. 135, 190, 262, 447). Two institutional responses emerged both dating from the 1820s. First of all, private banknotes were sometimes used in payments outside their locality. Many banknotes issued by country banks all over New England made their way into Boston, the regional commercial and financial center, and circulated there in competition with those of local banks. Out-of-town notes were typically valued at a discount from par or face value because of the costs involved in returning them to the issuing bank for redemption. The initial response to this "flood" of "foreign" notes was to get them out of circulation and also make a little money in the process. Local banks bought them at a discount and returned them for redemption at par, but competition narrowed the spread between their par value and market price so that such an operation was "hardly profitable." The obvious solution was collusion, a strategy in which a coalition of Boston banks (in 1824) would pool their resources for purchasing country notes and then send them back for redemption. The redemption agent was the Suffolk Bank (hence this was called the Suffolk system).

Rather than country banks facing unannounced calls by Suffolk Bank agents presenting unpredictable sums of notes for redemption in specie, the system was systematized by their maintaining (non-interest-bearing) clearing and redemption accounts with the Suffolk Bank. The resulting arrangement was one of net clearing. Instead of gross clearing in which notes were simply returned to the issuing bank for redemption, country banks sent their accumulated out-of-town notes to the Suffolk Bank. There the submitting country bank's account would be credited for notes received from it and debited for notes issued by it received from other banks. Consider the Suffolk system as a private payment network/clearing system with positive externalities, the benefit to an individual member (from being able to clear at par the notes of distant banks, for one thing) increases as more users join the network. Such a region-wide net clearing system was quite efficient, in the sense of minimizing movements of specie necessary for settlement and ensuring that the

notes of system banks circulated within the region at par. It has been widely admired by many later historians and economists for such reasons. Calomiris and Kahn (1996), for example, in support of the "sanguine" view of the Suffolk system, find that New England banks were able to issue more notes, which traded at uniform and low discount rates, while maintaining lower specie reserves, as compared to Middle Atlantic banks (i.e., ones not in the Suffolk system). Nevertheless, in spite of its admirable efficiency, the Suffolk system was heartily despised by most system members, many of which defected promptly when a rival (Bank of Mutual Redemption) appeared in the 1850s, prompting its rather precipitous collapse. Bodenhorn (2002) explains this paradox of the Suffolk system, admired by outsiders but strongly disliked by insiders. First note that in networks pricing at marginal cost might not produce efficient configurations or usage. In particular, members who place low values on participation but bring large external benefits to others should pay only a small part of common costs, perhaps even less than marginal cost. Based on a calculation of the implicit cost of network membership in the network, he argues that the clearing services of the network had indeed been mispriced – the smaller country banks, which benefitted the least from participation, paid proportionately the most, while the large Boston banks, which benefitted the most, paid the least.

Second, from the 1820s through the mid-1830s, the Second Bank of the United States (SBUS) under Nicholas Biddle with its network of interstate branches dominated the system of interregional payments through its role in the market for inland (domestic) bills of exchange (Catterall 1902, pp. 138–143; Knodell 2003). Internal bills of exchange drawn upon a SBUS branch could be discounted by the payee at his/her local SBUS. The funds at the drawn upon branch could in turn be sold to other local customers who needed to make payments in that city in the form of a bank draft, a check drawn by one bank against funds deposited in another. The price of a bank draft reflected an (domestic) exchange charge, the price of funds in a distant city in terms of local funds. The Second Bank served as a market maker in domestic exchange – it was always ready to be a party to a transaction. If someone wanted to buy, it would sell; if someone wanted to sell, it would buy and generally at rates below what would have prevailed in a private market, consistent with Biddle's desire to promote internal trade.

The SBUS branch network made for a quite efficient national payments network based on collective settlement, minimizing the interregional shipment of specie necessary to settle account imbalances. Between the demise of the Second Bank of the United States (in 1836) and the founding of the Federal Reserve System, the United States lacked a central monetary authority to orchestrate the clearing and collection of payments among highly localized banks. The prohibition against interstate branching furthermore precluded the formation of nationwide banks that could mediate a national payment system. The closing of the interstate branches of the SBUS thus ushered in a period of financial disintermediation in interregional payments. Knodell (1998) measures the rise in domestic exchange rates (Cincinnati and Cleveland on New York) in the wake of the SBUS closing, and initially they were substantial. Note brokers and private bankers then became active participants in the now-decentralized system of domestic exchanges (Knodell 1998, pp. 717–719;

Bodenhorn 2000, pp. 177–185). Individuals needing to make payments in New York, for example, could buy bills payable there through a broker.

Private banknotes began to be used increasingly in transactions at a distance, being cheaper and more convenient to ship than specie. Secondary markets reemerged in these out-of-town notes. Local note brokers, which had been overshadowed by SBUS operations, bought, sold, and/or sent these "foreign" notes back to the issuing bank for redemption. Out-of-town notes were usually valued at a discount from par or face value because of the costs involved in returning them to the issuing bank for redemption (and the risk that when it was presented at the counter the issuing bank might refuse to pay for it[4]) with their current prices published regularly in periodicals called banknote reporters. Using such data, Gorton (1996, 1999) studies the pricing of banknotes and the process of reputation formation for new banks in the late antebellum period, arguing that market discipline (pricing bank risk) prevented widespread wildcat banking.

Alternatively, networks of private bankers developed in lieu of the interstate branch network of the SBUS (Knodell 1998, 2010). Private banks increasingly maintained deposits in other banks in distant cities, the locations of which were determined by the needs of local trade (Weber 2003). These interbank or correspondent accounts allowed local banks to sell drafts drawn on them to customers needing to make payments in distant cities. As New York City emerged as the commercial and financial center of the country, it also became the focus of these correspondent bank networks. By just before the American Civil War, virtually every bank in the country maintained a New York correspondent, and the sight draft on New York, or New York exchange, had become the dominant financial instrument used in interregional or intercity payments. Payments between agents in different cities, say New Orleans and Cincinnati, were generally settled in New York funds, the standard means of interregional payment, since virtually all banks held accounts there. The earlier bilateral payment relationships had evolved into one centered on New York in essentially a hub and spoke network. New York exchange became *the* currency for settlement of long-distance transactions. This in turn allowed for a quite efficient system of essentially net collective settlement of nonlocal payments. That is, transactions between parties whose banks shared the same city correspondent could be settled as an intrabank "on us" transfer of funds; those between parties whose banks had different correspondents in the same financial center (i.e., New York) involved only a local movement of funds, greatly reducing the necessary shipment of reserves.

New York became the "clearing house of the country," and the crucial institution there in turn was the New York Clearing House (NYCH), founded in 1853. It was a net settlement system in which members were exposed to the risk of default by other members. Any losses at settlement were covered by the NYCH as a whole. As a

[4]The Second Bank of the United States notes however had been redeemable at par at any branch, regardless of the city of issue. The fact that their value did not decline with distance made them the paper money of choice for long-distance payments before 1836 (Temin 1969, p. 36).

result, membership was closely guarded, and applicant banks to the "club" were carefully scrutinized for soundness (James and Weiman 2011). In times of some financial crises in order to reduce pressure on reserves, the NYCH issued clearing house loan certificates to be used internally in settlement. Whether because of its select membership or the usefulness of clearing house loan certificates, the failure rates of NYCH banks during panics were very, very low. This ability and willingness to create high-powered money in times of panic have led some writers to hail the NYCH as an incipient quasi- central bank and lender of last resort (Gorton 1985; Timberlake 1984). Such a view, however, neglects the fact that because of internal conflicts among members, the NYCH was not in fact so effective in forestalling financial crises when they loomed. In both 1893 and 1907, NYCH banks nevertheless suspended payments, i.e., temporarily reneged on the commitment to pay out cash for deposits at par on demand (see Wicker 2000; Sprague 1910). In turn, these breakdowns of the payment system, which quickly spread nationwide, had serious consequences for real economic activity (e.g., see Sprague 1910). James et al. (2013) argue that such suspensions acted as severe adverse supply shocks and offer econometric evidence based on monthly data that contractions intensified during suspension periods (with a statistically significant decline in real activity of around 10–20%).

Business customers making payments in New York could buy a bank draft drawn on his/her local bank's correspondent there (New York exchange). In turn, such customers also sold exchange to their banks by depositing drafts or checks drawn on a New York bank. Banks would then remit these items to their correspondent for collection and receive payment usually in the form of ledger entries to their correspondent balances, rather than shipments of cash. In the course of providing routine payment services to business customers, banks would deplete and replenish their city correspondent balances. At any point in time, therefore they could find themselves with deficient or excess correspondent balances. To remedy these imbalances, interior banks could arrange to ship cash to or from their New York correspondents, but that again would then incur significant transaction costs. As a cheaper alternative, banks developed local wholesale or interbank markets in exchange where they bought and sold surplus correspondent balances, in other words local markets for domestic or New York exchange (Garbade and Silber 1979). The system of internal or domestic exchanges was a fixed exchange rate regime. The value of a dollar (in terms of gold) in New York was the same as that of one in Chicago. The spot price of New York funds in Chicago could differ (in normal times) from the mint parity exchange rate (one) within the currency points, which reflected the cost of shipping cash from Chicago to New York or vice versa, without eliciting an interregional/intercity currency flow. James and Weiman (2010) investigate longer-term changes in domestic exchange markets, particularly the dramatic decrease in variability or volatility over the late nineteenth century.

The practice of deposit banking, in which deposits rather than banknotes were the primary bank liability and means of payment, expanded rapidly from urban areas into interior regions in the decades after the Civil War. By the 1880s bank customers began to make payments to distant parties with checks drawn on their local banks

rather than through city drafts, undermining the centralization of payment operations made possible by drafts (James and Weiman 2010). The legal framework was not well designed for collecting checks used in interregional payments. Under the common law and commercial code, no bank could charge for payment on checks presented at its counter. As a consequence, the process of interregional clearing became more complicated, with New York banks utilizing their correspondent networks to act as collection agents for their country bank clients. Existing correspondent bank networks provided the structure to orchestrate their clearing, the process of transmitting, reconciling, and confirming payment orders. Settlement was still typically accomplished in New York by a transfer of funds between the payer and the payee banks' city correspondents. New York exchange therefore remained the means of settlement in long-distance transactions. Nevertheless, the resulting system was widely criticized for its perceived inefficiencies – primarily indirect routings of checks passed along from correspondent to correspondent on their (sometimes circuitous) way to be presented for collection and excessive charges by some (generally rural) paying banks for remitting settlement funds on checks not presented directly at their counter. More recent assessments however (Lacker et al. 1999; James and Weiman 2014) have suggested that such criticisms were exaggerated at best (see next section). Probably a more serious criticism was that the proliferation of interbank accounts for clearing and collection purposes led to substantial holdings of excess reserves overall under the *ancien régime* (Laughlin 1912).

The Twentieth Century

Soon after its founding in 1914, the Federal Reserve, the new US central bank, entered the check clearing market with the goal of displacing the myriad private correspondent and clearing arrangements that had been in place. Its stated goal was to enhance the efficiency of the payment system by standardizing banks' procedures for the clearing and collection of checks and centralizing their reserve holdings. How successful this effort was will be addressed in a few moments. But note here that the Fed system might be regarded as something of a template for modern clearing and collection operations. First of all, settlement between banks was accomplished through transfers of their balances or reserves held at the Fed. Central bank money replaced New York balances as the settlement medium of choice. To be sure, this was not entirely new – London banks, for example, settled among themselves by transfers of Bank of England notes and later deposits from early on in the nineteenth century. But notwithstanding, Norman et al. (2011) in their survey of the history of interbank settlement characterize the convergence of monetary systems around the world toward ones with interbank settlement in state-backed central bank money as a general twentieth-century phenomenon. Indeed, Charles Goodhart (1988) argued that pressure to devise efficient clearing and settlement procedures was very important in the development of central banking in general (and the centralization of reserves).

Second, from 1918 clearing and settlement was accomplished via an exclusive leased telegraph wire network known as Fedwire. This was an arrangement of real-time gross settlement (RTGS), the first, in which each payment is separately settled as it is sent. Every payment from one bank to another represents an irrevocable transfer of central bank funds at the time the order is transmitted, in continuous or real-time settlement, as opposed to a final net settlement at the end of the period. Although multilateral net settlement in clearing houses reduced the amount of funds needed to be transferred to settle accounts at the end of the period, it also exposed member banks to collective risks that another member might not be able to settle. In "secured" net settlement systems, all participants might be required to post collateral, which in total should be adequate to cover any failed positions. Gross settlement, while absolving individual banks of such collective risk bearing, imposes much greater liquidity demands on them. Over the course of a day, depending on the timing, there could be times when a bank's reserves might dip perilously low. Similarly, payment blockages might arise when one bank's outgoing payment waits on the arrival of an incoming payment from another bank. Smooth operation of a RTGS system then depends importantly on the provision of intraday credit by the central bank to offset temporary adverse movements in member bank's reserve positions. Since repayment is required by the end of the business day, such loans are called daylight overdrafts.

There has been "virtual agreement among major central banks that gross settlement makes wholesale payment systems more immune to widespread financial disruption" (Emmons 1997, p. 24), and over the late twentieth century, there has been a general move toward RTGS payment systems in most major industrialized countries. This has been particularly the case for countries participating in the European Economic and Monetary Union, which needed to establish common standards for interlinking in the TARGET (Trans-European Automated Real-time Gross Settlement Express Transfer system) payment network (later succeeded by second-generation TARGET2).

Fedwire is an example of a "wholesale" payment system, one which accomplishes periodic large-value transfers between financial institutions. Many of these represent a derived demand for settlement of claims arising from retail payments. In the United States, large-value settlement is not a government monopoly – Fedwire's major private rival/alternatives being CHIPS (Clearing House Interbank Payments System), a subsidiary of the New York Clearing House with a limited membership of large US and foreign banks using net deferred settlement (NDS), and CLS (Continuous Linked Settlement) in the foreign exchange market.

Now shift the focus from contemporary large-value payment networks back to smaller-value "retail" networks and back in time again to the founding of the Federal Reserve. In view of the perception of the then existing system of check clearing and collection as flawed and inefficient, virtually from its inception, the Fed vigorously pursued strategies to establish universal par clearing, in which checks drawn on any bank would be paid in full at par without deduction for any presentment or remittance fees. This was the final step in a true national monetary union. This was however vigorously resisted by so-called non-par banks. These were banks that were

not members of the Federal Reserve system, located primarily in rural areas, and which often relied heavily on such fees as important revenue sources. The battle was fought in the Congress, state legislatures, and the courts, making par clearing and collection one of the major issues of the early Federal Reserve period. The Supreme Court decided in 1923 that since Congress had not required the Fed to establish nationwide par clearing, it could not force nonmember banks to pay checks at par. The Fed then abandoned universal par collection as a policy objective, and non-par banks persisted for decades – in fact until 1980 when they were abolished by the Congress.

Some more recent writers have cast doubt on the wisdom of the Fed's objective of universal par clearing and collection of checks. In an influential paper William Baxter (1983) argued that in payment networks characterized by joint costs and interdependent demand side payments, or interchange fees, between financial institutions (imposing a gap between the price the buyer pays and the sum the seller receives) could be necessary to achieve equilibrium. Remittance fees, which in effect supported a more geographically extensive or comprehensive payment network than otherwise would have been possible, might be considered as such an interchange fee. Also along these lines, Lacker et al. (1999) characterize the pre-Fed check collection system as a payment network with externalities in which pricing could/should diverge from marginal cost. They take the dispute about remittance fees as one really in fact about who bore the costs of the payments network – the collecting (often city) or the paying (often country) bank. The advantage of the Fed was not in greater efficiency, but a legal one, being able to present checks for payment through the mail, a practice which was not available to private banks. Correspondent networks in turn were configured consonant with the needs of trade, so the odd indirectly routed check that passed from bank to bank was just an outlier. Gilbert (2000) on the other hand finds an increase in payment system efficiency due to the Fed as evidenced in decreased ratios of bank cash to total assets as banks overall no longer needed to hold as large or as many correspondent accounts for check clearing and collection.

In any case, the US check clearing and collection system did not, in fact, develop into one dominated by the Fed. While member banks could use the Fed's facilities for clearing and collection, they were not obliged to do so (the only obligation here was to pay at par on any checks presented by the Fed), and if the checks had been drawn on local banks, a more attractive option was often to continue to clear through local clearing houses, which still continued to function. Indeed a rough division of labor developed in which out-of-town checks were typically cleared through the Fed, while the clearing house handled local ones. Starting from zero, by 1934 the value of checks processed by the Fed had risen to around two-thirds of those handled by private clearing houses (Gilbert 1998, p. 133).[5]

[5]By 2010 Federal Reserve banks processed slightly less than half the value of interbank paid checks.

Furthermore, after 1980 when Federal Reserve banks were required to charge full costs for payment services, their share in both large- and small-value payments began to decline (but has bounced back more recently). This trend toward greater privatization has been noted with approval by some, for example, Lacker and Weinberg (1998) and Green and Todd (2001), who saw the provision of payments services as superfluous to the Fed's core central bank functions of conducting "monetary policy, banking regulation, and financial stabilization." James and Weiman (2005) disagree pointing to the Fed's significant roles in standards setting and coordination and as clearing house of last resort. Indeed, doubts had been raised that the risks of disruption in the operation of the payment system had increased due to the diminished role of the Fed in processing payments and the increased importance of private channels (Summers and Gilbert 1996). Such fears might be allayed, for now at least, in view of the decisive actions taken by the Federal Reserve in response to the terrorist attacks of September 11, 2001. Potential payment system gridlock as a result of disruptions to the Fedwire system was averted by the temporary provision of unprecedented amounts of liquidity to the banking system (McAndrews and Potter 2002).

The widespread use of checks in retail payments has been primarily a phenomenon of the Anglo-Saxon world. In contrast, in many/most Continental European countries (as well as in a number of others elsewhere), giro transfers dominated. In this case the payer ordered his/her institution to transfer funds from his/her account directly to the account of the payee with no action by the latter required (as compared with the case when a payer writes a check to the payee who must then present it to his/her bank, and the transaction is complete only when the check has been collected by the depositing bank from the bank upon which it was drawn). In the absence of well-developed concentrated banking systems accessible to many households on the one hand, or extensive correspondent banking networks on the other, payers in many European countries often turned to government institutions, such as the postal service or the central bank, which could offer payment services (credit transfers) through nationwide networks of branches (Hein 1959; Bank for International Settlements 1999).

With the development of electronic noncash payments methods (Bank for International Settlements 1999, pp. 13–14), there has been some convergence in the use of retail payment instruments across countries as check-using countries employ more direct transfers (obviously, rather than giro-based systems using more checks). In the United States, for example, the use of direct credit transfers in the form of the ACH (automated clearing house) system grew rapidly over the 2000s so that they constituted 51% of the value of total cash payments by 2009 (although only 18% of their number), along with the increasing use of direct debit transfers as well. Nevertheless, the distinctive (or anomalous) feature of the US retail payment system relative to those in other developed countries remains the continued relatively widespread use of paper checks. In 2009 checks accounted for 22% of the volume of noncash payments and 44% of their value in contrast with figures below 10% in most European countries (France being the exception with an 18.3% share of total transactions) and below 1% in some countries such as Germany.

Other payment instruments experiencing rapid growth in recent years are credit and debit cards. In the United States in 2009, they accounted for over half of all noncash transactions (20% for credit cards and 35% for debit cards). The average value per transaction was relatively low however, accounting for only 5% (credit cards 3%, debit cards 2%) of the total value of 2009 US noncash payments. Similarly, in most eurozone countries in 2008, they made up between one- and two-thirds of noncash transactions (excepting Austria, Germany, and Slovakia) (Kokkola 2010, p. 176).

Plastic card clearing and settlement networks (e.g., Visa or Master Card) are private joint ventures of depository institutions. In such networks interchange fees impose a gap between the amount buyers pay and sellers receive. So, notwithstanding the efforts of the Federal Reserve to stamp them out in the 1920s, non-par payment networks are back and occupy an integral part of the payment system.

Summary

There are trade-offs between risks and costs in payment systems. A system, for example, in which all payments were made in cash, might have low risk, but high costs, particularly for payments at a distance. On the other hand, payments based on credit arrangements, check or drafts, might reduce costs (of shipping cash) but would involve more risk. Innovations or improvements in the payment system then would shift the frontier or trade-off curve inward – increasing payment system efficiency at a given level of risk and/or decreasing risk at a given level of costs (see the model in Berger et al. 1996). Goodfriend (1991, p. 9) observed that "the evolution of the payments system has been, in large part, driven by efficiency gains from substituting credit, i.e., claims on particular institutions, for commodity money," but many payment system innovations did both – reducing risk and also improving efficiency (lowering costs).

Even the use of cash, the simplest form of payment, in the medieval period involved some risks. Locally there were problems in assessing the value of the myriad coins in circulation, as well as counterfeits. Money changers both provided expert assessments of such coins presented and offered a safe place to store them. Assignment of these deposit accounts held with banks or money changers in turn both reduced costs by decreasing necessary physical transfers of specie and also facilitated exchange between strangers. For nonlocal transactions, those at a distance, merchant networks dealing in bills of exchange fulfilled similar functions.

With the growth of banks in the more modern period, institutions developed to clear and settle bank liabilities, notes or deposits (i.e., checks), more efficiently. The Suffolk system both greatly reduced the need for physical transfers of cash and reduced risk, as evidenced in the elimination of discounts on New England banknotes. Local clearing houses similarly minimized physical cash transfers and offered timely final settlement for checks. Within the United States the development of correspondent bank networks focused on New York allowed settlement between distant banks to be accomplished simply through transfers of bankers' balances held

there rather than through large cash movements. In the twentieth century interbank settlement has increasingly been in terms of central bank money, providing finality, and accomplished through real-time gross settlement, free of systemic risk, rather than by net settlement at the end of the business day.

Acknowledgements I would particularly like to thank David Weiman for his thorough, penetrating, and most useful comments.

References

Bank for International Settlements, Committee on Payment and Settlement Systems (1999) Retail payments in selected countries: a comparative study. Bank for International Settlements, Basel

Baxter WF (1983) Bank interchange of transactional paper: legal and economic perspectives. J Law Econ 26:541–588

Berger AN, Hancock D, Marquardt JC (1996) A framework for analyzing efficiency, risks, costs and innovations in the payments system. J Money Credit Bank 28:696–732

Bodenhorn H (2000) A history of banking in antebellum America. Cambridge University Press, New York

Bodenhorn H (2002) Making the little guy pay: payments-systems networks, cross-subsidization, and the collapse of the Suffolk system. J Econ Hist 62:147–169

Calomiris C, Kahn CM (1996) The efficiency of self-regulated payments systems: learning from the Suffolk system. J Money Credit Bank 28:766–797

Catterall RCH (1902) The Second Bank of the United States. University of Chicago Press, Chicago

Colwell S (1859) The ways and means of payment. J. B. Lippincott, Philadelphia

de Roover R (1948) Money, banking and credit in mediaeval bruges. The Mediaeval Academy of America, Cambridge, MA

Denzel MA (2010) Handbook of world exchange rates, 1590–1914. Ashgate, Farham

Emmons WR (1997) Recent developments in wholesale payments systems. Fed Reserve St Louis Rev 79:23–43

Fantacci L (2005) Complementary currencies: a prospect on money from a retrospect on premodern practices. Financ Hist Rev 12:43–61

Fantacci L (2008) The dual currency system of Renaissance Europe. Financ Hist Rev 15:55–72

Garbade KD, Silber WL (1979) The payments system and domestic exchange rates: technological versus institutional change. J Monet Econ 5:1–22

Gilbart JW (1836) A practical treatise on banking, 4th edn. Longman, Ress, Orme, Brown, Green, & Longman, London

Gilbert RA (1998) Did the Fed's founding improve the efficiency of the U.S. payments system? Fed Reserve St Louis Rev 80:121–142

Gilbert RA (2000) The advent of the Federal Reserve and the efficiency of the payments system: the collection of checks, 1915–1930. Explor Econ Hist 37:121–148

Goodfriend M (1991) Money, credit, banking, and payments system policy. Fed Reserve Richmond Econ Rev 77:7–23

Goodhart CAE (1988) The evolution of central banks. MIT Press, Cambridge, MA

Gorton G (1985) Clearinghouses and the origin of central banking in the United States. J Econ Hist 45:277–283

Gorton G (1996) Reputation formation in early bank note markets. J Polit Econ 104:346–397

Gorton G (1999) Pricing free bank notes. J Monet Econ 44:33–64

Green EJ, Todd RM (2001) Thoughts on the Fed's role in the payments system. Fed Reserve Minneap Q Rev 25:12–27

Hein J (1959) A note on the giro transfer system. J Financ 14:548–554

James JA (2012) English banking and payments before 1826. In: Hanes C, Wolcott S (eds) Research in economic history, vol 28. Emerald, Bingley, pp 117–149
James JA, Weiman DF (2005) Financial clearing systems. In: Nelson R (ed) Complexity and limits of market organization. Russell Sage, New York, pp 114–155
James JA, Weiman DF (2010) From drafts to checks: the evolution of correspondent banking networks and the formation of the modern U.S. payments system, 1850–1914. J Money Credit Bank 42:237–265
James JA, Weiman DF (2011) The National Banking Act and the transformation of New York banking after the Civil War. J Econ Hist 71:338–362
James JA, Weiman DF (2014) Political economic limits to the Fed's goal of a common national bank money: the par clearing controversy revisited. In: Hanes C, Wolcott S (eds) Research in economic history, vol 30. Emerald, Bingley (forthcoming)
James JA, McAndrews J, Weiman DF (2013) Wall Street and Main Street: the macroeconomic consequences of New York bank suspensions. Cliometrica 7:99–130
Joslin DM (1954) London private bankers, 1720–1785. Econ Hist Rev 7:176–186
Juncker GR, Summers BJ, Young FM (1991) A primer on the settlement of payments in the United States. Fed Reserve Bull 77:847–858
Kahn CM, Roberds W (2009) Why pay? An introduction to payments economics. J Financ Intermed 18:1–23
Kashyap AK, Rajan R, Stein JC (2002) Banks as liquidity providers: an explanation for the coexistence of lending and deposit-taking. J Financ 57:33–73
Kinley D (1910) The use of credit instruments in payments in the United States. National Monetary Commission, Government Printing Office, Washington, DC
Knodell JE (1998) The demise of central banking and the domestic exchanges: evidence from antebellum Ohio. J Econ Hist 58:714–730
Knodell JE (2003) Profit and duty in the Second Bank of the United States' exchange operations. Financ Hist Rev 10:5–30
Knodell JE (2010) The role of private bankers in the US payments system. Financ Hist Rev 17:239–262
Kohn M (1999a) Bills of exchange and the money market to 1600. Working paper Department of Economics, Dartmouth College
Kohn M (1999b) Early deposit banking. Working paper Department of Economics, Dartmouth College
Kohn M (1999c) Medieval and early modern coinage and its problems. Working paper Department of Economics, Dartmouth College
Kokkola T (2010) The payments system. European Central Bank, Frankfurt
Kuroda A (2008) What is the complementarity among monies? An introductory note. Financ Hist Rev 15:7–15
Lacker JM, Weinberg JA (1998) Can the Fed be a payment system innovator? Fed Reserve Richmond Econ Q 84:1–25
Lacker JM, Walker JD, Weinberg JA (1999) The Fed's entry into check clearing reconsidered. Fed Reserve Richmond Econ Q 85:1–31
Laughlin JL (1912) Banking reform. National Citizens' League, Chicago
McAndrews J, Roberds W (1999) Payment intermediation and the origins of banking. NYFRB staff report: 40, New York
McAndrews J, Potter S (2002) Liquidity effects of the events of September 11, 2001. FRBNY Econ Policy Rev 8:59–79
Norman B, Shaw R, Speight G (2011) The history of interbank settlement arrangements: exploring central banks' role in the payments system, Bank of England working paper no 412. Bank of England, London
Postan MM (1973) Medieval trade and finance. Cambridge University Press, Cambridge
Quinn S (1997) Goldsmith-banking: mutual acceptance and interbanker clearing in restoration London. Explor Econ Hist 34:411–432

Quinn S (2004) Money, finance and capital markets. In: Floud R, Johnson P (eds) The Cambridge economic history of modern Britain, vol I. Cambridge University Press, Cambridge, pp 147–174
Quinn S, Roberds W (2003) Are on-line currencies virtual banknotes? Fed Reserve Atlanta Econ Rev 88:1–15
Quinn S, Roberds W (2006) An economic explanation of the early Bank of Amsterdam, debasement, bills of exchange, and the emergence of the first central bank. Federal Reserve Bank of Atlanta, Atlanta 2006–13
Quinn S, Roberds W (2007) The Bank of Amsterdam and the leap to central bank money. Am Econ Rev 97:262–265
Quinn S, Roberds W (2008) The evolution of the check as a means of payment: a historical survey. Fed Reserve Atlanta Econ Rev 93:1–28
Seyd E (1872) The London banking and banker's clearing house system. Cassele, Petter, & Gilpin, London
Sprague OMW (1910) History of crises under the national banking system. National Monetary Commission, Government Printing Office, Washington, DC
Spufford P (1988) Money and its use in medieval Europe. Cambridge University Press, Cambridge
Summers BJ, Gilbert RA (1996) Clearing and settlement of U.S. dollar payments: back to the future? Fed Reserve St Louis Rev 78:3–27
Temin P (1969) The Jacksonian economy. W.W. Norton, New York
Timberlake RH Jr (1984) The central banking role of clearinghouse associations. J Money Credit Bank 16:1–15
Usher AP (1914) The origin of the bill of exchange. J Polit Econ 22:566–576
van der Wee H (1977) Monetary, credit and banking systems. In: Rich EE, Wilson CH (eds) Cambridge economic history of Europe, vol V, The economic organization of early modern Europe. Cambridge University Press, Cambridge, pp 290–392
Weber WE (2003) Interbank payments relationships in antebellum Pennsylvania. J Monet Econ 50:455–474
Wicker E (2000) Banking panics of the gilded age. Cambridge University Press, New York

Interest Rates

Eric Monnet

Contents

Abstract

Many influential cliometric studies have examined interest rates in order to assess the investment efficiency, the integration of markets, the economic effects of changes in policies or institutions, the sources of macroeconomic cycles, and so on. The common feature of this approach to economic history is that it is based on the crucial assumption that interest rates are the market prices at which demand meets supply. In this perspective, most debates focus on how to calculate yields or compare different rates of return on capital. Cliometricians developed innovative methods to construct yields and lending rates that were not specified in historical sources. It is only quite recently that economic historians have turned to cases where interest rates are not market-clearing prices. In such cases, there is little connection between interest rates and the state of the economy. Highlighting market imperfections, some recent studies have challenged earlier historical interpretations that overlooked the potential disconnection between prices

E. Monnet (✉)
Banque de France, Paris School of Economics and CEPR, Paris, France
e-mail: eric.monnet@psemail.eu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_51

(interest rates) and quantities. They offer new insights into the historical functioning of credit markets, central banking, and government intervention in financial systems.

Keywords
Interest rates · Internal rate of return · Market yields · Return to capital · Central banks · Institutions · Market integration · Financial history · Market clearing · Credit rationing · Government bonds

Introduction

"The problem of interest has engaged the attention of writers for two thousand years, and of economists since economics began."[1] The first sentence of the famous book, *The Rate of Interest*, published in 1907 by the economist Irving Fisher (1907), reminds us that the rate of interest is not a concept recently formalized and used by historians to examine past economic practices. Interest rates were mentioned, applied, registered, and negotiated from ancient to modern times (Homer and Sylla 2005). Their use and the debates that surrounded them involved religious, cultural, and legal considerations (usury law being the most noted example; see, e.g., Hoffman et al. 2000; Botticini and Eckstein 2012). So quantitative economic historians are not the only historians interested in interest rates. Then, what makes the "cliometric" contribution to the study of interest rates specific?

Cliometricians have looked at interest rates as a market price in order either to infer some information about individual, collective, or policy preferences or to directly assess the efficiency of the economy. The reasons cliometricians have looked at interest rates, and the way they did it, are indeed much in line with what Deirdre McCloskey (1978, p.15) said about the general development of cliometrics: "Not counting but economic theory, especially the theory of price, is the defining skill of cliometricians, as for other economists." In light of price theory, interest rates are typically examined as an indicator of the integration of markets, the rationality of actors, their expectations, or the confidence in a corporation, a polity, or a government. Some important contributions to economic history that will be presented in this survey used interest rates (either profit rates or market rates) and built on economic theory to challenge conventional history, claiming that the information previously gathered by historians was too scattered or subjective. In this perspective, the study of interest rates is seen as a means of inferring

[1]Due to space limitations, many important articles on the subject had to be left out. Therefore, this survey should not be seen as an attempt to fully discuss the most recent work on this issue. What I have tried to achieve is to present the widest possible variety of methods, to highlight the pioneering work of early cliometricians which have inspired many subsequent researchers, and to highlight and criticize the strong theoretical assumptions that were the basis for standard studies of interest rates by economic historians. Although an attempt has been made to paint a general picture of the work of cliometricians, this survey obviously reflects the author's biases and research interests.

coherent and valuable information from the market, and it is considered superior to alternative methods based on limited qualitative sources, which may suffer from undergoing selection bias (Conrad and Meyer 1958; North and Weingast 1989; Rappoport and White 1993; Willard et al. 1996).

Various theories have been used to interpret interest rates. Neoclassical growth theory views the interest rate as the price that equates investment and saving and enables a productive allocation of factors. In microeconomic theory with rational agents but asymmetric information, the equilibrium interest reflects information problems. In financial theory (with perfect or imperfect information), the interest contains information about expectations. Building on Keynesian theory, the interest rate is the price that equals the demand and supply for money. Counting is not a minor issue, however. The reliance on theory, beliefs in market mechanisms, and the quest for historical market prices pushed quantitative economic historians to compute market yields that were not specified in contracts and to construct returns on investment based on actual or expected earnings. This is done even when historians have no specific information on whether and how economic agents actually computed such returns. Thus, as recalled by Homer and Sylla (2005), many rates of interest used by quantitative economic historians were not nominal rates specified in contract. They are not interest rates that contemporaries saw as such. They are estimations based on assumptions. Many innovative methods were applied to calculate yields when there was no information on prices or interest rates (Conrad and Meyer 1958; Davis 1965; Sussman and Yafeh 2006). Such methods created – and still create – many debates on the potential limitations of such computations and their underlying assumptions. As we will see in this chapter, this generated many controversies between cliometricians. Using historical interest rates led to provocative and fruitful research, but, in some areas, little consensus has been reached on how to construct and interpret the level and movements of interest rates in light of price theory.

Much less debated in the cliometric literature have been cases where interest rates – found in historical archives or calculated ex post – cannot be used to infer information about behavior or markets, because rates were not market prices. The practice of usury rates in the early modern and modern period has received a lot of attention in the historical literature (Hoffman et al. 2000; Temin and Voth 2005; Botticini and Eckstein 2012). This tends to obscure the fact that usury laws still exist today in many countries (not only in those where a religious law prevails) and, more generally, that many interest rates have been determined by regulation, manipulated by market markers (governments or oligopolies), or that markets cleared through quantities rather than prices. One position would be to say that, given its methodological premises, the cliometric literature has nothing to learn from fixed, nonmarket-determined interest rates. A more nuanced view is that there is still much to be learned about the functioning of an economy if one understands why some interest rates were not market-clearing prices. It is also a way to challenge earlier historical interpretations that overlooked the potential disconnection between prices (interest rates) and quantities (Hoffman et al. 2000, 2019; Temin and Voth 2005; Monnet 2014).

The Rate of Return on Investment and the Production Function

A standard way to assess profitability, or yield, in finance is to calculate the internal rate of return. This is the rate at which the total present value of the investment cost equals the total present value of future earnings. In Keynes's terminology, the internal rate of return is called the marginal efficiency of capital. For investment to take place (or to be rational), this has to be higher than the interest rate, which is the rate of return on a safe asset which can be purchased easily (i.e., a market rate). In one of the first cliometric studies published in 1958, Conrad and Meyer (1958) calculated the rate of returns on slaves in order to assess whether slavery was a profitable activity. By reconstructing investment costs and earnings of a cotton plantation in the south of the USA, they found rates that were much higher than in other activities. The issues at stakes behind such calculations were enormous and were further developed in Fogel and Engerman's landmark work, *Time on the Cross* (1974). If slavery in antebellum America was a profitable activity, then it meant that economic forces alone would not have brought slavery to an end without the necessity of war and political change.

Conrad and Meyer's study offers a prominent example of the new methods and questions, as well as new controversies, that the quantitative "new" economic history (or cliometrics) brought to history. As explained by Fogel and Engerman (1974, p. 65 et al.), discussions on the profitability of slaves had been limited by the lack of sources on the matter before the study of Conrad and Meyer. Leaving aside the question of whether slave owners themselves were making profitability calculations, Conrad and Meyer decided that "the basic problems involved in determining profitability are analytically the same as those met in determining the returns from any other kind of capital investment." They moved away from the debate framed by the accounting concept of profitability and turned to a purely economic concept of profitability, based on theoretical reasoning rather than on historical accounting practices. They derived a rate of return from estimates of the cost (investment) of slaves and their lifetime earnings. Their approach was not limited to calculating the rate of return from the investment and revenues of a typical cotton plantation. They had to compare it to alternative interest rates in order to assess whether slavery was an "efficient business." Hence, they had to find alternative rates of interest (whether commercial paper or investment in other activities) and, most of all, justify that such markets existed and were accessible to slave owners.

The underlying argument of Conrad and Meyer is that if an investment turns out to be profitable (in theory) but is not realized, it is because social or political constraints have prevented this investment. Agents are assumed to be sensitive to market incentives, and interest rates are assumed to be the price at which supply and demand meet. Then, if slavery disappeared despite being profitable, it was because of political choices. The calculations and comparisons of rates of return were, for these authors, a way to understand the determinants of historical changes. In a totally different context but in a similar perspective, Jean-Laurent Rosenthal (1990) applied the same method to examine the economic consequences of the French revolution. He calculated the rate of return of irrigation (building a canal) in

Provence and found similar values before and after the Revolution, whereas the construction of a canal started only after 1789. From this comparison of hypothetical interest rates based on financial theory, he stated that canals were not built before the revolution because political institutions prevented the transfer of property rights.

In a follow-up to Conrad and Meyer's (1958) article, Sutch (1965) described the method of the previous authors as "reconstructing the production function" of the cotton plantation. It means that the authors had to use historical sources to come up with detailed estimates of the cost of investment, as well as lifetime earnings. However, the method of calculating internal rates of return does not require the specification of a full-fledged production function formalizing how factors are combined to produce output. On the contrary, other economic historians have relied explicitly on a neoclassical production function to shed light on the relationship between the rate of return and the factors of production. The objective was not to compute an internal rate of interest based on the present value of cash flows but to express the marginal rate of return on capital, based on a production function. The theory underlying the interpretation of interest rates is thus different, although the terminology is sometimes quite similar. A typical way to proceed in this way is to take a market interest rate as given and assume that it equals the marginal return on capital in order to infer some information about the factors in the production function. In an influential but controversial paper, Peter Temin (1966) used such a method to assess whether labor scarcity (and hence high-real wage compared to the UK) was prevalent in the nineteenth-century US economy. Temin observed that market interest rates (yields on government bonds) were higher in the USA than in the UK. From this observation, making the assumption that market forces should equate different interest rates within a country, he concluded that rates of return in American manufacturing were higher than in British manufacturing. Since the rate of return of a neoclassical production function is positively related to the labor-capital ratio, Temin concluded that labor was more abundant in the USA and capital scarcer. One key assumption of Temin's result was that technologies were similar in the USA and the UK, so that labor and capital were used in the same way to produce manufacturing goods. Even if one is ready to accept this assumption, Fogel (1967) and Drummond (1967) showed that a higher interest rate in the USA is still compatible with a lower labor to capital ratio in this country, if one consider different factor prices or include land in the production functions. In Fogel's words, there was a specification problem in Temin's work: a single empirical observation was compatible with several theories. Moreover, the implicit assumption that two countries have reached their steady state is debatable.

The method of Conrad and Meyer (1958) was to calculate the internal rate of return based on estimations of capital and labor used for production, and then to compare this return to a market interest rate in order to assess efficiency. Temin's method, on the contrary, was to use the neoclassical theory of economic growth to make a statement about the relative importance of production factors, based on the observation of empirical interest rates. While no method is perfect, since they all rely on a specific set of important assumptions, the second method is more prone to

a specification or identification problem (i.e., several theoretical formulations are consistent with one empirical observation), as underlined by Fogel.

Theoretical and Effective Interest Rates

Critics of Temin's article raised an interesting question as whether market interest rates on financial assets are good proxies for the return to capital in the standard neoclassical model. Many pieces of evidence point to large discrepancies between yields on financial assets and the rate to capital that may correspond to the standard neoclassical model (Mulligan 2002). For this reason, some authors dealing with macroeconomic historical questions, like Allen (2009) and Piketty (2014), have preferred to compute profit rates from wealth estimates rather than use market interest rates in order to interpret historical economic evolution in light of growth theory.

Dealing with a more recent period, Caselli and Feyrer (2007) calculate the profit rate (marginal product of capital) for a panel of countries based on estimates of total income and the capital stock. They explain that comparing market interest rates at the international level is not informative and problematic "because in financially repressed/distorted economies, interest rates on financial assets may be very poor proxies for the cost of capital actually borne by firms" (p. 536). Many economic historians had already warned us not to interpret historically observed (real) market rates as the rate of return on capital in neoclassical growth theory (Harley 1977). Moreover, Harley's contribution to the debates around the Gibson paradox in the nineteenth-century Britain (i.e., the unexpected correlation between interest rates and the price level, rather than the rate of inflation) recalled that the assumed positive relationship between nominal interest rates and the rate of inflation is not warranted. This cast doubts on the relationship that economists usually expect between nominal and real variables.

Barsky and Summers (1988) proposed to solve the Gibson paradox by claiming that during the gold standard, the real interest rate was in fact determined by the relative price of gold. Since the writings of Keynes, the Gibson paradox has been seen as a prominent example showing that taking for granted some standard assumptions of economic theory about interest rates may lead to biased interpretations. It has been argued instead that economic theory should not be taken as a basis for interpretation but should be refined to provide predictions that are consistent with the historical evolution of interest rates. More generally, there are many monetary, financial, or institutional factors that may explain the differences between the observed real rate of return on financial assets and the return to capital derived from a standard neoclassical production function in an environment of perfect capital mobility. Some recent estimates of return on real and financial assets show that they often substantially differ (Jorda et al. 2017). It is therefore an open question as to whether the rates of return on financial assets can be interpreted in the light of neoclassical growth theory, and it is fair to say that no consensus has been reached by economic historians on this issue.

In their standard history of interest rates, Homer and Sylla (2005, p. 10) acknowledged the lack of agreement on the matter: "It is not the purpose of this book to analyze the causes of interest-rate levels and trends. There is a vast literature on this subject but little area of agreement." Partly for this reason, we shall see in the next sections that much of the cliometric literature has not focused on explaining the level of long-term interest rates but has been interested in explaining the short- or medium-term variations of these rates (or differences in rates) with the aim of clarifying historical market characteristics or agents' reactions to political or economic changes. Building new theories to explain the historical evolution of interest rates in the long run has not been the main objective of cliometricians. Consistent with McCloskey's (1978) principles, they are more inclined to use available standard interest rate theories to shed light on historical episodes.

Market Interest Rates: Sources and Calculation Methods

At first sight, it may seem much easier to calculate a market interest rate than the internal rate of return or the marginal rate of capital. Computations of the internal rate of return and the profit rate rely on sometimes-crude estimates of the values of investment, future earnings, or production factors. On the contrary, the market gives us a direct estimate of the value of an asset at a given time. The current yield is simply equal to the investment's annual income (interest, coupon, or dividends) divided by the current price of the security, as quoted on the market. It would be wrong, however, to think that this computation is an easy task. All textbooks in investment finance contain endless pages on how to calculate yields, depending on the assumption that are made on whether the bond is held toward maturity, whether investment returns are compounded, etc. In their introduction, Homer and Sylla (2005) give key definitions and highlight important pitfalls in the calculus of market interest rates. Many economic historians have formulated similar reflections, including Klovland (1994) in an important paper about the yield on British consols in the nineteenth century. Since the interest rates used by historians are often yields on government bonds (the yield on British consols being a prominent example) because these assets were traded continuously for a long time, debates on the methods to calculate the yields have received a lot of attention in the literature, and the pitfalls should be known to users. A conceptual clarification is first necessary. A distinction has to be made between what Homer and Sylla call the "nominal interest rate," which is the rate specified in the loan contract (the interest expressed as a percentage of the nominal, face, or par value of the loan), and the "market yield," which is the rate of return to the buyer at the market price. It is impossible to enter here into all details and pitfalls of the calculation of yield. However, it is enough to emphasize just how sensitive yield calculations are to the seemingly minor institutional details of price quotation and to assumptions about the date of redemption of the bonds (Klovland 1994).

The computation of average market yields relies on a chosen set of bonds. Issues of averaging and the choice of relevant bonds arise, creating considerable discussions for both market investors and historians (Homer and Sylla 2005).

Short-term interest rates on government bills do not face the same issues of redemption as do long-term interest rates. However, there are still issues regarding the choice of the quoted purchase price of the bill as well as the best method to calculate the yield. We have mentioned earlier formulas to calculate yields on long-term bonds that offer a coupon, but many short-term bills do not. Such bills, like the US treasury bills, are issued at a discount from par value. For such cases it is debatable whether to use the discount yield formula [((par value – purchase price)/par value) ∗ 360/days to maturity)] or the investment yield method [((par value – purchase price)/purchase price) ∗ 360/days to maturity)]. For all these methods to be reliable, the market needs to be sufficiently liquid for prices to be quoted with regular frequency. This is why we have so few long-run yield series. The exceptions are government securities in major countries with a stable government and well-developed financial markets since the nineteenth century. Thus, the available long-term series of interest rates on financial assets have a strong selection bias (Homer and Sylla 2005; Jorda et al. 2017): they exist only when financial markets were sufficiently developed and liquid. Little is known about lending practices and contracts in less developed markets.

Money market (interbank) rates are another type of animal. Since the interbank market is an over-the-counter market, there is no quoted price. The rates at which funds were borrowed are known from declarations of market participants, as for the LIBOR (London Interbank Offered Rate) today. For historians using such rates, it is not always easy to identify the types of assets that were traded on the market and the actors involved. For example, the widely used interest rates published by the British newspaper, *The Economist*, in the nineteenth century, were limited to a premium market whose conditions were presumably different from most credit conditions in the country (Bazot et al. 2016). The recent dataset of Mitchener and Weidenmier (2015) reminds us that short-term interest rates (either government bills or interbank markets) are available for few countries in the late nineteenth century, despite the period being quite integrated with developed capital markets worldwide. For lack of a better alternative, the discount rate of the central bank is often used as the best proxy for short-term interest rates, but this is hardly a market price. Few sources provide a call money or a repo rate.

Long-term series of bank lending (or deposit) rates do not exist. This is simply because very few countries had banking regulation before the interwar period, so that no central authority collected and registered the interest rates applied by banks. Banks themselves did not make public a series of their average interest rates. Even in the USA – which is a notable exception because banking regulation existed as early as the nineteenth century, meaning that banks had to send their balance sheets to the regulator – such data do not exist before the second half of the twentieth century. Yet, armed with unique bank balance sheet data that have no equivalent anywhere in the world, US scholars have been able to compute proxies of interest rates for the postbellum era by dividing bank earnings by their stock of earning assets (Davis 1965). In other countries, such as England, estimates of bank lending rates have often been based on few case studies (see Temin and Voth 2005 for a review).

Market Integration and Market Risk: Differences Between Rates in Several Areas

Collecting and estimating interest rates is not an easy task. But this hard work often generated landmark studies that shed a new light on historical debates. An influential study in this area was the article by Lance Davis (1965), which estimated regional interest rates for the USA from 1870 to 1914 and showed a striking and persistent divergence between these rates until the war. The evidence of such low integration in US capital markets led to countless studies trying to understand why it was the case and how growth could occur despite low integration (see Landon-Lane and Rockoff 2007 for a review).

Many authors have looked at interest rates differentials in order to study the integration of national or international capital markets. At the international level, spreads between market interest rates (when available) or estimations of profit rates (marginal return to capital) have also been used as a measurement of financial integration and its evolution over time (e.g. Obstfeld and Taylor 2004; Flandreau et al. 2009; Caselli and Feyrer 2007).

One problem with comparing interest rates or average returns between different geographical areas is that it is not always possible to compare interest rates on assets with the same risk, as Eichengreen (1984) showed in his study of mortgage lending in the USA. And when this is possible, it is not always enough to assess financial integration. McCloskey (1970) underlies the fact that British rates on safe bonds (railway or government) were close to rates abroad with the same risk, but overall, the average rate of return perceived by British investors on British assets in 1911–1913 was twice as high as the return on capital abroad. McCloskey recalls that contemporaries had already noted such discrepancy and had not interpreted it as the result of a bias toward safer assets abroad but as a deliberate policy of the financiers of the City of London to favor foreign capital and maintain high rates in the country. She quotes J.M Keynes writing that the effect London's investment policy was "to starve home developments by diverting savings abroad and, consequently, to burden home borrowers with a higher rate of interest than they would need to pay otherwise" (McCloskey 1970, p. 452). McCloskey assumes that the maximum spread between foreign and domestic rates of return could be a measure of market imperfection created by the City of London. She calculates whether, in this case, British national income could have been much higher without such an imperfection. She finds that, had imperfections been smaller, the growth of British national income would not have been much higher. Thus, she refuted the argument that slow growth in Britain during the Victorian era could be attributed to the outflows of capital to other countries.

Besides assessing whether capital markets were efficient and estimating the degree of integration, differences between interest rates of several countries have also been used to discuss the autonomy of domestic monetary policy relative to the world. Full integration of international financial markets would mean that countries have no ability to set their interest rates independently from others. As is the case with regional interest rates, there were in fact differences in worldwide interest rates,

either because of voluntary or involuntary capital market imperfections. Economic historians are interested in understanding what these differences reveal about the geopolitics of finance. Morys (2013) showed that the discount rates of central banks in Eastern Europe under the gold standard were influenced by discount rates in Berlin and London to a similar degree. This finding suggests that the functioning of the gold standard was less London-centered than had been hitherto assumed and that Germany was the policy reference for most countries in the "periphery." Obstfeld and Taylor (2004) interpreted the spreads between the domestic interest rates and an index of several leading international rates as a measure of the constraint of international finance (the "trilemma"). A high spread is evidence of higher autonomy. They looked at how this autonomy varied over time and depended on the international monetary regime.

An influential study by Bordo and Rockoff (1996) looked at the spread between the domestic long-term interest rate in several countries and the British rate during the gold standard. They showed that gold standard adherence lowered the spread between domestic and British rates, and they interpreted this finding as evidence that the gold standard was a signal of financial rectitude, a "good housekeeping seal of approval," that facilitated access by peripheral countries to capital from the core countries of Western Europe. A large literature followed from this article (enlarging the sample of countries and using different definitions of interest rates) whose common point has been to discuss the potential benefits of gold standard adherence by looking at spreads between the domestic interest rate and the leading international rate (see Alquist and Chabot 2011; Mitchener and Weidenmier 2015; Chavaz and Flandreau 2017 for recent additions and a literature review). Bordo and Rockoff's argument was international in nature but was influenced by neo-institutionalist theory that discussed how a change in institutional settings (the gold standard in this case) affected the cost of government borrowing by creating a credible commitment to financial and fiscal rectitude ("good housekeeping seal of approval" in this case). The landmark paper in this literature was North and Weingast (1989), to which we now turn our attention.

Interest Rates and Political Regimes

In a controversial and groundbreaking paper, North and Weingast (1989) famously claimed that the new institutional setting arising from the 1688 revolution in England allowed the government to commit credibly to upholding property rights and that this made the financing of the public debt cheaper and contributed to the development of private markets as well. They used interest rates (on public debt) as a way to assess the influence of institutions on the government's ability to borrow and to discuss the benefit of institutional reforms.

Few historians still accept the conclusions of North and Weingast (Coffman et al. 2013), but the paper remains influential in shaping a new approach to examining interest rates and interpreting them in the light of neo-institutional theory. The North and Weingast thesis was attacked for many reasons. Some argued that there was no

evidence that the 1688 revolution guaranteed property rights and that it was mainly because of the parallel development of financial markets that sovereign debt became a safe and liquid investment vehicle with a low interest rate (Coffman et al. 2013). Others went on to build new interest rate series to challenge North and Weingast's conclusions about a decline in government bond yields after the 1688 revolution (Sussman and Yafeh 2006).

There was no high frequency listed price of government bonds for this period, so the estimates rely on fragmentary data. To overcome this limitation, Sussman and Yafeh (2006) calculated the cost of British government debt as the ratio of debt service payments to total government debt (dividing government debt service expenditures by a series of the stock of total debt). They found that the interest rate on British debt remained high for several decades after the 1688 revolution, especially in relation to the Dutch interest rate. Based on various sources from private financial institutions, several other authors have also found no change in interest rates after the revolution (see Temin and Voth 2005; Coffman et al. 2013 for a review).

Not all economic historians share the neo-institutionalist perspective of North and Weingast on the link between the cost of sovereign debt and the "quality" of institutions. But it has become a common practice to look at changes in interest rates to interpret the effects of political events. Willard et al. (1996) look at the Greenback market during the US Civil War and use econometric techniques to identify breaks in the price. The USA issued an inconvertible currency called the Greenback starting in 1862. Its value in gold fluctuated over time, reflecting the expectation of future war costs. The authors use data on the gold price of Greenbacks and compare the reactions of participants in financial markets to significant military events during the Civil War. Their findings highlight the importance of some events (especially expectations of victories and end of the war) that had not been viewed as crucial by historians, but seemed to have created major changes in the expectations of contemporaries.

Ferguson (2006) studied the behavior of interest rates of long-term bonds quoted in London between 1848 and 1914 and found that political events had a much smaller effect on the bond market after 1880. He interpreted this result as the consequence of the deepening of national and international financial markets in the late nineteenth century. Most of all, he showed that the outbreak of the First World War was not anticipated by the market. He used this result to challenge the traditional view, which tended to over-determine the beginning of the war.

In a similar vein, Hautcoeur and Sicsic (1999) examined the interest rate of French and foreign bonds in interwar France to shed new light on the monetary unrest and political troubles of this period. They showed that two types of information can be extracted from interest rates: expectations of taxation (a capital levy should lead to a drop in the price of taxed capital assets) and expectations of a devaluation (through the price of long-term bonds whose coupons were indexed on the pound/franc exchange rate). Expanding the work of Klovland (1994), they also emphasize the role of the political context and the importance of assumptions on the risk of redemption and conversion to provide meaningful calculus of the interest rate on perpetual bonds.

These three papers made different uses of interest rate series and discussed in a different way the potential limitations of the calculus of interest rates, but they shared the premise that expectations of major political events are reflected in market prices. Looking at interest rates in this way is a novel method of assessing what contemporaries really thought about the likelihood and importance of political events. This method is more precise than looking at other sources – such as the press or parliamentary debates – which are difficult to exploit in a comprehensive way. Since the interpretation relies crucially on the assumption of efficient markets, the articles quoted above devoted a lot of attention to justify that the market they were looking at was efficient, in the sense that they were liquid and without barriers to entry.

Interest Rates, Financial, and Macroeconomic Cycles

The link between politics and interest rates has not just been studied through the lens of institutions and expectations. The interest rate is a key variable in any macroeconomic model and, as such, deserves attention to understand macroeconomic fluctuations and their mechanisms. It is standard for macroeconomists to look at whether interest rates respond to economic shocks in a way which is consistent with theory and to model monetary shocks as an increase in the interest rate. These common questions and methods have been applied to historical data to inform both economic modeling and our understanding of the past. Robert Barro (1987) studied the response of interest rates to changes in government spending in the nineteenth-century England. His main motivation was to test for Ricardian equivalence, which predicts that an increase in public debt will be matched by a higher present value of future taxes, and thereby has no effect on desired national saving or interest rates. Barro's results show that interest rates rose after increases in spending only during wars, but not during the few episodes where budget deficits were caused by other reasons (compensation payments to slave owners in 1835–1836 and political disputes over the income tax in 1909–1910). During the gold standard, there was no relationship between monetary growth and changes in government spending.

Economic historians have also used the econometric methods of macroeconomists and references to theory to study the impact of central bank decisions on the economy ("monetary policy shocks"). Monnet (2014), Bazot et al. (2016), and Lennard (2017), for example, applied standard VAR (vector autoregressions) to assess the impact of changes in the central bank discount rate during periods whose institutional features of central banking were very different from what we know today. One advantage of these methods is to provide an estimation of the share of variance of main macroeconomic or financial variables, which is explained by changes in the discount rate of the central bank.

The use of recent econometric techniques has helped economists and economic historians to better understand the channels of transmission of main fluctuations and to assess the importance and impact of policy choices in these

fluctuations. Although short- or medium-term macroeconomic cycles have attracted much attention, seasonal variations have not been left aside. As with many variables reflecting economic activity, interest rates are often seasonal. Seasonality was particularly strong when agriculture accounted for a large share of national income and credit increased during the storage period of crops. This well-known pattern has been the topic of numerous studies (Homer and Sylla 2005). It was also recognized by contemporaries and, in some cases, generated policy interventions of governments or central banks, whose objective was to smooth and harmonize rates within a country (Miron 1986; Bazot et al. 2016). Miron (1986) shows that the size of the seasonal movements in nominal interest rates declined substantially after the US central bank – the Federal Reserve System – was established during the First World War. The Fed conducted seasonal open market policy to eliminate the seasonality in interest rates. These Fed operations not only mitigated the seasonality in interest rates but also the frequency of financial panics, which were themselves partly caused by the seasonality of interest rates. Bazot et al. (2016) pointed out that the Bank of France – the French central bank – was smoothing seasonal fluctuations in the interbank market. In addition, they provided evidence that central bank interventions were partly absorbing the effects of international financial shocks on the French money market rate.

The link between interest rates and financial panics has also been the topic of considerable interest. One key question is whether cheap credit or loose credit conditions – reflected in interest rates – caused a financial bubble and then a crash. In a paper entitled "Was There a Bubble in the 1929 Stock Market?", Rappoport and White (1993) used interest rates as a test to answer this question. They look at interest rates for brokers' loans, which investors used to fund stock purchases. During the 1928–1929 stock market boom, lenders required a large premium in this market (over other money market rates) because they thought that stock prices might collapse during the term of a loan and jeopardize the collateral. They interpret this finding as evidence of a bubble in the stock market at this time, which was in fact expected by some market investors.

At first sight, the studies on interest rates and the macroeconomy may seem somewhat more agnostic about the efficiency of financial markets than the studies which merely aim to extract relevant information from market prices. Yet, the macroeconomic studies still assume that interest rates are prices that clear markets and, thus, that they reflect changes in demand and supply. When they do not share the neoclassical view of interest rates as the price that equals savings and investment, they at least subscribe to Keynesian principles that interest rates reflect demand and supply for liquidity. Such an assumption is usually made about the central bank leading interest rate, so that the central bank would move the quantities necessary to reach the target rate, and agents would adjust their behavior to that rate. These assumptions may be valid in some contexts, but certainly not at anytime and anywhere. Thus, to avoid unduly narrowing their field of study, economic historians have had to deal with situations where interest rates provide little information on the functioning of the economy. The last part of the survey is devoted to research in this area.

When Interest Rates Do Not Clear Markets

Hoffman et al. (2000) chose a highly revealing title for their groundbreaking study of private credit markets in Paris over two centuries (the mid-seventeenth to mid-nineteenth centuries): *priceless markets*. The title follows from the finding that client-specific information – and thus quantity rationing – was far more important than price in clearing the credit market. Interest rates did little to allocate capital or inform participants in the credit market. Such findings were based on a wide sample of private and public loan contracts taken from Parisian notarial records. The study was then extended at the national level in Hoffman et al. (2019). Notaries were the most important financial intermediaries during this period. Loans were usually granted at a rather similar rate, whatever the maturity and the risk. But willingness to lend and non-price conditions of the loans reflected characteristics of the borrowers. Although usury laws (until the French revolution) and later on interest rate regulations were an incentive to keep stable interest rates, the authors argue that the practices of notaries and the informational structure of the market mainly explain the insignificance of interest rates. This conclusion about the unimportance of interest rates for financial markets was based on microeconomic evidence and a detailed study of the market infrastructure and bears important consequences for macroeconomic or political economy analyses. It means that we only observe a stable interest rate, from which it is impossible to infer information about demand and supply. Hence, there is no single average interest rate that could be interpreted as the price of capital in the economy, and there is no interest rate differential that could be interpreted as evidence of borrower selection or imperfect capital integration. Was there a price – any other interest rate – to which we can relate the amounts of credit observed on the Parisian credit market? Could we say, for example, that notarized loans were not offset by interest rates, but that a market interest rate from other financial transactions was nevertheless a good indicator of credit conditions in the private credit market mediated by notaries? The authors say no. They examined the available short-term and long-term government bond yields over this period and concluded that "it is remarkable how little a connection there was between interest rates and the state of the economy" and "following the variations of either the long or short-term interest rate series is thus unlikely to tell us much about the scarcity of credit except when the scarcity is driven by politics" (Hoffman et al. 2000, p. 43).

The unimportance of interest rates was not specific to French notaries. Temin and Voth (2005) challenge the premise of the debate that followed North and Weingast (1989) (see supra) by claiming that interest rates in the seventeenth- and eighteenth-century England were not indicative of credit conditions. Not only did Temin and Voth (2005) address the debate on the cost of financing after the 1688 revolution but they also joined in the discussion of crowding out during the British Industrial Revolution. A common explanation for slow growth during the eighteenth and nineteenth centuries is that wartime financing crowded out private investment, which should have translated into higher interest rates in wartime. Temin and Voth (2005) first review the work of previous scholars and conclude that, in both debates (effects of the 1688 institutional change and crowding out), evidence based on

interest rates has been scarce and contradictory. According to them, using interest rates to solve these debates is illusory because "private-sector interest rates are not the right indicator of scarcity in the case of 18th-century finance – for both practical and conceptual reasons. In contrast to good markets, where price is an efficient way of allocating scarce goods, credit markets rarely reach equilibrium through changes in interest rates alone" (Temin and Voth 2005, p. 326). Rationing occurred because of usury laws and because of asymmetric information, which pushed financial intermediaries to discriminate by means other than interest rates. Borrowers willing to pay very high interest rates were inherently bad risks, so banks needed to find other margins on which to allocate credit. Based on the analysis of one English bank (Hoare's bank) whose balance sheets and archives have been kept, the authors show that 92% of all loans were made against interest at the usury limit. In such conditions, a negative shock to the private credit market (such as the issuance of government debt, which reduced bank deposits and the ability of banks to lend) was not reflected in higher interest rates.

Nonmarket-clearing interest rates are not specific to usury laws of early modern Europe. It has been typical of government interventions to attempt to smooth fluctuations of interest rates and maintain them at a low and stable level. This can be accomplished either by regulation or financial markets interventions that manipulate the level of market interest rates. Such policies can achieve several objectives, such as providing a low cost of financing for the government (Reinhart and Sbrancia 2015), harmonizing interest rates across the country, allowing individuals to borrow at cheap rates, or protecting savings against inflation. The central bank is likely to play an important role in maintaining low and stable interest rates in a way that differs from what would have been the market outcome (Miron 1986; Bazot et al. 2016).

In economies with high financial regulation and capital controls – which have been widespread throughout the twentieth century, including in non-Communist countries (Reinhart and Sbrancia 2015) – the central bank typically maintains stable interest rates and use quantitative rationing (credit controls, reserve requirements, etc.) to fight inflation. Studying monetary policy in postwar France (1945–1973), Monnet (2014) showed that the discount rate of the central bank provides a misleading indicator of the stance of monetary policy. The French central bank did not move interest rates. Rather, they usually relied on credit ceilings (limits on bank credit growth) to limit credit creation and curb inflation. Such central banking practices took place in a planned economy with segmented credit markets where many sectors benefited from subsidized loans, such that interest rates were not market-clearing prices. It was also consistent with the objective of the government to maintain low and stable interest rates. The use of quantitative credit controls led to a disconnection between the level of interest rates and the overall monetary policy stance (or credit conditions).

Since many interest rates were regulated and credit markets were segmented, a tight credit policy did not lead to higher market interest rates. Hence, looking at interest to assess the impact of monetary policy leads to a mismeasurement of the policy stance and misleading results. Acknowledging such central bank strategies is also crucial to interpret gaps between interest rates at the international level. Central

banks could assign the interest rate to the external side while managing domestic credit expansion with direct quantitative controls. As a result, a low spread between national interest rates cannot be interpreted as a lack of autonomy of monetary policy (Monnet 2018). The method of Obstfeld and Taylor (2004), which looks at interest rate differentials across countries to assess the autonomy of domestic policy, is only valid when the interest rate is the main policy instrument on the domestic side. There is much historical evidence that over the long run this was not always the norm, especially outside of the UK and the USA.

The various studies mentioned in this section criticize the common use of interest rates by economic historians. They point out that either because of imperfect information, government policies, or institutional rules and norms, interest rates were rarely revealing anything other than market imperfections. They were not a market-clearing price. If this is the case, interest rates cannot be used to assess the market integration, the autonomy of domestic policy, the impact of institutional reforms, etc. Earlier, I quoted the work of economists (Caselli and Feyrer 2007) who decided to estimate the marginal return to capital in order to assess the efficiency of capital allocation at the international level, because market interest rates were deemed imperfect indicators of actual credit conditions. "Because in financially repressed/distorted economies, interest rates on financial assets may be very poor proxies for the cost of capital actually borne by firms" (Caselli and Feyrer 2007, p. 536).

What economic history has recently shown is that repression and financial distortions were indeed so prevalent in the past – and still are today in many countries – that one should be very cautious in drawing conclusions about market returns or interest rates set by financial intermediaries, governments, or central banks. However, these conclusions are not all negative. Assessing the unimportance of interest rates provides important historical information on how markets functioned and how government interventions took place in the past. It also pushes researchers to develop innovative ways of circumventing the absence of significant changes in interest rates. However, since it is no longer possible to infer market information from a single price, researchers must look for data on quantities. Such work is often much more data intensive and time consuming (Hoffman et al. 2000, 2019; Temin and Voth 2005; Monnet 2014).

Conclusion

From Conrad and Meyer's (1958) study on the rate of return of slaves, to Davis (1965) on the integration of financial markets in the USA, to North and Weingast (1989) on the 1688 revolution in England, interest rates have been central to some of the most influential cliometric studies. As McCloskey (1978) points out, the theory of price was the defining skill of early cliometricians. The interpretation of historical interest rates on the basis of standard price theory has been a constituent element of cliometrics. This method was widely and provocatively applied in order to attack traditional history, arguing that the sources of earlier work had been too scattered and

biased to provide sound economic conclusions. Considering the interest rate as a market price was one way to draw more rigorous conclusions from sometimes-Spartan historical sources. Since cliometric studies were – and still are, to a large extent – focused on the UK and the USA in the modern period, the assumptions underlying standard price theory were often seen as valid. The yield on British consols has probably been the most observed series of interest rates in history and epitomizes what is a market price in well-developed financial markets.

Cliometric studies of the rate of interest have not been without criticism. Questions were raised about the most relevant interest rate to consider, how to calculate yields on financial assets, and how to relate observed prices to the neoclassical theory of growth, which Fogel (1967) saw as a prime example of the specification problem in economic history. It is only quite recently that some studies have begun to discuss nonmarket-clearing interest rates and incorporate them into a quantitative perspective of economic history. The common feature of these various studies is that they declare that some significant historical interest rates cannot be interpreted as market prices. It is therefore not possible to obtain precise information on preferences, demand, and supply from interest rates. Therefore, it is necessary to go beyond interest rates and accumulate data and information on quantities, since markets did not clear through prices. Some of these studies directly question the earlier results of cliometricians, such as the possibility of assessing the impact of institutional changes on financial markets (Temin and Voth 2005) or of measuring monetary policy autonomy by examining the gap between national and international rates (Monnet 2018). Yet their main message is not to get rid of price theory and interest rates in economic history. Some interest rates were undoubtedly market prices that can be considered as such, but, in the end, they could be exceptions rather than a rule. Looking at how interest rates were set by private actors and governments out of the market – and what were the economic effects of such manipulation – is an area of further research.

References

Allen RC (2009) Engels' pause: technical change, capital accumulation, and inequality in the British industrial revolution. Explor Econ Hist 46(4):418–435

Alquist R, Chabot B (2011) Did gold-standard adherence reduce sovereign capital costs?. J Monet Econ 58(3):262–272

Barro RJ (1987) Government spending, interest rates, prices, and budget deficits in the United Kingdom, 1701–1918. J Monet Econ 20(2):221–247

Barsky RB, Summers LH (1988) Gibson's paradox and the gold standard. J Polit Econ 96(3):528–550

Bazot G, Bordo MD, Monnet E (2016) International shocks and the balance sheet of the Bank of France under the classical gold standard. Explor Econ Hist 62:87–107

Bordo MD, Rockoff H (1996) The gold standard as a "good housekeeping seal of approval". J Econ Hist 56(2):389–428

Botticini M, Eckstein Z (2012) The chosen few: how education shaped Jewish history, 70-1492. Princeton University Press, Princeton

Caselli F, Feyrer J (2007) The marginal product of capital. Q J Econ 122(2):535–568

Chavaz M, Flandreau M (2017) "High & dry": the liquidity and credit of colonial and foreign government debt and the London Stock Exchange (1880–1910). J Econ Hist 77(3):653–691
Coffman D, Leonard A, Neal L (2013) Questioning credible commitment: perspectives on the rise of financial capitalism. Cambridge University Press, Cambridge, UK
Conrad AH, Meyer JR (1958) The economics of slavery in the ante bellum South. J Polit Econ 66(2):95–130
Davis LE (1965) The investment market, 1870–1914: the evolution of a national market. J Econ Hist 25(3):355–399
Drummond IM (1967) Labor scarcity and the problem of American industrial efficiency in the 1850's: a comment. J Econ Hist 27(3):383–390
Eichengreen B (1984) Mortgage interest rates in the populist era. Am Econ Rev 74(5):995–1015
Ferguson N (2006) Political risk and the international bond market between the 1848 revolution and the outbreak of the First World War. Econ Hist Rev 59(1):70–112
Fisher I (1907) The rate of interest: its nature, determination, and relation to economic phenomena. The MacMillan Company, New York
Flandreau M, Galimard CJ, Jobst CC, Nogues-Marco P (2009) The bell jar: Commercial interest rates between two revolutions, 1688–1789. The Origins and Development of Financial Markets and Institutions. From the Seventeenth Century to the Present, Cambridge University Press, Cambridge, UK, pp 161–208
Fogel RW (1967) The specification problem in economic history. J Econ Hist 27(3):283–308
Fogel R, Engerman S (1974) Time on the cross, 2 vols. Little, Brown and Co, Boston
Harley CK (1977) The interest rate and prices in Britain, 1873–1913: a study of the Gibson Paradox. Explor Econ Hist 14(1):69
Hautcoeur PC, Sicsic P (1999) Threat of a capital levy, expected devaluation and interest rates in France during the interwar period. Eur Rev Econ Hist 3(1):25–56
Hoffman PT, Postel-Vinay G, Rosenthal JL (2000) Priceless markets: the political economy of credit in Paris, 1660–1870. University of Chicago Press, Chicago
Hoffman PT, Postel-Vinay G, Rosenthal JL (2019) Dark matter credit: the development of peer-to-peer lending and banking in France. Princeton University Press, Princeton
Homer S, Sylla R (2005) History of interest rates, 4th edn. Wiley, Hoboken
Jordà Ò, Knoll K, Kuvshinov D, Schularick M, Taylor A (2017) The rate of return on everything. NBER Working Paper No. 24112
Klovland JT (1994) Pitfalls in the estimation of the yield on British Consols, 1850–1914. J Econ Hist 54(1):164–187
Landon-Lane J, Rockoff H (2007) The origin and diffusion of shocks to regional interest rates in the United States, 1880–2002. Explor Econ Hist 44(3):487–500
Lennard J (2017) Did monetary policy matter? Narrative evidence from the classical gold standard. Explor Econ Hist 68:16–36
McCloskey DN (1970) Did Victorian Britain fail? Econ Hist Rev 23(3):446–459
McCloskey DN (1978) The achievements of the cliometric school. J Econ Hist 38(1):13–28
Miron JA (1986) Financial panics, the seasonality of the nominal interest rate, and the founding of the Fed. Am Econ Rev 76(1):125–140
Mitchener KJ, Weidenmier MD (2015) Was the classical gold standard credible on the periphery? Evidence from currency risk. J Econ Hist 75(2):479–511
Monnet E (2014) Monetary policy without interest rates: evidence from France's Golden Age (1948 to 1973) using a narrative approach. Am Econ J Macroecon 6(4):137–169
Monnet E (2018) Credit controls as an escape from the trilemma. The Bretton Woods experience. Eur Rev Econ Hist 22(3):349–360
Morys M (2013) Discount rate policy under the classical gold standard: Core versus periphery (1870s–1914). Explor Econ Hist 50(2):205–226
Mulligan CB (2002) Capital, interest, and aggregate intertemporal substitution (No. w9373). National Bureau of Economic Research, Cambridge, MA

North DC, Weingast BR (1989) Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century England. J Econ Hist 49(4):803–832
Obstfeld M, Taylor AM (2004) Global capital markets: integration, crisis, and growth. Cambridge University Press, Cambridge, UK
Piketty T (2014) Capital in the 21st century. Harvard University Press, Cambridge, MA
Rappoport P, White EN (1993) Was there a bubble in the 1929 stock market? J Econ Hist 53(3):549–574
Reinhart CM, Sbrancia MB (2015) The liquidation of government debt. Econ Policy 30(82):291–333
Rosenthal JL (1990) The development of irrigation in Provence, 1700–1860: the French revolution and economic growth. J Econ Hist 50(3):615–638
Sussman N, Yafeh Y (2006) Institutional reforms, financial development and sovereign debt: Britain 1690–1790. J Econ Hist 66(4):906–935
Sutch R (1965) The profitability of ante bellum slavery: revisited. South Econ J:365–377
Temin P (1966) Labor scarcity and the problem of American industrial efficiency in the 1850's. J Econ Hist 26(3):277–298
Temin P, Voth HJ (2005) Credit rationing and crowding out during the industrial revolution: evidence from Hoare's bank, 1702–1862. Explor Econ Hist 42(3):325–348
Willard KL, Guinnane TW, Rosen HS (1996) Turning points in the civil war: views from the greenback market. Am Econ Rev 86(4):1001–1018

The Great Depression in the United States

Christopher Hanes

Contents

C. Hanes (✉)
Department of Economics, State University of New York at Binghamton, Binghamton, NY, USA
e-mail: chanes@binghamton.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_40

Abstract

This chapter describes and explains the course of aggregate real activity and inflation in America from 1929 through 1937–1938 within the theoretical framework of "new Keynesian" models with "financial market frictions." Those models point to plausible causes for the initial downturn, the depth of the subsequent decline in real activity, and the path of inflation and real wages over 1929–1933. One key factor was Federal Reserve interest-rate policy as constrained by the "zero bound" on nominal interest rates and adherence to the international gold standard. Another was the development of a massive financial crisis in which the Federal Reserve System failed to act as lender of last resort, mainly because Fed policymakers did not believe a lender of last resort was needed. The course of the economy after 1933 presents some open questions, especially about the extremely high inflation rates observed over 1934–1937 and causes of the second downturn in 1937–1938.

Keywords

Great depression · Business cycles · New keynesian models

Introduction

In the United States, most people use the term "Great Depression" to refer to a long stretch of depressed real activity that began in late 1929 and did not end until after the outbreak of the Second World War. Figure 1 plots estimates of real GDP and unemployment over 1928–1939 (annual) along with the Federal Reserve Board's index of industrial production (monthly). Vertical lines mark business cycle turning points (NBER chronology). The Depression began with a downturn from a cyclical peak in August 1929 to a trough in March 1933. A brisk expansion followed, but before employment and output had recovered to their pre-1929 trends, there was another downturn, from a peak in May 1937 to a trough in June 1938. These were two distinct business cycles and appeared as such to contemporaries. They referred to the first as "the Depression" and to the second as "the Recession." But in American historical memory, the two cycles together constitute one big slump. In the historical memory of economists, such as it is, the American Great Depression is a case study in bad monetary and fiscal policy. Economists of our day often blame the Depression mainly on decisions of policymakers in the Federal Reserve System. In 2002, Ben Bernanke, then a Federal Reserve Board Governor, half-jokingly took responsibility for the Great Depression on the part of the Federal Reserve System: "we did it. We're very sorry...but we won't do it again" (Bernanke 2002).

In this chapter I review economic literature on the Great Depression to explain the course of aggregate real activity and inflation in the United States from 1929 through the beginning of the 1937–1938 recession. There are many excellent short overviews of the Great Depression in the United States (e.g., Romer 1993; Temin 2000). This one is different in that it is specifically directed at readers who are new to economic history but well-acquainted with current macroeconomic theory, especially "new

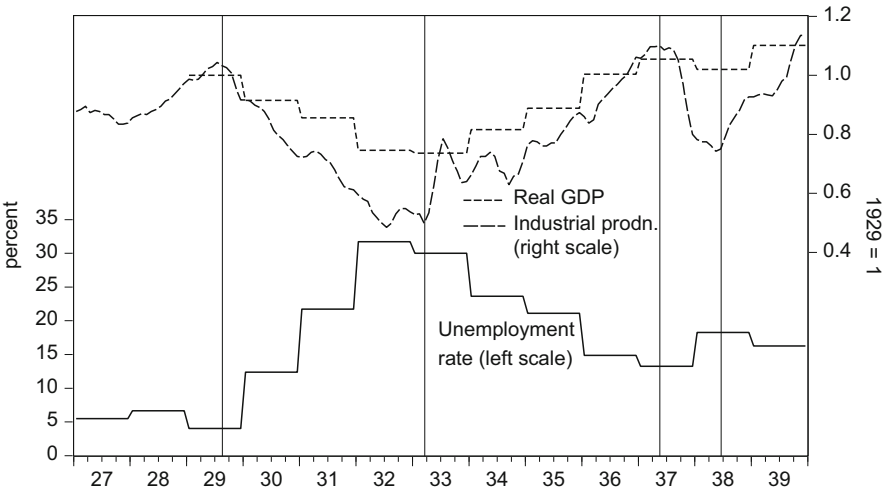


Fig. 1 Unemployment rate and industrial production index 1927–1939, real GDP 1929–1939. (Sources and notes: real GDP from United States Bureau of Economic Analysis (https://www.bea.gov/iTable/index_regional.cfm). Federal Reserve Board index of industrial production (seasonally adjusted) from Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/series/INDPRO). Private nonfarm unemployment rate from Weir 1992, Table D3. The household surveys that underlie postwar unemployment rate series were not taken in the 1920–1930s. Weir estimates annual employment from a variety of sources and defines the unemployment rate in ratio to an estimate of the "usual labor force" constructed from Census data. Weir excludes agriculture, government, and relief workers from both the employment and labor force figures. Weir excludes agriculture because, absent survey data, there is no reliable way to estimate short-term variations in agricultural employment (since so much farm labor is family labor). By excluding government workers, Weir sidesteps a debate as to whether the large numbers of Federal relief workers in the 1930s should be classified as employed or unemployed)

Keynesian" models with financial market "imperfections" or "frictions." In those models, the channel from monetary policy and financial market conditions to real activity runs through the economy's default-free real interest and/or the spread between the interest rate and the cost of borrowing to "credit-constrained" (or "liquidity-constrained") agents. I argue that it is easy to explain the 1929–1933 Depression in terms of such models. They point to plausible reasons for the beginning of the downturn and the depth of the subsequent decline in real activity. One key factor was Federal Reserve interest-rate policy, as constrained by the "zero bound" on nominal interest rates and America's adherence to the international gold standard. Another was the development of a massive financial crisis in which the Federal Reserve System failed to act as lender of last resort. New Keynesian models also account for the course of wage inflation, price inflation, and real wages over 1929–1933. About the 1933–1936 recovery and the 1937–1938 recession, the models define some puzzles and open questions.

There have been attempts to account for the Depression using "real business cycle" (RBC) models, as the result of an exogenous negative shock to production technology or unusual institutional developments over 1929–1933 that widened the

"wedge" between the marginal disutility of labor and the cost of labor to firms. I ignore these because economic historians have found no evidence for either phenomenon.[1]

With more misgiving, I also ignore monetarist explanations of the Great Depression, though monetarism frames many of the best studies of the Depression starting with Friedman and Schwartz (1963). In monetarist theory, fluctuations in aggregate demand affect real activity and inflation much as in Keynesian theory. But the key variable determining aggregate demand is not interest rates and borrowing costs as in Keynesian models. It is rather the "money supply": the quantity of currency held by the public *plus* checkable deposits *plus* any other similarly liquid assets. Monetary policy and financial market conditions affect aggregate demand through changes in this money supply (Woodford 2010, pp. 22–23). Keynesian economists complain that monetarists generally, and Friedman and Schwartz (1963) particularly, never laid out exactly what they had in mind as the transmission mechanism from the money supply to aggregate demand (Bordo and Schwartz 2004; Romer and Romer 2013).[2]

In the first section of this chapter, I review key elements of new Keynesian macroeconomic models and also models that describe the nitty-gritty mechanics

[1]Cole and Ohanian (2007) observe that conventionally measured aggregate total factor productivity fell a lot from 1929 to 1933. They argue this may represent an exogenous shock and can account for over half of the 1929–1933 decline in real GDP. Ohanian (2001, p. 37) speculates this TFP decline was due to "a decrease in organizational capital, the knowledge and know-how firms use to organize production...There are a number of reasons why this large stock of capital could have fallen, including breakdowns in relationships with suppliers that lead to changes in production plans, and breakdowns in customer relationships that lead to changes in marketing, distribution, and inventory plans." Studies that account for variations in labor and capital utilization find little or no decline in productivity over 1929–1933 (Inklaar et al. 2011, Watanabe 2016). Across the entire 1929–1939 decade, average productivity growth was remarkably rapid for reasons clearly related to the introduction and diffusion of new technologies (Field, 2011). On the labor wedge, Ohanian (2009) argues that Hoover administration policies prevented large industrial employers from cutting nominal wages after 1929. Thus price deflation, due to a drop in nominal aggregate demand, raised real manufacturing wages in a practically exogenous and historically unique way. In fact, as I will argue below, there was nothing at all unusual about the behavior of nominal wages over 1929–1933; manufacturing wages fell just as one would predict from the usual "Phillips curve" relationship. Rose (2010) looks for evidence that Hoover's policies affected the timing or magnitude of wage cuts. He finds none.

[2]Some authors (e.g., Bordo et al. 2000; Christiano et al. 2005) attempt to represent monetarist views within a new Keynesian model. In these models the interest rate that governs aggregate demand is indirectly determined by the quantity of a variable called "money" that is directly controlled by a central bank. The interest rate is related to "money" because "money" pays no interest, and the ratio of money to the price level is an argument of the representative agent's utility function. Bringing the model to data, the authors match the model's "money" to monetary aggregate statistics such as M1, which correspond (at least roughly) to the monetarist notion of the money supply. That does not do justice to the data: most assets within M1 pay interest at market-determined rates. It does not do justice to monetarists: they never argued that the money supply affected aggregate demand only (or even primarily) through interest rates on financial assets (Bordo and Schwartz 2004). And it gives a misleading notion of the way a central bank controls interest rates.

of interest-rate control by a central bank ("monetary policy implementation"). The latter are probably unfamiliar to the reader, but they are needed to understand the nature of the zero bound constraint in the Depression. In the second section, I describe peculiar institutional features of the American financial system in the 1920s, a necessary background for understanding the Depression. In the third section, I describe, narrate, and explain the 1929–1933 Depression. In the fourth I cover the 1933–1936 recovery. In the final section, I briefly discuss the 1937–1938 recession.

Theoretical Background

New Keynesian Macroeconomic Models

Consider a closed-economy "New Keynesian IS/LM" model of the simplest kind (e.g., King 2000) in which price adjustment is subject to a Rotemberg (1982) or Calvo (1983) constraint and all agents can borrow and lend at the same default-free interest rate per period. Let x^edenote today's expected value for a future variable x. The new Keynesian IS and AS ("new Keynesian Phillips curve") equations are:

$$y_t = y^e_{t+1} - \alpha r_t + \epsilon_{1t} \tag{1}$$

$$\pi_t = \delta \pi^e_{t+1} + \gamma y_t \quad \text{where} \quad 0 < \delta < 1 \tag{2}$$

where y is the "output gap," here defined as the difference between aggregate output and "flex-price equilibrium" output in the long-run steady state. ϵ_1 describes effects of temporary disturbances to preference parameters or government purchases of output. r is the spread between the short-term (one-period) real interest rate and the long-run "natural rate of interest" (the rate at which $y = y^e{}_{+1} = 0$ when $\epsilon_1 = 0$).

The model is closed with a mechanism representing the monetary regime that determines r subject to a lower bound on the nominal interest rate usually assumed to be zero – hence the "zero bound." The mechanism (usually an interest rate rule or a loss function minimized by a central bank as in Clarida et al. [1999]) must ensure the existence of a unique long-run steady state with $\pi^e = 0$. Generally:

$$y_t = \sum_{\tau=0}^{\infty} \left[-\alpha r^e_{t+\tau} + \epsilon^e_{1t+\tau} \right] \tag{3}$$

The output gap depends on the path of expected values of the short-term real interest rate from the present through the distant future. Meanwhile inflation depends on expected values of the output gap over the same horizon:

$$\pi_t = \gamma \sum_{\tau=0}^{\infty} \delta^\tau y^e_{t+\tau} \tag{4}$$

In standard models expectations are fully rational, but a weaker condition, the "law of iterated expectations," is actually sufficient for (3) and (4) (Adam and Padula 2011).

Many studies find that, in reality, expected values of future output (e.g., real GDP) generated by the most sophisticated forecasting methods are usually quite close to simple autoregressive forecasts (literature surveyed by Chauvet and Potter 2013). Thus, it is realistic to assume that the public's expected values for future output are close to AR(1) forecasts except under special circumstances. Describe the public's expected value of the output gap in a future period $t + \tau$ as:

$$y^{e}{}_{t+\tau} = \rho^{\tau} y_t + z^{e}{}_{t+\tau} \tag{5}$$

where ρ is the serial correlation coefficient from an AR(1) estimate. $z^{e}_{t+\tau}$ summarizes effects of unusual events that cause the public's forecast for $y^{e}_{t+\tau}$ to differ from the AR(1) forecast. Most of the time, $z^{e}_{t+\tau}$ must be small relative to fluctuations in y – otherwise AR forecasts would not work as well as they do. Together with (5), (3) and (4) give:

$$y_t = -\frac{\alpha}{1-\rho} r_t + \epsilon_{2t} \quad \text{where} \quad \epsilon_{2t} = \frac{1}{1-\rho}\left[\epsilon_{1t} + z^{e}_{t+1}\right] \tag{6}$$

$$\pi_t = \gamma \frac{1}{1-\delta\rho} y_t + \gamma \sum_{\tau=1}^{\infty} \delta^{\tau} z^{e}_{t+\tau} \tag{7}$$

Exogenous increases in the real interest rate should be associated with decreases in real activity. Decreases in real activity should be associated with declines in inflation. Unusual events that cause expectations of future real activity to deviate from their usual relationship with current real activity, that is, nonzero values of z^{e}, cause output to deviate from its usual relationship with the current real interest rate and also cause inflation to deviate from its usual relationship with real activity.

In more complicated new Keynesian models, there are mechanisms that create *lags* in effects of interest-rate shocks and *persistence* in disturbances to real activity, such as "habit formation" in consumption and costs of changing the rate of investment in capital (e.g., Christiano et al. 2005; Smets and Wouters 2007). The labor market is imperfectly competitive, and nominal wage adjustment is subject to the same kinds of frictions that apply to prices (following Erceg et al. 2000). Wage and price inflation can be affected by "wage markup shocks," that is, fluctuations in the spread that the wage setting process effectively seeks to maintain between wages and workers' opportunity cost of employment (Gali et al. 2012). In open economy models (Gali and Monacelli 2005), real activity is affected by foreign demand for domestic output. There can be a variable corresponding to the unemployment rate which is negatively related to output and has a long-run steady-state value corresponding to the "natural rate of unemployment" (e.g., Gali 2011). Many new Keynesian models include mechanisms to create persistence in inflation (e.g., "indexing"), but those are not needed to understand the Depression as I will explain below.

New Keynesian models can include financial market "imperfections" or "frictions." The real interest rate within r above is the return to default-free assets. With realistic constraints on contracting and "asymmetries" in information, some agents cannot borrow at that rate. If these "credit-constrained" agents can borrow at all, it is through contractual relationships with pledges of collateral to particular lenders who incur costs to collect private information about the borrower and enforce the contracts. In models, the spread between the default-free rate within r and the cost of funds to a credit-constrained borrower depends on the borrower's net worth or "wealth." A "financial intermediary" is an agent that lends to credit-constrained borrowers with funds the intermediary borrows himself at the default-free rate or through yet another set of contractual lending relationships. In the latter case, the cost of funds to a credit-constrained borrower depends not only on the borrower's own wealth but also on the net worth or "capital" of financial intermediaries (which affects the spread between the default-free interest rate and the rate *intermediaries* must pay for funds). Several models place financial market imperfections within a new Keynesian macroeconomic setting (e.g., Bernanke et al. 1999; Gertler and Karadi 2011; Curdia and Woodford 2016). In these models, financial market imperfections magnify effects of changes in r on real activity – the "financial accelerator" – and create several additional factors that can affect real activity, including current taxes ("Ricardian equivalence" fails), wealth of households or net worth of firms producing output, and financial intermediaries' capital.

Generally, more complicated new Keynesian models imply:

$$y_t = -\sum_{\tau=0}^{i} \beta_{yr\tau} r_{t-\tau} + \sum_{\tau=1}^{j} \beta_{yy\tau} y_{t-\tau} + \epsilon_{yt} \tag{8}$$

$$\pi^P{}_t = \beta_{\pi py} y_t + \epsilon_{\pi pt} \tag{9}$$

$$\pi^W{}_t = \beta_{\pi wy} y_t + \epsilon_{\pi w t} \tag{10}$$

π^P and π^W are, respectively, price and wage inflation. ϵ_y still reflects fluctuations in preference parameters or government purchases of output and unusual expectations of future real activity. But it can have many additional components: exogenous factors affecting demand for exports, taxes, and government transfer payments, wealth of households and firms producing output, and capital of financial intermediaries or anything that causes a withdrawal of lending to financial intermediaries (Woodford 2010). $\epsilon_{\pi P}$ and $\epsilon_{\pi W}$ can be affected by wage markup shocks (Gali 2011). Equations 8, 9, and 10 can alternatively be written in terms of the "unemployment gap," the difference between the unemployment rate and the natural rate of unemployment.

New Keynesian models have no necessary implication for the relative magnitude of $\beta_{\pi wy}$ versus $\beta_{\pi py}$, that is, for the cyclical behavior of *real wages* (the ratio of the wage level to the price level). Real wages may be procyclical ($\beta_{\pi wy} > \beta_{\pi py}$), countercyclical ($\beta_{\pi wy} < \beta_{\pi py}$), or acyclical ($\beta_{\pi wy} \approx \beta_{\pi py}$) depending on the relative degree of nominal rigidity in wages versus prices, among other things (Gali 2013).

Runs on Financial Intermediaries

A financial intermediary that borrows short-term and holds illiquid assets such as loans can be vulnerable to "runs." In a run, short-term lenders *to* an intermediary withdraw their loans *en masse* (or lend less to the intermediary on a given amount of collateral). To repay its creditors, the intermediary must sell off assets including the loans it has made. As assets, however, most loans made by intermediaries are "illiquid" – hard to sell quickly at the highest possible price. A buyer needs private information to assess the probability of default on the loans – that is why the borrowers had to take out loans in the first place rather than borrow at the safe interest rate. If an intermediary suffering a run is unable to find informed buyers quickly, it may have to accept low prices from uninformed buyers (in a financial "fire sale") and be left unable to pay off all its creditors even if it would have been solvent absent the run (Shleifer and Vishny 2011). The intermediary must then go out of business or at least "suspend payments." In a suspension of payments, an intermediary subject to a run refuses to pay off its short-term borrowings as originally pledged but promises to pay later in hope that the run will stop or it will be able to find informed buyers for its assets in the meantime. Either way, the intermediary stops lending at least for a while. A wave of runs on many intermediaries at once is a "financial crisis." Gertler et al. (2017) present a new Keynesian macroeconomic model with financial intermediaries subject to runs, in which a financial crisis depresses real activity.

In models runs are set off by conditions that would arise in a business cycle downturn (e.g., Allen and Gale 1998; Gertler et al. 2017). Thus runs may be another mechanism amplifying the effect of an increase in r (increasing β_{ry} and $\beta_{lag\ y}$). The relationship between economic conditions and runs is abruptly discontinuous because an individual lender to an intermediary has greater incentive to withdraw his loan if other lenders withdraw ("strategic complementarity") (e.g., Morris and Shin 2000; Goldstein and Pauzner 2005). In some models runs can be triggered by events not obviously related to macroeconomic conditions (e.g., Chari and Jagannathan 1988) or even occur *randomly* ("caused" by "sunspots") as in the well-known model of Diamond and Dybvig (1983). That implies runs can be a source of exogenous shocks to output (more components to ϵ_y in 8).

Generally, models of runs imply they can be prevented by an insurance system that pays off lenders to an intermediary if the intermediary defaults. Runs can also be prevented or mitigated by a "lender of last resort." A lender of last resort provides funds to an intermediary in exchange for its illiquid assets, valuing the illiquid assets at the relatively high prices they would fetch eventually from informed buyers. A lender of last resort can provide funds either by buying the illiquid assets outright or by taking illiquid assets as collateral for loans. Because a lender of last resort needs either stupendous wealth or the ability to create funds at will, the most obvious real-world institution that can act as a lender of last resort is a central bank.

Mechanics of Monetary Policy Implementation

Macroeconomic models skim over *minutiae*. New Keynesian macroeconomic models ignore, or describe in wildly unrealistic ways, the mechanism by which the value for a short-term interest rate desired by central bank policymakers becomes the market rate. This mechanism is, however, described in realistic models of monetary policy implementation. Most central banks today use variants of the same basic setup (described by Ennis and Keister 2008). Firms and households make payments to each other by transferring ownership of short-term loans – "deposits" – to certain financial intermediaries – "banks." Banks make payments to each other mainly by transferring balances in "reserve accounts" at a central bank. "High-powered money" is the total of these reserve balances and currency. The central bank pays a "reserve interest rate" on reserve-account balances and charges a higher "penalty rate" for short-term loans to banks. The market overnight rate cannot fall below the reserve interest rate (the return to lending overnight funds to the central bank) and cannot rise above the penalty rate (the cost of borrowing overnight funds from the central bank).

A bank that suffers a shortfall in its reserve account is required to borrow at the penalty rate to cover the shortfall. A shortfall might be an overdraft or a failure to meet a regulatory minimum "reserve requirement." A bank's "excess reserve" is its reserve balance less its reserve requirement. A bank's "free reserve" is its excess reserve *not* including funds borrowed from the central bank at the penalty rate. The total supply of free reserves is equal to high-powered money less currency less banks' required minimum balances less their penalty-rate borrowing. The central bank adds to (subtracts from) free reserve supply when it buys (sells) assets on the open market.

There is a cost to a bank of holding free reserves: it loses the spread between the reserve interest rate and the market overnight rate. But there is also a benefit. It is hard to predict exactly when some reserve-account payment orders will go through. A bank that aims to keep a free reserve of zero may actually suffer a reserve shortfall and have to borrow at the high penalty rate. As a bank trades off the cost of holding free reserves against the benefit, aggregate demand for free reserves is a downward-sloping function of the market overnight rate, between the upper bound of the penalty rate and the lower bound of the reserve interest rate. This reserve-demand curve is shifted up (down) by an increase (decrease) in the penalty rate and/or the reserve interest rate. The market overnight rate equates the demand for free reserves to the supply.

A central bank can control the market overnight rate because it controls the supply of free reserves, the penalty rate, and the reserve interest rate. A central bank can reduce the market overnight rate by adding to free reserve supply or by lowering the penalty rate and/or reserve interest rate. A sufficiently large free reserve supply can drive the market rate all the way down to the reserve interest rate. At this point banks will be indifferent to holding even more funds in their reserve accounts, so further increases in reserve supply will have no effect on market short-term rates.

Also, there will be no borrowing at the penalty rate to cover reserve-account shortfalls, because each bank holds a large enough free reserve to entirely eliminate the danger of a shortfall.

The ultimate lower bound on short-term nominal rates is the lowest possible value the central bank can pay for reserve balances. This might be the nominal interest rate on currency, which is zero, hence the "zero lower bound." When the market overnight rate has been pushed to the ultimate lower bound, the economy is in a "liquidity trap."

Historical Background: The American Economy in the 1920s

America's Banks

On the eve of the Great Depression, the American financial system was more vulnerable to crises than other countries with sophisticated financial markets such as Britain, Australia, and Canada. By the late 19th century, each of those countries had a few large banks with nationwide branch networks which lent to businesses and households all over the country. Defaults by borrowers in a particular region did not threaten such a bank's solvency. Bank assets included illiquid loans, long-term bonds, and two types of liquid short-term assets: "call money" loans and "acceptances." Acceptances were bills (transferrable short-term debts) which had been guaranteed – "accepted" – by a firm of well-known solvency, the "acceptor." Some acceptances financed purchases of securities. Others – "real bills" – financed purchases of materials, inventories, or goods for shipment. An accepted bill was usually liquid, easy to sell on secondary markets, because its default risk depended on the perceived solvency of its acceptor not the original issuer. Call money loans were overnight loans collateralized by stocks, bonds, or acceptances. They were taken out mainly by speculators and dealers in cities with big financial markets. Banks made call money loans through their branches in those cities.

In the world's largest financial market, London, the Bank of England had taken on the functions of a central bank and lender of last resort. The Bank could lend to practically anyone on almost any kind of collateral and did so in crises (Bagehot 1873, pp. 51, 196). However, the Bank could do a great deal to stabilize financial markets without lending directly to anyone in particular. In a crisis acceptances could become illiquid as doubts arose about acceptors' solvency. The Bank helped to maintain bills' liquidity, hence intermediaries' ability to pay off their creditors on short notice, by making a standing offer to purchase ("rediscount") acceptances at a standard rate (Bagehot 1873; Clapham 1970; Bignon et al. 2012).

The United States was different. Prior to 1914 it had no central bank. American regulations prevented a bank from operating offices in more than one state (or even, in some states, more than one city). Most banks lent mainly to businesses and households within a small region. A merely regional downturn could threaten the solvency of a bank. Fewer types of liquid assets were available to banks. Accepted bills did not exist, partly due to regulations (national banks were not allowed to

guarantee bills).[3] An American bank could make call money loans, but there was little demand for such loans outside New York City, the country's main securities market. Most banks outside New York made few if any call money loans. Their main liquid asset was interest-paying "interbank" demand deposits in New York banks which lent the funds on in the New York call money market (James 1978). The prevalence of interbank deposits added an extra source of potential instability to the American financial system. Just as depositors in the hinterland might run on their banks, hinterland banks might run on New York banks (Chari 1989). In the decades before 1914, America suffered several national financial crises in which depositors started to withdraw deposits *en masse* from hinterland banks, hinterland banks started to withdraw their deposits *en masse* from New York banks, New York banks suspended payment, and banks throughout America suspended payment because their main liquid asset had become illiquid (Wicker 2000).

These crises occurred in the wake of cyclical downturns and stock market crashes, which caused fears of defaults on bank assets (Calomiris and Gorton 1991). The cyclical downturns were in turn caused by hikes in American interest rates due to foreign financial crises or domestic monetary factors (Hanes and Rhode 2013). A national financial crisis accelerated a downturn, not just because it restricted the supply of loans from financial intermediaries as in new Keynesian models but also because widespread suspension of bank payments made it harder to hire labor and sell things. Factories closed because they could not get cash for weekly payrolls. Trade declined because it was harder to make payments between cities. James et al. (2013) describe these effects. They show that widespread suspensions of bank payment were coincident with sharp drops in real activity. Real activity stabilized or turned up when banks resumed payment.

At the beginning of the Great Depression, American banks were still prohibited from setting up branches in more than one state. Several state governments had attempted to set up systems of deposit insurance that could prevent bank runs by paying off depositors of a failed bank. But all these schemes had failed (Federal Deposit Insurance Corporation 1998).

The Federal Reserve System

The Federal Reserve System – the "Fed" – started operation in 1914. It was a confederation of 12 "Reserve banks" placed in major banking cities. One was in New York. Each Reserve bank had a geographically defined "district." The whole system was overseen by a central Board (later known as the Board of Governors) that

[3]The closest thing to accepted bills in American financial markets was a type of bill called "commercial paper." Commercial paper was less liquid than accepted bills, with no active secondary market, because it had no guarantor (James 1978). Its value depended on the perceived solvency of its original issuer, typically a firm that was relatively large but not as widely known as firms that accepted bills in Europe.

met in Washington. It was not clear exactly which potential actions of a Reserve bank required Board approval.

A bank could become a "member" of the Federal Reserve System. A member bank was required to hold a reserve balance in its district Reserve bank subject to a required minimum balance called a "reserve requirement." A member bank could use its reserve balance for payments to and from other banks and the federal government. Some banks (those chartered by the federal government, called "national" banks) *had* to become members. Other banks (chartered by state governments) could choose whether or not to join. Many chose not to because members were subject to costly extra regulations, and the Fed did not pay interest on reserve balances (Wingfield 1941).

The designers of the Fed hoped that it would eventually be able to operate as the Bank of England did, interacting with banks primarily through a market for acceptances. To that end American regulations were changed to encourage creation of acceptances, called "banker's acceptances" in America. Each Reserve bank set an "acceptance rate" at which it bought acceptances. By 1929 a market in acceptances had begun to develop, but banks outside New York still did not hold many of them (Ferderer 2003).

The Federal Reserve's ability to lend to financial intermediaries was strictly limited by the law setting up the Fed, the "Federal Reserve Act," and later amendments to that law. A Federal Reserve Bank could ordinarily lend only to a member bank. To lend to a nonmember bank, a Reserve bank needed special permission from the Board in Washington. A Reserve bank could buy, or take as collateral for a loan, *only* banker's acceptances, federal debt, or certain types of bank loans – "eligible paper." To be "eligible" a bank loan had to be short-term (original maturity 90 days or less) and "for producing, purchasing, carrying or marketing goods in one or more of the steps of the process of production, manufacture or distribution." A loan was *not* eligible if it had been used to finance purchases of stocks or bonds or "for permanent or fixed investments of any kind, such as land, buildings or machinery" (Steiner 1926). The rates at which Reserve banks bought ("rediscounted") or lent against eligible paper could vary across Reserve banks. They became known as Federal Reserve "discount rates." A bank that acquired funds this way was "borrowing through the discount window."

Discounting of eligible paper and purchases of banker's acceptances occurred at the initiative of member banks. At its own initiative, a Federal Reserve Bank could buy and sell federal government debt in "open-market operations." By the late 1920s, most Reserve banks' open-market operations were managed through a system-wide committee (later called the Federal Open Market Committee [FOMC]). Federal Reserve Banks also made a standing offer to buy and sell gold at a fixed price. This put the United States "on the gold standard" as described further below.

By the late 1920s, a market had developed in uncollateralized overnight loans of Federal reserve funds – "federal funds" (Meltzer 2003, p. 164), which operated alongside the long-established market for overnight call money. The level of market overnight rates for federal funds and call money was influenced by two factors under

the control of the Federal Reserve System: Reserve bank discount rates and the supply of free reserves. The mechanism (described by Hanes 2006) corresponded to the models described in section "Mechanics of Monetary Policy Implementation" except that the reserve interest rate was always zero. The "penalty rate" facing a member bank was its Reserve bank's discount rate (lacking acceptances, a bank short of funds usually had to borrow through the discount window). Free reserve supply was total reserve balances in excess of reserve requirements less funds supplied through the discount window. Fed purchases (sales) of gold, or of federal debt in open-market operations, added to (reduced) free reserve supply. A sufficiently large free reserve supply could push the market return to overnight lending down to the zero rate paid on reserve balances. At this point the economy would be in a liquidity trap.

Fed policymakers knew that their operations influenced short-term rates. But they had a different theory of the mechanism. Today their theory is called the "Burgess-Riefler doctrine" after two Fed staffers who developed it (Wheelock 1990, Meltzer 2003, pp.141, 161). This theory assumed banks wanted to hold a certain total quantity of reserve balances. If the Fed supplied less than this quantity through gold purchases and open-market operations, banks had to borrow the difference through the discount window. But banks were "reluctant to borrow." A bank that had borrowed through the discount window would curtail lending to pay its debt to the Fed. An increase in free reserve supply reduced discount borrowing and, hence, made banks more willing to lend and reduced market short-term interest rates. A decrease in discount rates had the same effect because it made banks more willing to remain in debt to the Fed.

The Burgess-Riefler doctrine implied that there was a limit to the Fed's influence on interest rates similar to the liquidity trap. Once free reserve supply was great enough to entirely eliminate discount-window borrowing, further increases in reserve supply could have no effect on financial market conditions (Wheelock 1991).

The Gold Standard and the Fed's Monetary Policy Strategy

In the late 1920s, the United States and most of its international trading partners were in an international gold standard system. Each country had a monetary authority, usually a central bank, that exchanged currency and central bank reserve balances for gold at a fixed price. This held foreign exchange rates close to the values implied by the countries' relative gold prices ("gold parities"). The authority of a country running a balance of payments deficit could cover it by selling reserves of foreign assets, but once those were gone, it had to sell gold. A country's long-run equilibrium price level was determined by its gold price and world prices, in gold, of tradable goods. The world gold price level was in turn determined by the balance of world gold supply against gold standard countries' demand for gold reserves. Overall the gold standard monetary regime appears to have kept inflation rates close to zero in the very long run (Taylor 1999). Thus it satisfied the precondition for new Keynesian models that the long-run expected inflation rate be close to zero.

For an individual country, the gold standard was in most respects simply a commitment to fixed exchange rates. This placed a constraint on monetary policy. If international capital mobility had been "perfect" (so that "uncovered interest-rate parity" held), the constraint would have been very tight for all countries all the time: each country's nominal interest rates would have to equal foreign rates. In fact international mobility was "imperfect" (as in textbook models, e.g., Romer 2006, pp. 236–241; Mankiw 2010, pp. 153–161). A country's interest rates need not equal foreign rates, but its net inflow of international investment, hence its balance of payments, decreased if foreign interest rates rose relative to the country's domestic interest rates. Thus the gold standard constraint on monetary policy was "asymmetric," tight on a country running a balance of payments deficit and loose on a country running a surplus (Temin 2000). Authorities of a surplus country did not have to eliminate the surplus. They could just keep buying up gold (or foreign assets); to prevent the high-powered money supply from increasing and reducing domestic interest rates, they could "sterilize" gold purchases by simultaneously selling bonds in open-market operations. Authorities of a deficit country, on the other hand, had to eliminate the deficit because they must run out of gold sooner or later. To eliminate the deficit, they had to raise domestic interest rates. In the short run, that did the trick by drawing in international investment. In the long run, it depressed real activity, lowered the domestic inflation rate, lowered the country's nominal wage and price level, and hence depreciated its real exchange rate. The gold standard's "rules of the game" called for a surplus country to allow its purchases of gold (or foreign assets) to boost its high-powered money supply and lower its interest rates. That would help other countries get back into balance. But there was nothing to *force* this policy on a surplus country.

Over most of the 1920s, the United States ran a balance of payments surplus. Federal Reserve banks usually held no foreign asset reserves, just gold. The most influential person in the Federal Reserve System was Benjamin Strong, head of the New York Reserve bank. Sometimes Strong called for decreases in US interest rates to help deficit countries remain on the gold standard. More often he wanted to sterilize gold purchases and use the Fed's influence on financial market conditions to stabilize the American price level. Strong and his staff monitored indicators of inflation and real activity. When there were signs of inflation or too-rapid growth, Strong pushed for increases in discount rates and/or open-market sales of securities to reduce free reserve supply. When there were signs of deflation or slow growth, Strong pushed for the opposite.[4] Strong usually managed to get support for his

[4]Strong explained the transmission mechanism from monetary policy to the price level this way: "when we have very cheap money, corporations and individuals borrow money in order to extend their business. That results in plant construction; plant construction employs more labor, brings in to use more materials for plant construction, and gives more employment. It may cause some elevation of wages. It creates more spending power; and with that start it will permeate through into the trades and the general price level" (quoted in Hetzel 1985, p. 7).

preferred policies (Meltzer 2003, pp. 169, 177, 209, 230). Thus, over the 1920s the Fed followed something close to a Taylor rule with a zero inflation target (Orphanides 2003).

But Strong's monetary policy strategy was not universally accepted among Federal Reserve policymakers. Many Reserve bank heads and most members of the Board in Washington were adherents of the "real bills doctrine." According to this doctrine, a central bank should not, and probably could not, use free reserve supply and discount rates to stabilize the price level.[5] Its discount rates should follow market rates, not control them. The important thing was to discourage "speculation" by hindering the supply of loans to finance purchases of securities. Thus, a European central bank should accept only real bills for rediscount. The Federal Reserve System should accept as collateral at the discount window only bank loans analogous to real bills. That was the idea behind the definition of eligible paper in the Federal Reserve Act.

Benjamin Strong died in 1928. His successor as head of the New York Reserve bank, George Harrison, shared Strong's views on strategy but lacked Strong's influence in the system (Eichengreen 2016).

The 1929–1933 Depression

1928: The Federal Reserve Hikes Short–Term Interest Rates

In 1928 there were no signs of inflation. Real activity was sluggish, if anything. But there was a developing boom in the stock market which Federal Reserve policymakers wanted to stop. They began to raise discount and acceptance rates and sold bonds to reduce free reserve supply. The idea was to raise short-term rates and choke back bank lending, in order to hinder the supply of loans to finance stock purchases (Meltzer 2003, pp. 224–225, 228, 235). Some Fed policymakers warned that this could have an undesirable side effect: it could cause a recession and unwanted deflation (pp. 225, 230, 239). Hoping to stop bank lending to finance stock purchases *specifically*, the Board directed Reserve banks to apply "direct

[5] Adolph Miller, the "dominant personality at the Board" (Meltzer p. 138), argued that arguments for price-level targeting relied on faulty assumptions: "One of those assumptions is that changes in the level of prices are caused by changes in the volume of credit and currency; the other is that changes in the volume of credit and currency are caused by Federal reserve policy. Neither one of those assumptions is true...undertaking to regulate the flow of Federal reserve credit by the price index is a great deal like trying to regulate the weather by the barometer. The barometer does not make the weather; it indicates what is in process" (quoted in Hetzel 1985, p. 10). Miller further argued that there was nothing the Fed could do to stop a recession: "you can not stop the recession by the lowering of the discount rate, the cheapening of the cost of credit, or making credit more abundant" (p. 12). The head of the Philadelphia Reserve bank said he did not believe the Fed should attempt to resist changes in the price level: "When the movement of prices is underway, it seems to me that it is always a doubtful and generally a dangerous thing for any outside agency to interfere with and attempt to alter the current" (p. 12).

pressure" against such lending. This meant "restricting the [discount] borrowings from the federal reserve banks by those member banks which were increasingly disposed to lend funds for speculative purposes" (Miller 1935, p. 454) and requiring "interviews between Federal Reserve Bank officials and member banks whose [discount-window] borrowing appeared to be excessive or too continuous" (Burgess 1930, p. 16). In effect, Federal Reserve Banks were to *ration* discount-window credit to banks that appeared to be lending on stock collateral.

The immediate result was a big increase in overnight rates. From December 1927 to May 1929, the call money rate rose by more than 4%. The overnight federal funds rate rose by about 2% (Willis 1957, p. 10). Both call money and federal funds rates rose well above the New York Reserve bank's discount rate, which had previously set a ceiling on market overnight rates, perhaps because "direct pressure" rationing raised the effective cost of discount borrowing above the posted discount rate (Beckhart and Smith 1932, p. 46; Wheelock 1990). A relatively clear indicator of the return to overnight lending (or at least the return to overnight lending expected to prevail within the next few weeks) is the rate on very short-term Treasury debt. This is plotted in Fig. 2 on a monthly frequency, along with the yield on long-term Treasury bonds.

The Initial Downturn 1929–1930

On October 24, 1929 ("Black Thursday"), the stock market crashed. Real activity had turned down before that. The NBER cyclical peak is August 1929. The obvious

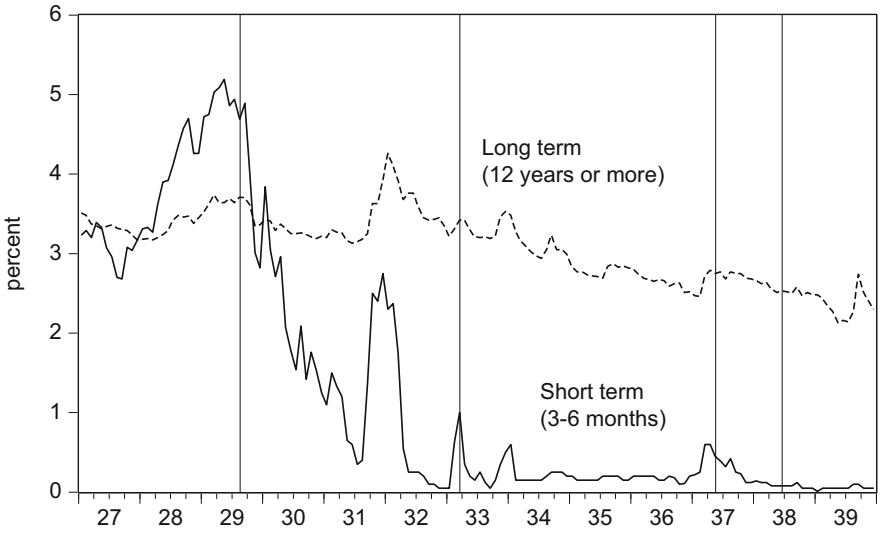


Fig. 2 Treasury interest rates 1927–1939. (Sources and notes: short-term 1927–1928 from US Federal Reserve Board (1943), Tables 122, 3- to 6-month Treasury notes and certificates. Short-term 1929–1939 from Cecchetti (1988), Table A1, 3 Months. Long term from US Federal Reserve Board (1943), Table 128, US government)

explanation of the downturn is the preceding hike in interest rates. This was plausibly an exogenous shock to real activity because it was not a response to inflation or rapid growth. From December 1927 to May 1929, short-term Treasury rates rose about 2% (Fig. 2). This compares with increases in short-term Treasury rates in postwar (after the Second World War) policy tightenings engineered by the Fed. Romer and Romer (1989) identify six instances in which the Fed hiked short-term interest rates for reasons unrelated to current or forecast inflation. Around some of these tightenings, across a 6- or 12-month window around the date of the policy change, 3-month Treasury rates rose more than 2%. Across many of them, 3-month Treasury rates rose *less* than 2%. *All* of the tightenings were followed by cyclical downturns.

After August 1929, however, the decline in real activity was extraordinarily sharp. On current estimates by the US Bureau of Economic Analysis (BEA), from 1929 to 1930, real GDP fell by about 8 1/2%. After the end of the Second World War, real GDP never fell that much in 1 year, even in the most severe recessions. The decline in real GDP at the beginning of the "Great Recession," from 2008 to 2009, was less than 3%. What can explain the violence of the 1929 downturn? Was there something about the economy in 1929 that magnified the effect of an increase in interest rates? Were there other exogenous shocks that contributed to the downturn?

New Keynesian models imply that the immediate decline in real activity associated with a hike in interest rates is unusually large if people expect the current recession to be unusually bad, so that expected future real activity deviates from its usual relationship to current real activity (negative values of z^e in (5)). That cannot explain 1929–1930. There is no evidence expectations were unusually pessimistic at the beginning of the Great Depression. It appears people believed they were facing a normal cyclical downturn (Mathy and Stekler 2017). Business managers forecast a slowdown in real activity less severe than the one that actually occurred (Klug et al. 2005). There was nothing unusual about that. At the start of deep recessions in the postwar era, forecasts of real activity have been similarly overoptimistic (Mankiw 2010, p. 449).

The stock market crash may have worsened the downturn. It was a big hit to households' net worth (Mishkin 1978). In New Keynesian models with financial market frictions, a reduction in wealth reduces real activity. To the degree the crash reflected a decline in rational expectations of future corporate earnings, this was part of the financial accelerator. If the crash was the bursting of an irrational bubble – a debatable point (White 1990) – it was an additional cause.

Had the American economy built up some type of "financial fragility" that supercharged the financial accelerator? In models of financial frictions, the decline in real activity resulting from a negative shock is greater if households have previously taken on lots of debt in a "credit boom," so that their net worth is already low at the time the shock hits (Gertler and Gilchrist 2017). Perhaps such a credit boom had taken place in the 1920s (Eichengreen and Michener 2004). New forms of debt had been created to finance purchases of new "consumer durables" such as automobiles and radios. Much of this debt was on the "installment plan," which gave borrowers strong incentive to cut back on other spending in response to a decline in current income. This may have caused household spending to decline more after 1929 (Olney 1999). Also, around 1925 there had been a nationwide boom in residential real estate: an increase in house construction, rising house prices, and

an unprecedented increase in mortgage debt. The boom ended before 1929: over 1925–1927 prices fell off a bit, and construction fell off a lot. But the boom may have left something behind, such as more mortgage debt that affected spending after 1929. Over 1929–1933 the cities that had seen the biggest booms around 1925 suffered the biggest declines in house prices and the highest mortgage foreclosure rates (Brocker and Hanes 2014).

Continued Decline 1930–1933

Real activity continued to decline through 1933 – 4 years in a row. This is what made the Great Depression great. In most postwar recessions, real GDP fell for just 1 year. There were just two postwar recessions in which real GDP fell for as long as 2 years (1973–75, 2008–2010).

One factor that probably did *not* contribute to continued decline was fiscal policy. The response of the federal government led by the Hoover administration was to boost spending while raising tax rates as needed to maintain a balanced budget (Fishback 2010, pp. 402–403). Figure 3 plots real purchases of goods and services by all levels of government (federal, state, and local). Government purchases rose from 1929 to 1931 and fell only modestly from 1931 to 1933. Figure 4 plots taxes, government transfer payments to households, and the ratio of taxes net of transfers to national income. New Keynesian models with financial frictions imply that an increase in this ratio might reduce spending of liquidity-constrained agents. The ratio did increase after 1931 but not much. Marginal personal income tax rates were lower over 1930–1933 than they had been in 1929 (Meltzer 2003, p. 563).

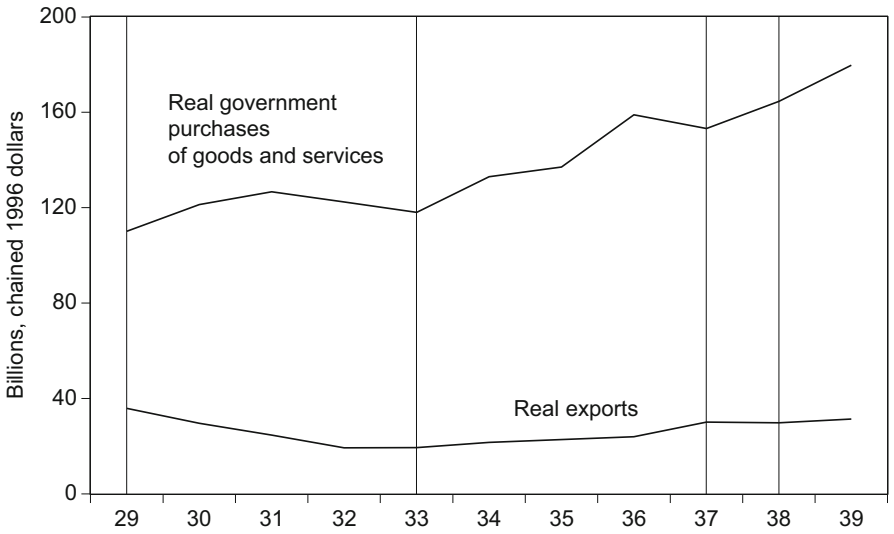


Fig. 3 Real government purchases of goods and services, real exports 1929–1939 (Sources and notes: both series from Carter et al. (2006), government purchases series Ca89, exports series Ca87)

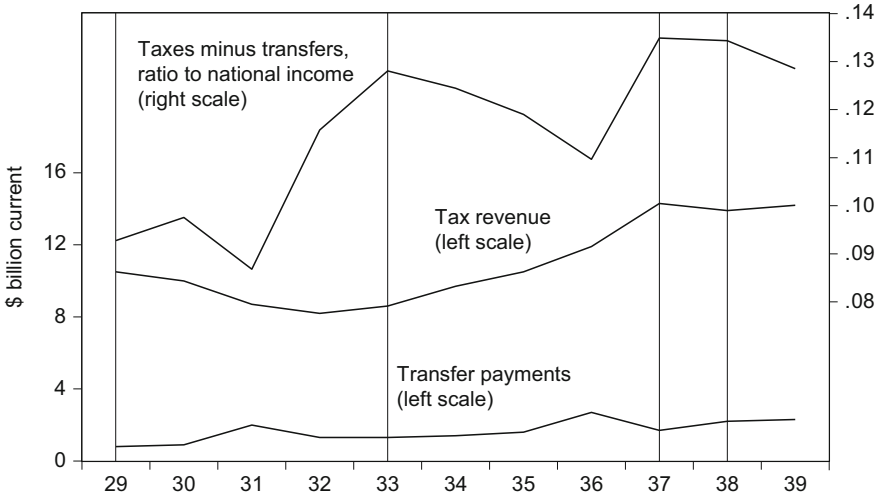


Fig. 4 Taxes, government transfer payments, and ratio of taxes minus transfers to national income 1929–1939. (Sources and notes: United States Bureau of Economic Analysis (https://www.bea.gov/iTable/index_regional.cfm) Table 3.1 for taxes and transfers, Table 1.17.5 for national income. Taxes are current receipts including social insurance contributions. Transfers are government transfer payments to persons. National income is gross national income)

A factor that did contribute to decline but only in a small way was a drop off in demand for American exports. Real exports, plotted in Fig. 3, fell steadily from 1929 through 1932. This was largely blowback from the American interest-rate hike of 1928–1929. Along with the stock market boom, the hike in American interest rates drew in international investment, increased the American balance of payments surplus, and drained gold from foreign monetary authorities. To save their gold reserves, they raised domestic interest rates. That, and a series of financial crises in Europe, put foreign economies into recession and decreased their demand for American goods (Eichengreen 1992 p. 246). But exports were a small portion of American GDP. Thus this (and the "Smoot-Hawley" tariff increase of 1930) cannot have made much difference (Temin 2000, p. 305).

So what explains the persistence and depth of the decline in real activity after 1930? One obvious cause was a perverse hike in short-term interest rates engineered by the Fed in the midst of the downturn. Another was an increase in bank failures that started in late 1930 and ended in 1933 in a massive financial crisis. New Keynesian models imply both of these factors could prolong and deepen the Depression.

Federal Reserve Interest Rate Policy

On the heels of the October 1929 stock market crash policymakers at the Federal Reserve Bank of New York bought bonds in open-market operations, hoping to help stabilize financial markets. They did not ask the Board in Washington for

permission, which annoyed members of the Board. At the beginning of November, the Board allowed the New York Reserve bank to cut its discount rate but only on condition that it suspend open-market purchases (Meltzer 2003, pp. 243–44).

By January 1930 it became apparent that a recession had begun (Meltzer 2003, 291). As the recession deepened, Fed policymakers became aware of its severity (p. 315). Fed policymakers did not agree on what their response should be. Some, including George Harrison of the New York Reserve bank, wanted to lower market interest rates and spur bank lending to fight the recession (consistent with Benjamin Strong's strategy). Others, adherents of the real bills doctrine, opposed any such actions (Meltzer 2003, p. 290). Harrison's side mostly prevailed. From 1930 through the middle of 1931, they pushed through several cuts in discount and acceptance rates. Free reserve supply grew rapidly as the system purchased bonds in open-market operations and refrained from sterilizing gold inflows. Overall, Fed interest-rate policy in the first year of the Depression was similar to policy in postwar downturns including the post-2008 Great Recession. In the Great Recession, the Fed began to cut its target federal funds rate in July 2008; over the following year, the market federal funds rate and 3-month Treasuries both fell about 3 1/4%. From October 1929 to October 1930, federal funds rates fell by about 4 1/4% (Willis 1957, p. 10); short-term Treasuries fell about 2 3/4% (Fig. 2).

In late 1931 all the Fed's information indicated that real activity was still falling. But the Fed took steps to raise short-term rates. As shown in Fig. 2, short-term Treasury rates rose almost 2 1/2% from August to December 1931. The increase in short-term rates was accompanied by an unusually large increase in long-term rates.

Why did the Fed act in this perverse way? To maintain the gold standard. Over 1930–1931 many foreign central banks had raised domestic interest rates to prevent gold outflows. In summer 1931 German authorities stopped outflows of international investment and gold with regulations (exchange controls). In September 1931 authorities in Britain and most dominions of the British Empire stopped paying out gold for domestic currency, effectively floating their currencies (Temin 2000 p. 312; Meltzer 2003, p. 342). International investors feared the United States would follow suit. They sold American assets and bought gold. The Federal Reserve's gold reserve was draining out. To show commitment to the gold standard and stop the drain, the Federal Reserve System acted to raise American short-term interest rates, mainly by hiking discount rates (Eichengreen 1992, pp. 293–294).

The interest-rate hikes did what they were supposed to do. Gold outflow stopped in November and reversed in December (Meltzer 2003, p. 354). Relieved, Fed policymakers began to discuss steps to lower interest rates. But they were stymied by a provision of the Federal Reserve Act which required each Reserve bank to hold a minimum gold reserve (valued at the official price) equal to 35% of the balances held in that Reserve bank plus 40% of the Federal Reserve notes (a type of currency) issued by that Reserve bank. Some Reserve banks were up against these minimums. Without "free gold" they could not cover the increases in reserve balances that would result from open-market bond purchases or increased discount-window lending to banks (Meltzer 2003, pp. 354, 357).

In February 1932 Congress passed a revision to the Federal Reserve Act that loosened the free gold constraint. The Federal Reserve began to buy bonds in open-market operations. The New York Reserve bank cut its discount rate and cut it again in June. By late spring 1932, Federal funds rates were at the lowest level ever observed in the 1920s–1930s (1/8 of 1%), and there was almost no discount borrowing (Willis 1957, p. 10). Three-month Treasuries were down to 1/4% (Fig. 2).

In July 1932 the Fed stopped buying bonds. Why? Eichengreen (1992, pp. 315–316) argues that Fed policymakers feared further bond purchases might raise doubts about American authorities' commitment to the gold standard and trigger outflows of international investment and gold like those seen in 1931. Epstein and Ferguson (1984) argue that Fed policymakers feared that low short-term interest rates would depress bank profits. According to Hsieh and Romer (2006, p. 169), "the Federal Reserve decided to slow the monetary expansion in mid-June in part because its model of monetary policy led it to believe that monetary conditions were already loose and that further purchases would be of little use." They refer to the Burgess-Riefler doctrine which implied that open-market purchases and discount-rate cuts lost their stimulative power once free reserve supply was so great that banks had no debt to the discount window. Hsieh and Romer call the Burgess-Riefler doctrine "a flawed model of the economy" (p. 172).

The Burgess-Riefler doctrine was indeed a flawed model, but new Keynesian models imply that Federal Reserve policymakers were right about one thing: further increases in reserve supply or discount-rate cuts were useless. Short-term rates were at the zero bound. As shown in Fig. 2, by the end of 1932, 3-month Treasuries were less than a tenth of a percent. This is about the same as the 3-month Treasury rate that prevailed in 2010, when the Federal Reserve's overnight rate target was practically zero. The absence of discount borrowing in 1932 confirms that the economy was indeed in a liquidity trap. Recall that in the correct model of overnight rate determination described in "Mechanics of Monetary Policy Implementation," when the return to overnight lending is equal to the reserve interest rate (zero in this case), banks stop borrowing at the penalty rate to cover reserve-account shortfalls.

Unfortunately, the Fed was unable to leave interest rates on the floor for long. The gold standard took one more bite. In February 1933 international investors again began to fear that American authorities would devalue the dollar. The balance of international investment again turned against the United States. The New York Reserve bank, which sold most gold for export, was running out of gold. In accordance with gold standard orthodoxy, Fed policymakers allowed market interest rates to rise as gold sales reduced free reserve supply by refraining from sterilization and hiking discount rates (Meltzer 2003, p. 379). But this time the increase in short-term rates did not stop the gold drain (Friedman and Schwartz 1963, p. 326; Wigmore 1987, p. 748; Meltzer 2003, p. 387). On March 6, 1933, the incoming president, Franklin Roosevelt, solved the problem by suspending Federal Reserve Banks' obligation to exchange gold for dollars (Friedman and Schwartz 1963, p. 328).

Bank Failures and the Financial Crisis of 1933

From late 1930 through 1932 there were waves of bank failures. Their timing is indicated by Fig. 5, which plots total deposits in banks that suspended payment and reopened later or went permanently out of business in a given month, as a fraction of deposits in American banks. The first wave hit in October 1930. Another peaked in October 1931, while the Fed was raising short-term rates in the wake of Britain's devaluation. Until late 1932 failures remained confined to a few cities or regions, and the fraction of deposits in failed banks was very small. These failures may nonetheless have significantly disrupted financial intermediation. Total deposits, also plotted in Fig. 5, began to fall at the time of the first wave of failures and plummeted at the time of the second wave. At the same points in time currency held by the public increased (Friedman and Schwartz 1963, pp. 311, 313). It looks like regional bank failures caused nationwide withdrawal of funding from banks.

At the end of 1932, things got worse. A wave of failures that began in the far west spread to the Midwest and the then-important city of Detroit. In February 1933 New York banks began to see mass withdrawals of interbank deposits (Friedman and Schwartz 1963, pp. 324–326). In March 1933 there was a crisis worse than any of the pre-1914 era. Pre-1914 crises had been resolved by suspensions of payments organized by private bank associations (clearing houses) during which banks stopped handing out currency but still provided many other services. The 1933 crisis

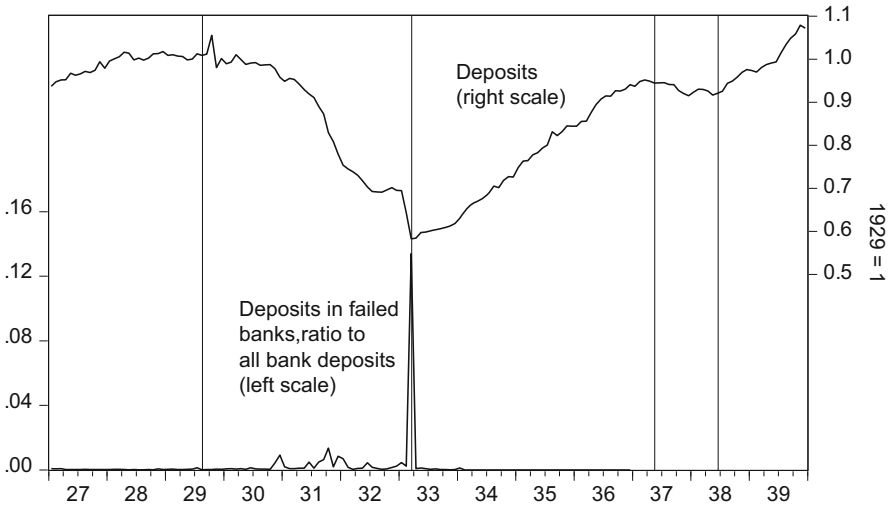


Fig. 5 Deposits in failed banks 1927–1936, deposits in all banks 1927–1939. (Sources and notes: deposits from Friedman and Schwartz (1963) Table A-1, total adjusted deposits in commercial banks. Deposits in failed banks from Federal Reserve Bulletin December 1937, p. 909. The failed bank series underestimates the intensity of failure in some months as it does not include banks that suspended or partially defaulted on deposits by "agreement" with depositors. Such agreements were common in 1931 and 1932 (Federal Reserve Bulletin September 1937, p. 1206). More importantly, it fails to include banks temporarily shut down by bank holidays)

was resolved by government-ordered shutdowns of all banks, cheerfully called "bank holidays," that stopped nearly all bank services. By the beginning of March 1933, nearly all states' governors had declared holidays. On March 6, incoming president Franklin Roosevelt declared a national bank holiday that shut down all banks in the country. At that time households and businesses in much of the country had already been without bank services for weeks.

Bank failures may have been an amplification mechanism more than an independent cause of the Depression. Up to 1933, at least, banks that failed appear to have been relatively close to insolvency *before* depositors began withdrawing *en masse* because of interactions between their prior investment policies and the Depression's effects on loan defaults and prices of relatively risky bonds (Calomiris and Mason 1997, 2003b). In any case bank failures and withdrawal of funding from banks must have deepened the Depression. In models of financial frictions, hindrances to the operation of financial intermediaries depress real activity by raising the cost of funds to liquidity-constrained borrowers. Bernanke (1983) argues that this mechanism was at work over 1930–1933. Christiano et al. (2003) present a new Keynesian model with financial intermediation by banks, specifically meant to describe the Depression era. In the model, withdrawal of deposits and increased demand for currency reduces aggregate output.[6] Recall that in pre-1914 financial crises, suspensions of payments appear to have depressed real activity by making it harder to produce and trade. There is abundant evidence that held over 1930–1933 (Rockoff 1993; Kennedy 1973, pp. 161–164). Over 1929–1933, across time and across geographic regions, bank failures were strongly associated with lower real activity (Bernanke 1983, Calomiris and Mason 2003a).

Where Was the Lender of Last Resort?

Some banks that failed over 1930–1932 were merely "illiquid" rather than "insolvent" (Richardson 2007). They closed only because their depositors had withdrawn *en masse* creating a financial fire-sale problem. Certainly, it was a liquidity crisis, not a wave of ordinary insolvency, that hit in 1933. In a liquidity crisis, a lender of last resort can prevent closure of financial intermediaries by purchasing their illiquid assets or taking them as collateral for loans. A central bank is supposed to act as a lender of last resort. What was the Federal Reserve System doing all this time?

Provisions of the Federal Reserve Act hindered the ability of the Federal Reserve System to act as a lender of last resort. In the regional crises of 1930–1932, many banks that needed to borrow were not member banks. Some member banks that

[6]In their model the shift from deposits to currency occurs because of a shock to preferences, not a response to bank failures. They also assume that the supply of high-powered money would not respond to the consequences of the shock, which is not realistic. But the model does illustrate how deposit withdrawal can affect real activity within an otherwise-conventional new Keynesian model.

needed to borrow did not hold enough assets the Fed would accept as collateral for a loan – Federal debt, banker's acceptances, and eligible paper (Carlson and Wheelock 2016). The Fed was not allowed to lend against many types of bank assets that were illiquid or became so in a crisis, such as long-term corporate bonds.

To act effectively as a lender of last resort, a Reserve bank had to "improvise and test the limits of the Federal Reserve Act" (Bordo and Wheelock 2013, p. 88). Some Reserve banks, most prominently the Federal Reserve Bank of Atlanta, did push the limits. Other Reserve banks, such as the Federal Reserve Bank of St. Louis, did not. Richardson and Troost (2009) examined a crisis in late 1930 that affected banks in a region that straddled the line between the Atlanta and St. Louis districts. They found banks were less likely to fail, and business activity remained stronger, within the Atlanta district. "If other Federal Reserve Banks had pursued similar strategies [to Atlanta's], fewer banks would have failed, and the depression may have followed a different course" (p. 1034). Unfortunately, most Reserve banks were like St. Louis.

In the absence of Federal Reserve action, President Hoover and Congress set up other institutions that could lend to more types of financial intermediaries and on more types of collateral. The most important was the Reconstruction Finance Corporation (RFC) established in January 1932. Unfortunately, the RFC's effectiveness was crippled by a law passed in July 1932 that "was interpreted as requiring the publication of the names of banks to which the RFC had made loans...The inclusion of a bank's name on the list was correctly interpreted as a sign of weakness, and hence frequently led to runs on the bank" (Friedman and Schwartz 1963, p. 325).

Why did Federal Reserve policymakers not try harder to act as lender of last resort? Because most of them did not think that was their job. Even the head of the New York Federal Reserve bank, George Harrison, opposed proposals to let Reserve banks lend on more types of assets or lend indirectly through less-constrained agencies (Meltzer 2003, p. 347). In February 1932 the same legislation that loosened the Fed's free gold constraint (the Glass-Steagall Act of February 1932) made it possible for Reserve banks to lend on any kind of collateral, under certain conditions (only to relatively small member banks and on approval of Board members [Carlson and Wheelock 2016]). Incredibly, Reserve banks failed to take advantage of this (Meltzer 2003, p. 358). The head of the Fed's Board from 1930 through early 1933, Eugene Meyer, supported the creation of the RFC and even served concurrently as chairman of the RFC's board. He did not appear to believe there was any overlap between the functions of the RFC and the Fed.

When the Fed was founded in 1914, its supporters argued that it would prevent the kinds of financial crises that had plagued the United States. But they did not expect it to do so by acting as a lender of last resort. They believed the United States had suffered financial crises before 1914 mainly because its supply of high-powered money did not adjust to seasonal fluctuations in money demand; by following the real bills doctrine, the Federal Reserve System would automatically adjust money supply and create an "elastic currency" so that "such a thing as a currency and credit panic can not exist under the Federal Reserve System" (Miller, quoted in Hetzel 1985, p. 12).

Some influential people in the Federal Reserve System believed that a lender of last resort actually did more harm than good. Henry Parker Willis was the Secretary of the Federal Reserve board 1914–1918 and research director at the board 1918–1922. In a book published in the wake of the 1933 crisis, he wrote (Willis 1936):

> The commonest and least precise idea of central banking is that which regards it as a form of "emergency relief" or "panic insurance." According to this viewpoint, ordinary banks are likely from time to time to get into "trouble".. . due to the acquisition of "frozen" [that is illiquid] assets...lack of confidence in which depositors or other creditors, believing they cannot at will convert their claims into cash, may bring about a "run," thereby forcing a bank into an embarrassed condition, or even forcing it to suspend...it may be possible to afford "relief"...through rediscounting or buying paper held in the embarrassed institution which it could not otherwise dispose of (pp. 5, 6)

> the view of central banking which considers it a means of helping out hard-pressed banks that have become "frozen" is not only theoretically unsound but is actually found in most cases to be injurious...When a central bank does so it merely tends to make a bad matter worse. (p. 15)

Willis believed that the constraints on Fed lending imposed by the definition of eligible collateral were a good thing *because they made it harder for the Fed to act as lender of last resort*:

> The fact that throughout the whole history of Federal Reserve banking, there has been continuous resistance to the observance of eligibility rules shows conclusively the need of them, and illustrates the danger involved in making the central bank merely a medium of emergency relief, to be availed of in times of stress...As long as this conception of central banking prevails, there will, of course, be continuous danger. (p. 138)

Meltzer (2003, p. 731) concluded that Fed policymakers were influenced by a "firm belief" that the system "could, or should, do nothing to prevent bank failures... Failures, they believed, were the inevitable consequence of bad decisions and speculative excesses that had to be purged before stability could return."

Aggregate Supply in the 1929–1933 Downturn: Wage Inflation, Price Inflation, and Real Wages

While real activity declined over 1929–1933, what happened to inflation? Pretty much what one would expect based on new Keynesian models and experience in other cyclical downturns.

In simple new Keynesian models, inflation rises and falls with the output gap (expressions (9), (10)). One of the conditions generating that relationship is that the public's expected value for the inflation rate that will prevail in the long-run future is zero. That condition seems to have held at the outset of the Great Depression. Inflation had been close to zero for several years. Judging from contemporary

literature, most people believed the gold standard would assure a roughly stable price level in the long-run future. Experts assumed the dollar's gold value would remain fixed and forecast a stable or slightly decreasing price level based on the balance of world gold supply and demand (Nelson 1991, pp. 6–7). Thus, new Keynesian theory implies that wage and price inflation should have fallen along with real activity over 1929–1933. That is what happened. Figure 6 plots inflation (percent change from the same month or quarter of the previous year) in a monthly index of average hourly earnings in manufacturing (the only high-frequency indicator of wage inflation from the 1930s that can be compared with series from other historical eras) and two price indices: a monthly CPI and a quarterly GNP deflator. Inflation started to fall almost immediately after the cyclical peak and fell more as the depression deepened. Because inflation was about zero at the start, this meant deflation.

There have been other historical eras in which expected long-run inflation was probably close to zero. Before 1914 the international gold standard could reasonably be believed to ensure a stable price level in the long run. From the 1950s through the mid-1960s and again after the 1980s, expected future long-run inflation appears to have been "anchored" at a low value by public confidence in monetary policy. Data from both of those eras are consistent with (9) and (10): wage and price inflation are positively (negatively) correlated with estimates of the output (unemployment) gap (Gordon 1990; Alogoskoufis and Smith 1991; Allen 1992; Gali 2011; Ball and Mazumder 2015; Blanchard 2016).[7]

The magnitude of deflation over 1929–1933 was about what one would expect from patterns in those other eras. Figure 7 is a scatterplot of annual data on wage inflation and unemployment rates from 1924 to 1939 and 1891–1914 and 1954–1965. I chose these particular sets of years to compare with the Great Depression because they are covered by comparable series on wage inflation and unemployment rates, and they were unaffected by wage and price controls.[8] The observations for 1930–1932 are in line with those from 1891 to 1914 and 1954–1965.[9] Observations from 1933 to 1938 are different: those years' annual wage inflation is anomalously high. I will return to this point later.

Returning to Fig. 6, note that over 1930–1931, the two measures of price inflation fell faster than wage inflation: real wages were countercyclical. That may seem odd. Most studies of postwar data find real wages to be procyclical or acyclical. But it was the usual pattern in recessions prior to the Second World War. It was not anything special about the Great Depression. Whenever one can compare similarly

[7]In some postwar years, from the later 1960s through the 1980s, real activity was correlated with the *change* in inflation – that is, inflation was strongly "persistent." Inflation persistence can be generated in new Keynesian models by "indexing" wages and prices to lagged inflation (e.g., Christiano et al. 2005). It can also be generated in models where the expected long-run future inflation rate varies over time. The latter possibility is tricky; see Ascari (2004), Kozicki and Tinsley (2002), and Cogley and Sbordone (2008).

[8]The Korean War controls were lifted in February 1953 (Rockoff 1984).

[9]It is sometimes claimed that nominal wages were unusually rigid in the 1929–1933 downturn (e.g., O'Brien 1989; Ohanian 2009). Obviously not true.

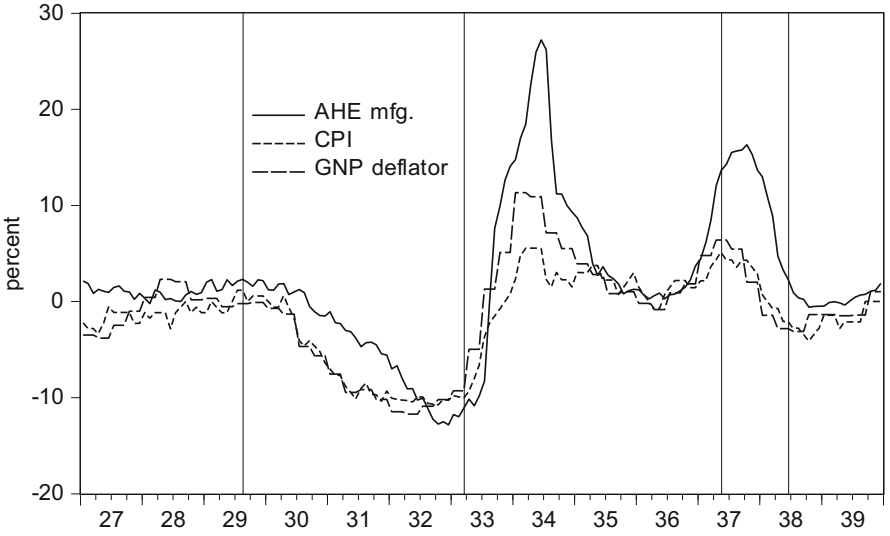
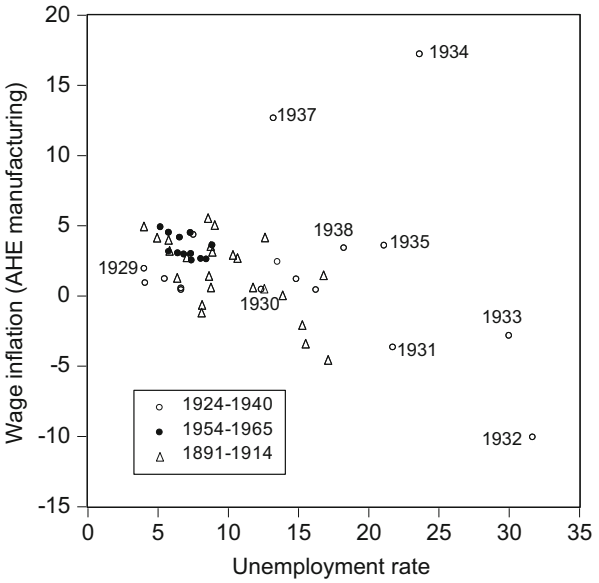


Fig. 6 Inflation in average hourly earnings, consumer prices, GNP deflator 1927–1939. (Sources and notes: average hourly earnings from Hanes (1996). Consumer price index from Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/series/CPIAUCNS). GNP deflator from Balke and Gordon (1986) Table 2. Both the CPI and GNP deflators are rougher estimates than their postwar counterparts constructed from lower-frequency estimates by interpolation on very limited data)

Fig. 7 Wage inflation and unemployment 1924–1940, 1954–1965, and 1891–1914. (Sources and Notes: annual average hourly earnings in manufacturing 1891–1914 from Rees (1961) "nine-industry" index; 1924–1940 and 1954–1965 from Hanes (1996). Private nonfarm unemployment rate from Weir 1992, Table D3)



constructed wage and price series across historical eras, real wages appear less procyclical, more countercyclical in eras prior to the Second World War (Hanes 1996). Huang et al. (2004) show that this historical development is consistent with new Keynesian models that allow for multiple stages of production, so that sticky-priced output is used partly as an input to production of more sticky-priced output. In these models real wages are more countercyclical if there is fewer number of stages or rounds of production before final sale. In the United States, both consumer goods and aggregate final output (GDP) were less finished, in this sense, in the 1930s than in postwar decades (Hanes 1999).

The Recovery 1933–1937

The recovery from the March 1933 cyclical trough was spectacular. On current BEA estimates, in *every* year from 1934 through 1936, annual real GDP grew faster than in *any* year after the Second World War. Three factors undoubtedly contributed to recovery. First, fiscal policy was mildly stimulative. Second, short-term nominal interest rates were as low as they could go: monetary policy pushed them back down to the floor and, this time, kept them there. Third, financial intermediaries were reactivated. Banks reopened and were reformed in a way that created confidence they would not fail in the future. Deposits flowed back in. That is, funding for loans to credit-constrained borrowers flowed back into financial intermediaries. Yet another factor may have contributed to recovery, but its existence is hard to prove. Monetary and exchange-rate policies adopted by the incoming Roosevelt administration may have created expectations of future inflation. In new Keynesian models, that would give an immediate boost to current real activity.

Fiscal Policy

Real government purchases of goods and services rose from 1933 through 1936 (Fig. 3), while the ratio of taxes net of transfers to national income fell (Fig. 4). Thus fiscal policy contributed to recovery. But most analysis has concluded that the contribution was modest at best (e.g., Brown 1956). In 1936 the federal government made a one-time "bonus" transfer payment (apparent in Fig. 4) to veterans of the First World War. Hausman (2016) argues that the bonus allowed credit-constrained recipients to borrow against expected future income growth and boosted 1936 GDP growth by as much as 1.6%. But that was not much relative to total real GDP growth in 1936, which was almost 13%.

Revival of the Banking System

Most banks were back in business less than 2 weeks after Roosevelt's proclamation of the national bank holiday on March 4, 1933 (Friedman and Schwartz 1963,

pp. 421–428; Awalt 1969; Kennedy 1973, pp. 179–202; Meltzer 2003, pp. 421–435). Within that short time, legislation had been passed that allowed the RFC to lend on security of low-quality assets and to make equity investments in banks to boost bank capital. Also, all banks had been examined by agents of existing bank regulators and triaged. Those that were clearly solvent and liquid were reopened over March 13–15. Deeply insolvent banks were sent into liquidation; owners lost their investment, depositors and other creditors received a fraction of what was owed them. (The process usually involved a loan from the RFC, so that illiquid assets could be sold off slowly and get a better price.). Marginally solvent banks were reorganized and reopened gradually (often with a capital contribution from the RFC). In January 1934 a new deposit insurance scheme started operation, backed by the federal government (Federal Deposit Insurance Corporation 1998, pp. 27–31).

Deposits began to flow back into banks quickly starting in January 1934 (Fig. 5). Of course, that does not mean the spread between safe interest rates and the cost of funds to credit-constrained borrowers snapped right back to pre-1929 levels. There is evidence loan supply remained impaired after 1933. Many banks still had less capital than before the depression, and they appear to have required higher expected returns on risky assets such as loans relative to safe assets such as government bonds (Friedman and Schwartz 1963, pp. 449–462; Calomiris and Wilson 2004).

The Path of Short-Term Interest Rates

In January 1934 the Roosevelt administration put the United States back on the gold standard, in the sense that American monetary authorities resumed exchange of dollars for gold at a fixed price. It was a new, higher price, however, which devalued the dollar against currencies of the many countries which continued to fix, or at least tightly manage, their gold exchange rates. Under the new institutional arrangements, the Treasury rather than the Fed exchanged gold for dollars (buying gold from anyone, selling gold only to foreign monetary authorities). But just as before, gold purchases added to high-powered money (Meltzer 2003, p. 458). Also as before the United States usually ran a balance of payments surplus. The Federal Reserve did not sterilize the effect of Treasury gold purchases, so high-powered money and free reserves grew rapidly. Overnight rates fell. In early 1934 the federal funds rate was back on the floor. Discount borrowing to cover reserve shortfalls was nil (Hanes 2006). This situation prevailed through the beginning of 1937, as indicated by the low Treasury bill rate in Fig. 2.

Expected Future Inflation

Many economists have argued that Roosevelt's suspension of gold payments in March 1933 and devaluation of the dollar against gold in January 1934 created expectations of future inflation (Temin and Wigmore 1990; Jalil and Rua 2016). That

would certainly have been true if expectations were sufficiently rational. Generally, exchange-rate devaluation raises import prices and the domestic price level (Eichengreen 2004). A commitment to a devalued exchange rate is effectively a promise on the part of the authorities that a higher domestic price level is coming sooner or later (Svensson 2003). Over 1933–1934, Roosevelt not only depreciated the dollar but pledged to "reflate" prices back to pre-Depression levels. Romer (1992) argues that realized increases in high-powered money supply from 1934 on also raised expected future inflation.

In standard new Keynesian models, when short-term nominal rates are stuck at the zero bound, news that inflation is coming in the future boosts real activity immediately. Recall that current real activity depends not only on the current short-term real interest rate but on the path of expected short-term real rates from now into the distant future (expression 3 above). Given the path of expected future nominal rates, a belief that inflation must come *sometime* lowers expected real rates for at least some future periods. Krugman (1998, p. 161) and Svensson (2004, p. 90) argue that Roosevelt's policy moves of 1933–1934 may have spurred real activity in this way. Eggertsson (2008) presents a new Keynesian model of the 1930s economy in which a mechanism of this type accounts for most growth in output over 1933–1937.

Unfortunately, there is no simple way to confirm that Roosevelt's policies and pronouncements really did affect beliefs of households and business managers in this way. There were no surveys of inflation expectations in the 1930s like today's Survey of Professional Forecasters or Michigan Survey.

Aggregate Supply over 1933–1937: Anomalous Inflation

Figure 6 shows that wage and price inflation picked up as real activity recovered after March 1933, as one would expect. But the rates of inflation, especially wage inflation, that prevailed after March 1933 were anomalously high. Output, though rising, was still far below the pre-Depression trend. Unemployment, though falling, was still high. In Fig. 7, annual wage inflation looks wildly high in 1934 and 1937.

One possible explanation of this anomalous inflation is the same new Keynesian expected-inflation mechanism that could have contributed to recovery of real activity. In new Keynesian models, extraordinary events that raise expected future inflation at any horizon, however distant, give an immediate boost to current inflation at any given level of current real activity. In terms of (7), such events mean higher values of $z^e_{t+\tau}$.

Another explanation is Roosevelt administration policies that fixed minimum wages, banned wage cuts, encouraged union formation, and strengthened union bargaining power (Samuelson and Solow 1960, p. 188; Weinstein 1980). There were historically unprecedented increases in the fraction of workers belonging to unions and engaging in strikes, especially over 1934–1935 and 1937–1938. It is plausible that these things were causes and effects of an increase in workers' bargaining power. In new Keynesian models, increases in worker bargaining

power correspond to "wage markup shocks" that should raise inflation at any given level of real activity (positive values of $\epsilon_{\pi P}$ and $\epsilon_{\pi W}$ in 9, 10).

The 1937 Downturn

At the end of 1936, full recovery was in sight. If output had continued to grow as it had over 1935–1936, real GDP would have been back to pre-Depression trend by the end of 1938 (Melzter 2003, p. 571). Instead there was another downturn, from a cyclical peak in May 1937. This recession was short; the trough was just a year later in June 1938. But it was sharp. Annual real GDP fell more than 3 1/4% from 1937 to 1938, more than in any postwar year. The 1937–1938 downturn has been studied much less than 1929–1933. Its causes are still unclear.

It must have been at least partly due to fiscal policy. From 1936 to 1937, real government purchases fell (Fig. 3). Taxes net of transfers rose as a fraction of national income (Fig. 4) due to three events that were all exogenous to macroeconomic conditions: the end of the veterans' bonus payment, legislation that raised personal income tax rates (the Revenue Act of June 1936), and the first collections of payroll taxes for the new Social Security system (Meltzer 2003, pp. 521, 563; Velde 2009). However, it would be hard to argue that the fiscal tightening alone can account for the magnitude of the decline in real activity (Romer 1992, p. 766).

What about monetary policy? In 1936 Fed policymakers had come to fear that the increasing supply of high-powered money must create uncontrollable inflation at some point in the future, because it was getting to be too large to drain with sales of Fed assets. According to the Burgess-Riefler doctrine, there would soon be no way to tighten policy because it would be impossible to reduce free reserves enough to force banks into the discount window. To head off this problem, the Federal Reserve Board reduced free reserves by hiking reserve requirements. In July 1936 the Fed announced a hike to become effective in August (Meltzer 2003, pp. 493–503). In January 1937 it announced more hikes to become effective in March and May 1937 (Meltzer 2003, pp. 507–510).

Meanwhile Treasury policymakers had come to fear foreign gold inflows might reverse in the future, creating uncontrollable *deflation* (Blum 1959, pp. 359–60). They developed a method by which the Treasury could sterilize gold inflows and build up a gold reserve to cover a future balance of payments deficit. The scheme was simultaneously announced and put into effect in late December 1936, at which time high-powered money growth ceased.

Together, the hikes in reserve requirements and cessation of high-powered money growth reduced free reserve supply substantially. In March 1937 short-term rates rose above the floor (Fig. 2). But they did not rise much, and the cyclical downturn came just 2 or 3 months later. The lag between an interest-rate hike and a consequent decline in real activity is usually longer than that, in postwar data at least (Christiano et al. 2005). The first hike in reserve requirements back in 1936 was not associated with an increase in short- or long-term interest rates (Fig. 2). Thus, the channel

through which reserve-requirement hikes and cessation of money growth could have actually caused the recession is not obvious. Eggertsson and Pugsley (2006) argue that the channel was the new Keynesian expected-inflation mechanism: the policy changes raised doubt about policymakers' commitment to a future increase in the price level, hence raised expected future real rates.

Conclusion

Graduate students should always be on the lookout for open questions to answer. So I conclude by pointing out some open questions about the Great Depression. In 1933, did suspension of bank payment services depress real activity by making it harder to produce and trade? How much did this matter relative to other effects of bank failures such as increases in costs of funds to credit-constrained borrowers? Why was inflation so high over 1933–1937? In 1937–1938, how did apparently small increases in short-term interest rates and a mild tightening of fiscal policy cause such a large recession so quickly? Some proposed answers to those questions rely on propositions about the response of the public's long-term inflation expectations to historically unique events. How can such propositions be tested?

References

Adam K, Padula M (2011) Inflation dynamics and subjective expectations in the United States. Econ Inq 49(1):13–25

Allen F, Gale D (1998) Optimal financial crises. J Financ 53(4):1293–1326

Allen SG (1992) Changes in the cyclical sensitivity of wages in the United States, 1891-1987. Am Econ Rev 82(1):122–140

Alogoskoufis GS, Smith R (1991) The Phillips curve, the persistence of inflation and the Lucas critique: evidence from exchange-rate regimes. Am Econ Rev 81(5):1254–1275

Ascari G (2004) Staggered prices and trend inflation: some nuisances. Rev Econ Dyn 7(3):642–667

Awalt FG (1969) Recollections of the banking crisis in 1933. Business History Review 43(3):347–371

Bagehot W(1920) [1873] Lombard street: a description of the money market. John Murray, London

Ball L, Mazumder S (2015) A Phillips curve with anchored expectations and short-term unemployment. IMF Working Papers 15(39)

Balke N, Gordon RJ (1986) Appendix B historical data. In: Gordon RJ editor. The American business cycle: continuity and change. University of Chicago Press for NBER, Chicago

Beckhart BH, Smith JG (1932) The New York money market, volume II: sources and movements of funds. Columbia University Press, New York

Bernanke BS (1983) Non-monetary effects of the financial crisis in the propagation of the great depression. Am Econ Rev 73(3):257–276

Bernanke BS (2002) On Milton Friedman's ninetieth birthday. Remarks at a conference to honor Milton Friedman, University of Chicago, Chicago, Illinois. November 8, 2002

Bernanke BS, Gertler M, Gilchrist S (1999) The financial accelerator in a quantitative business cycle framework. In: Taylor JB, Woodford M (eds) Handbook of macroeconomics, vol 1. Elsevier, Amsterdam

Bignon V, Flandrea M, Ugolini S (2012) Bagehot for beginners: the making of lender-of-last-resort operations in the mid-nineteenth century. Economic History Review 65(2):580–608

Blanchard O. (2016) The US Phillips curve: back to the 60s? Policy Brief PB16-1. Peterson Institute for International Economics
Blum JM (1959) From the Morgenthau diaries: years of crisis, 1928–1938. Houghton-Mifflin, Boston
Bordo MD, Erceg CJ, Evans CL (2000) Money, sticky wages and the great depression. Am Econ Rev 90(5):1447–1463
Bordo MD, Schwartz AJ (2004) IS-LM and monetarism. Hist Polit Econ 36(4):217–239
Bordo MD, Wheelock DC (2013) The promise and performance of the federal reserve as lender of last resort 1914-1933. In: Bordo MD, Roberds W (eds) A return to Jekyll Island: the origins, history and future of the federal reserve. Cambridge University Press, New York
Brocker M, Hanes C (2014) The 1920s real estate boom and the downturn of the great depression: evidence from city cross-sections. In: White EN, Snowden K, Fishback P (eds) Housing and mortgage markets in historical perspective. University of Chicago Press for NBER, Chicago
Brown EC (1956) Fiscal policy in the thirties: a reappraisal. Am Econ Rev 46(5):857–879
Burgess WR (1930) The money market in 1929. Rev Econ Stat 12(1):15–20
Calomiris CW, Gorton G (1991) The origins of banking panics: models, facts, and bank regulation. In: Hubbard RG (ed) Financial markets and financial crises. University of Chicago Press for NBER, Chicago
Calomiris CW, Mason JR (1997) Contagion and bank failures during the great depression: the June 1932 Chicago banking panic. Am Econ Rev 87(5):863–883
Calomiris CW, Mason JR (2003a) Consequences of bank distress during the depression. Am Econ Rev 93(3):937–947
Calomiris CW, Mason JR (2003b) Fundamentals, panics, and bank distress during the depression. Am Econ Rev 93(5):1615–1647
Calomiris CW, Wilson B (2004) Bank capital and portfolio management: the 1930s 'capital crunch' and the scramble to shed risk. J Bus 77(3):421–455
Calvo G (1983) Staggered prices in a utility-maximizing framework. J Monet Econ 12(3):383–398
Carlson MA, Wheelock DC (2016) The lender of last resort: lessons from the fed's first 100 years. In: Humpage OF (ed) Current federal reserve policy under the lens of economic history. Cambridge University Press, New York
Carter SB, Gartner SS, Haines MR, Olmstead AL, Sutch R, Wright G (2006) Historical statistics of the United States millennial edition. Cambridge University Press, Cambridge, UK
Cecchetti SG (1988) The case of the negative nominal interest rates: new estimates of the term structure of interest rates during the great depression. J Polit Econ 96(6):1111–1141
Chari VV (1989) Banking without deposit insurance or bank panics: lessons from a model of the U.S. national banking system. Federal Reserve Bank of Minneapolis Monthly Review 13(3):3–19
Chari VV, Jagannathan R (1988) Banking panics, information, and rational expectations equilibrium. J Financ 43(3):749–761
Chauvet M, Potter S (2013) Forecasting output. In: Elliott G, Granger C, Timmermann A (eds) Handbook of economic forecasting, volume 2. Elsevier, Amsterdam
Christiano LJ, Eichenbaum M, Evans CL (2005) Nominal rigidities and the dynamic effects of a shock to monetary policy. J Polit Econ 113(1):1–45
Christiano L, Motto R, Rostagno M (2003) The great depression and the Friedman-Schwartz hypothesis. J Money Credit Bank 35(6:2):1119–1197
Clapham J (1970) The Bank of England: a history, volume II. Cambridge University Press, Cambridge, UK
Clarida R, Gali J, Gertler M (1999) The science of monetary policy: a new Keynesian perspective. J Econ Lit 37(4):1661–1707
Cogley T, Sbordone A (2008) Trend inflation, indexation, and inflation persistence in the new Keynesian Phillips curve. Am Econ Rev 98(5):2101–2126
Cole HL, Ohanian LE (2007) A second look at the U.S. great depression from a neoclassical perspective. In: Kehoe TJ, Prescott EC (eds) Great depressions of the twentieth century. Federal Reserve Bank of Minneapolis, Minneapolis

Curdia V, Woodford M (2016) Credit frictions and optimal monetary policy. J Monet Econ 84(1):30–65
Diamond DW, Dybvig P (1983) Bank runs, deposit insurance and liquidity. J Polit Econ 91(3):401–419
Eggertsson GB (2008) Great expectations and the end of the depression. Am Econ Rev 98(4):1476–1516
Eggertsson GB, Pugsley B (2006) The mistake of 1937: a general equilibrium analysis. Monet Econ Stud 24(S-1):1–40
Eichengreen B (1992) Golden fetters: the gold standard and the great depression. Oxford University Press, Oxford, UK, pp 1919–1939
Eichengreen B (2004) Viewpoint: understanding the great depression. Can J Econ 37(1):1–27
Eichengreen B (2016) Doctrinal determinants, domestic and international, of federal reserve policy 1914-33. In: Bordo MD, Wynne MA (eds) The Federal Reserve's role in the global economy: a historical perspective. Cambridge University Press, New York
Eichengreen B, Mitchener KJ (2004) The great depression as a credit boom gone wrong. Res Econ Hist 22:182–237
Ennis HM, Keister T (2008) Understanding monetary policy implementation. Federal Reserve Bank of Richmond Economic Quarterly 94(3):235–263
Epstein G, Ferguson T (1984) Monetary policy, loan liquidation and industrial conflict: the federal reserve and the open market operations of 1932. J Econ Hist 44(4):957–983
Erceg CJ, Henderson DW, Levin AT (2000) Optimal monetary policy with staggered wage and price contracts. J Monet Econ 46(2):281–313
Federal Deposit Insurance Corporation (1998) A brief history of deposit insurance in the United States. International Conference on Deposit Insurance, Washington, DC, September 1998
Ferderer JP (2003) Institutional innovation and the creation of liquid financial markets: the case of bankers' acceptances. J Econ Hist 63(3):666–694
Field A (2011) A great leap forward: 1930s depression and U.S. economic growth. Yale University Press, New Haven
Fishback P (2010) US monetary and fiscal policy in the 1930s. Oxf Rev Econ Policy 26(3):385–418
Friedman M, Schwartz AJ (1963) A monetary history of the United States. Princeton University Press for NBER, Princeton
Gali J, Smets F, Wouters R (2012) Unemployment in an estimated new Keynesian model. NBER Macroecon Annu 26:329–360
Gali J (2011) The return of the wage Phillips curve. J Eur Econ Assoc 9(3):436–461
Gali J (2013) Notes for a new guide to Keynes (I): wages, aggregate demand and employment. J Eur Econ Assoc 11(5):973–1003
Gali J, Monacelli T (2005) Monetary policy and exchange rate volatility in a small open economy. Rev Econ Stud 72(3):707–734
Gertler M, Gilchrist S (2017) What happened: financial factors in the great recession. New York University working paper, October 2017
Gertler M, Karadi P (2011) A model of unconventional monetary policy. J Monet Econ 58(1):17–34
Gertler M, Kiyotaki N, Prestipino A (2017) A macroeconomic model with financial panics. NBER working paper 24126, December 2017
Goldstein I, Pauzner A (2005) Demand-deposit contracts and the probability of bank runs. J Financ 60(3):1293–1327
Gordon RJ (1990) What is new-Keynesian economics. J Econ Lit 28(3):1115–1171
Hanes C (1996) Changes in the cyclical behavior of real wage rates, 1870-1990. J Econ Hist 56(4):837–861
Hanes C (1999) Degrees of processing and changes in the cyclical behavior of prices in the United States, 1869-1990. J Money Credit Bank 31(1):35–53
Hanes C (2006) The liquidity trap and U.S. interest rates in the 1930s. J Money Credit Bank 38(1):163–194
Hanes C, Rhode PW (2013) Harvests and financial crises in gold standard America. J Econ Hist 73(1):201–246

Hausman JK (2016) Fiscal policy and economic recovery: the case of the 1936 veterans' bonus. Am Econ Rev 106(4):1100–1143
Hetzel RL (1985) The rules versus discretion debate over monetary policy in the 1920s. Federal Reserve Bank of Richmond Economic Review 71(6):3–14
Hsieh C-T, Romer CD (2006) Was the federal reserve constrained by the gold standard during the great depression? Evidence from the 1932 open market purchase program. J Econ Hist 66(1):140–176
Huang KXD, Zheng L, Phaneuf L (2004) Why does the cyclical behavior of real wages change over time? Am Econ Rev 94(4):836–856
Inklaar R, De Jong H, Gouma R (2011) Did technology shocks drive the great depression? Explaining cyclical productivity movements in U.S. manufacturing, 1919-1939. J Econ Hist 71(4):827–858
Jalil A, Rua G (2016) Inflation expectations and recovery in spring 1933. Explor Econ Hist 62(C):26–50
James JA, McAndrews J, Weiman DF (2013) Wall street and main street: the macroeconomic consequences of New York bank suspensions, 1866-1914. Cliometrica 7(2):99–130
James JA (1978) Money and capital markets in postbellum America. Princeton University Press, Princeton
Kennedy SE (1973) The banking crisis of 1933. University of Kentucky Press, Lexington
King RG (2000) The new IS-LM model: language, logic and limits. Federal Reserve Bank of Richmond Economic Quarterly 86(3):45–103
Klug A, Landon-Lane JS, White EN (2005) How could everyone have been so wrong? Forecasting the great depression with railroads. Explor Econ Hist 42(1):27–55
Kozicki H, Tinsley PA (2002) Alternative sources of the lag dynamics of Inflation. Federal Reserve Bank of Kansas City Working Papers 02–12
Krugman PR (1998) It's baaack: Japan's slump and the return of the liquidity trap. Brook Pap Econ Act 2:137–206
Mankiw NG (2010) Macroeconomics, 7th edn. Worth, New York
Mathy G, Stekler H (2017) Expectations and forecasting during the great depression: real-time evidence from the business press. J Macroecon 53(1):1–15
Meltzer AH (2003) A history of the federal reserve, volume I. University of Chicago Press, Chicago
Miller AC (1935) Responsibility for federal reserve policies: 1927-1929. Am Econ Rev 25(3):442–458
Mishkin FS (1978) The household balance sheet and the great depression. J Econ Hist 38(4):918–937
Morris S, Shin HS (2000) Rethinking multiple equilibria in macroeconomic modeling. NBER Macroecon Annu 15:139–161
Nelson DB (1991) Was the deflation of 1929-30 anticipated? The monetary regime as viewed by the business press. Res Econ Hist 13:1–65
O'Brien AP (1989) A behavioral explanation for nominal wage rigidity during the great depression. Q J Econ 104(4):719–735
Ohanian LE (2001) Why did productivity fall so much during the great depression? Am Econ Rev 91(2):34–38
Ohanian LE (2009) What - or who - started the great depression? J Econ Theory 144(6):2310–2335
Olney M (1999) Avoiding default: the role of credit in the consumption collapse of 1930. Q J Econ 114(1):319–335
Orphanides A (2003) Historical monetary policy analysis and the Taylor rule. J Monet Econ 50 (5):983–1022
Rees A (1961) Real wages in manufacturing 1890–1914. Princeton University Press for NBER, Princeton
Richardson G (2007) Categories and causes of bank distress during the great depression, 1929-33: the illiquidity versus insolvency debate revisited. Explor Econ Hist 44(4):588–607
Richardson G, Troost W (2009) Monetary intervention mitigated banking panics during the great depression: quasi-experimental evidence from a federal reserve district border, 1929-1933. J Polit Econ 117(6):1031–1073

Rockoff H (1984) Drastic measures: a history of wage and price controls in the United States. Cambridge University Press, New York
Rockoff H (1993) The meaning of money in the great depression. NBER working paper H0052
Romer CD (1992) What ended the great depression? J Econ Hist 52(4):757–784
Romer CD (1993) The nation in depression. J Econ Perspect 7(2):19–39
Romer DH (2006) Advanced macroeconomics, 3rd edn. McGraw-Hill, New York
Romer CD, Romer DH (1989) Does monetary policy matter? A new test in the spirit of Friedman and Schwartz. NBER Macroecon Annu 4:121–170
Romer CD, Romer DH (2013) The missing transmission mechanism in the monetary explanation of the great depression. Am Econ Rev 103(3):66–72
Rose JD (2010) Wage rigidity in the onset of the great depression. J Econ Hist 70(4):843–870
Rotemberg J (1982) Monopolistic price adjustment and aggregate output. Rev Econ Stud 49 (4):517–531
Samuelson PA, Solow RM (1960) Analytical aspects of anti-inflation policy. Am Econ Rev Paper Proc 50(2):177–194
Shleifer A, Vishny R (2011) Fire sales in finance and macroeconomics. J Econ Perspect 25(1):29–48
Smets F, Wouters R (2007) Shocks and frictions in US business cycles: a Bayesian DSGE approach. Am Econ Rev 97(3):586–606
Steiner WH (1926) Paper eligible for rediscount at federal reserve banks: theories underlying federal reserve board rulings. J Polit Econ 34(3):327–348
Svensson L (2003) Escaping from a liquidity trap: the foolproof way and others. J Econ Perspect 17(4):145–166
Svensson L (2004) Comment on Bernanke, Reinhart and Sack. Brook Pap Econ Act 2:84–93
Taylor JB (1999) A historical analysis of monetary policy rules. In: Taylor JB (ed) Monetary policy rules. University of Chicago Press for NBER, Chicago
Temin P (2000) The great depression. In: Engerman SL, Gallman RE (eds) The Cambridge economic history of the United States. Cambridge University Press, Cambridge, UK
Temin P, Wigmore BA (1990) The end of one big deflation. Explor Econ Hist 27:483–502
U.S. Federal Reserve Board of Governors (1943) Banking and monetary statistics. National Capital Press, Washington, DC
Velde FR (2009) The recession of 1937-a cautionary tale. Federal Reserve Bank of Chicago Economic Perspectives 33(4):16–37
Watanabe S (2016) Technology shocks and the great depression. J Econ Hist 76(3):909–933
Weinstein M (1980) Recovery and redistribution under the NIRA. North-Holland, Amsterdam
Weir DR (1992) A century of U.S. unemployment, 1890-1990. Res Econ Hist 14:301–386
Wheelock DC (1990) Member bank borrowing and the fed's contractionary monetary policy during the great depression. J Money Credit Bank 22(4):409–426
Wheelock DC (1991) The strategy and consistency of Federal Reserve monetary policy. Cambridge University Press, Cambridge, UK, pp 1924–1933
White EN (1990) The stock market boom and the crash of 1929 revisited. J Econ Perspect 4(2):67–83
Wicker E (2000) Banking panics of the gilded age. Cambridge University Press, Cambridge, UK
Wigmore BA (1987) Was the bank holiday of 1933 caused by a run on the dollar? J Econ Hist 47(3):739–755
Willis HP (1936) The theory and practice of central banking. Harper and Brothers, New York
Willis PB (1957) The federal funds market: its origin and development. Federal Reserve Bank of Boston, Boston
Wingfield BM (1941) Deterrents to membership in the Federal Reserve System. In: Banking studies. Federal Reserve Board, Washington, DC, p 1941
Woodford M (2010) Financial intermediation and macroeconomic analysis. J Econ Perspect 24(4):21–44

Central Banking

Jon Moen

Contents

Abstract

The study of central banking is an important part of cliometrics today. However, economists and bankers have been studying central banking since well before the cliometric revolution began. Early eighteenth century British economists were studying the structure of the Bank of England in detail, and economists were closely involved in the debates leading up to the creation of the Federal Reserve System. Starting in the 1970s and 1980s, cliometricians began studying central banking and related issues in great detail. The study of bank panics of the National Banking Era and Great Depression in the United States has revealed that lender of last resort activities have been undertaken rather timidly, ranging from the incomplete responses of clearinghouses to the failure of the Federal Reserve during the Depression. Cliometricians have also added several large and rich sources of evidence to the study of central banking.

J. Moen (✉)
Department of Economics, The University of Mississippi, Oxford, MS, USA
e-mail: jmoen@olemiss.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_50

Keywords
Clearinghouse · Cliometrics · Loan certificates · Real bills · Federal Reserve

Introduction

Cliometricians have not always been involved in the study of central banking. But the cliometric approach, that of using the economic and quantitative tools available at the time, has been used from early on. Even before there were formally recognized economists, banking practitioners were as much involved in the analysis of the effects of central banks as were historians. This sets the study of central banking apart from the study of other fields, such as slavery or railroads, fields that were eventually dominated by cliometrics and cliometricians. To examine the contributions of cliometrics and cliometricians to the study of central banking, I look primarily at the economic study of central banking in the United States, prefaced with a brief foray into eighteenth and nineteenth century Britain.

The functions of a central bank can be divided broadly into banking versus monetary policy. Banking policy deals with functions like note issue, check clearing, lending, discounting, and emergency lender of last resort functions. Monetary policy deals more with macroeconomic issues. Before the twentieth century, monetary policy tended to deal with the relationship between note issue and the price level or exchange rates. Issues related to modern macroeconomic theory and policy, especially a central bank's ability to alter business cycles and economic activity in general, became predominant after the Great Depression and Keynes' General Theory. I will leave modern macroeconomic theory to macroeconomists and will focus on banking policy and early monetary policy.

Cliometricians have made several important contributions to the understanding of central banking. The first set of contributions deals with the structures and institutions of central banking. As important precursors to central banking in the United States, clearinghouses and the correspondent banking system are all better understood because of cliometrics. We have a clearer understanding of bank panics, what happens during a panic, the actions of lenders of last resort that may or may not have helped and, hence, these same behaviors in more recent panics. Cliometricians have also made new sources of evidence available to researchers, either by discovering new sources or by making old sources computer readable. Having a better grasp of what happens during panics also sets the stage for better theories of panics.

I will start with a brief survey of pre-cliometric studies of banks and central banks. A mainly chronological survey of cliometrics and central banking will follow, starting with the antebellum United States. While the postbellum United States had no formal central bank, institutions arose that took on some central bank-like functions. In particular, I will examine the cliometric study of clearinghouses and their nascent central bank functions during the postbellum period. The actions of the New York Clearing House during the Panic of 1907 influenced the subsequent creation of the Federal Reserve System in the panic's aftermath. The remainder

of the survey examines cliometric studies of the Fed and central banking. It finishes with a review of databases relevant to central banking contributed by cliometricians.

Early Studies of Central Banking

The systematic study of central banking and monetary policy by contemporary observers goes back to at least the late eighteenth century. Henry Thornton and Thomas Tooke were early nineteenth century English economists and bankers who were involved in the bullionist controversy. Politicians in Great Britain were concerned about the effects of the issuance of paper currency on the price level and exchange rates during the Napoleonic Wars after Britain had suspended gold convertibility in 1797. This was a debate about the relationship between the level of prices and changes in issue of paper money by the Bank of England. The bullionists argued that convertibility was necessary to prevent the over-issue of currency relative to the gold in the bank's vault that otherwise would result in inflation. The anti-bullionists argued that the issue of banknotes was limited by the fact that they were issued in exchange for bills of exchange (real bills). Thornton's *An Enquiry into the Nature and Effects of the Paper Credit of Great Britain* (1802) is an early look at the relationship between currency issue and prices. He takes the anti-bullionist position to defend the bank of England from charges of excessive issue, although he argued that issuing currency for bills of exchange was not a sure guarantee against inflation. Adam Smith, in contrast, was a strong supporter of the anti-bullionist position. Thomas Tooke (1844) was more in the bullionist camp, supporting a return to convertibility, although he opposed the strict limits imposed on the Bank's discretion to issue currency by the Bank Charter Act of 1844. Both Thornton and Tooke held the position that bank-created money was the result of increases in prices, not the cause. Robert Hetzel provides a fine overview of Thornton as a central banker (Hetzel 1987).

David Ricardo was well within the bullionist camp and gold convertibility. He also presented a somewhat obscure plan for a formal central bank in Great Britain, *Plan for the Establishment of a National Bank* (1824). At that time, the bank of England was neither a central bank nor lender of last resort. Rather, it consisted of two sections, a bank of note issue and a lending bank whose customers included the government of Great Britain. Ricardo proposed splitting the two functions, making the government, or national, bank the sole issuer of paper currency supervised by an independent board of commissioners. Such a split would have no negative economic effects, but in Ricardo's view it would allow for better regulation of note issue. Quite recently, the bullionist controversy has been revisited by Joshua Hendrickson (2018). Using Bayesian econometrics, he provides support for the bullionist position. Pamfili Antipa (2014) comes to similar conclusions.

Several decades later, Walter Bagehot, a journalist and economist, examined the issue of currency and prices in *Lombard Street* (1873), arguing for a larger

reserve at the Bank of England. Here he made his well-known prescription for fighting a bank panic – that the bank of England should lend freely at a high rate of interest. As Milton Friedman and Anna Schwartz (1963) have argued in the context of the United States during the Great Depression, this prescription does not seem to have been followed that well, although the US Federal Reserve seems to have finally learned the lesson. Under the leadership of Chairman Ben Bernanke, in response to the disruption to financial markets in 2008/2009, the Fed did lend more freely than in the past and to intermediaries that had not been narrowly defined as banks. It appears to have barely lent just enough to have avoided a collapse of banking and the money supply, unlike during the Great Depression.

Even in history of thought, the early economists were contributing to the understanding of the relationship between money and the price level and short run vs. long run neutrality at a theoretical level. This goes back to at least David Hume (1985) and later Knut Wicksell (1934). See Blaug(1996) for an overview of the evolution of early banking theory. Both emphasize that changes in money can have short run effects on some real activity, but in the long run the effects are neutral and only affect the price level. All of these early discussions relate to the relationship between money and the price level, or level of various prices. How money changed real activity and the business cycle was not yet in the forefront of economic discussion.

Antebellum US Central Banking

Early in US history the role and impact of a national bank were widely discussed. Alexander Hamilton and Andrew Jackson, before there were even economists in the United States, sparred on the effects of the Bank of the United States on credit to small business owners and farmers in addition to the extent of the inflationary effects of centralized note issue. Richard Timberlake (1993) notes that the First Bank of the United States' high reserve ratio allowed it to manage currency and credit availability. It could restrain credit by presenting state bank notes for redemption in specie, while it could increase credit by loosening lending to businesses. It was, after all, in part a private bank, and the fact that it was competing with state banks contributed to its charter not being renewed.

The Second Bank of the United States was structured in ways similar to those of the first bank. One difference was that under the more active leadership of Nicholas Biddle, the Bank appeared to have been more aggressively managing credit. But Timberlake (1993) points out that the Bank was actively maintaining the workings of the metallic standard and the movements of international exchange, which may have been interpreted as intervening in credit markets for its own ends.

Jane Knodell (2003) has shown that the Second Bank of the United States was able to balance the trade-off between private profitability and performing its services for the federal government. Its large-scale dealings in domestic exchange and foreign transactions allowed it to profitably and efficiently provide regional and international payments, a basic function of a central bank.

In the midst of this period, the Suffolk banking system (1825–1858) flourished in the New England area as what may be referred to as an "operational" central bank, especially in the absence of the Second Bank after 1837. Although the key features of the system foreshadowed some characteristics of modern central banking procedures (lending reserves to other banks, maintaining the operation of the payments system, presenting the notes of country banks for redemption as a means to control note issue), the Suffolk system was not at the time viewed as a model for a central bank. Research by Rolnick et al. (1998) suggests that the actions of the Suffolk Bank during the Panic of 1837 helped the New England economy fare better than the rest of the country. New England banks were able to maintain loans and maintain the operation of the regional payments system. The Suffolk Bank may have played a useful monitoring role for the member banks in the system as well. Calomiris and Kahn (1996) show that the Suffolk system operated efficiently, with little rent-seeking or exploitation of the smaller, country banks within the note exchange system. Howard Bodenhorn (2002), however, questioned the ability of the Suffolk system to control note issue.

Postbellum Central Banking and Clearinghouses

As mentioned above, the Suffolk system provided some of the elements of central banking to the banks of its region, but these functions supplied mainly payments system efficiency without the central bank features of liquidity creation. From 1837 to 1914, the United States had no formal lender of last resort, an unusual feature for an advanced economy at this time. Esther Taus (1943) and Richard Timberlake (1993) describe how the Treasury employed some limited, central bank-like powers before the Federal Reserve System. Timberlake (1993) notes that the pre-Civil War independent Treasury developed monetary policy tools like open market operations.

Of these early central bank-like institutions to arise, the most important are the clearinghouses that appeared across the United States in response to a system of unit banking in the absence of a formal lender of last resort. Before the Federal Reserve System, the United States nevertheless had institutions in the form of the clearing houses across the country that in many instances served as proto-central banks. The most prominent is the New York Clearing House, although others have been studied as well. An early analysis of the Clearing House was carried out by O.M.W. Sprague (1910), writing for the National Monetary Commission in his famous *History of Crises under the National Banking System*. He highlights the unfortunate delay of the Clearing House in issuing clearing house loan certificates during the early stages of the panic of 1907. James Cannon, a president of the NYCH, produced a volume analyzing the structure of the more important clearing-houses in several large US cities (Cannon 1910). Even though Sprague and Cannon did not write much about central banking directly, the clearinghouses have been viewed as precursors to a modern central bank and as models for the Federal Reserve's initial decentralized structure. These two early authors have provided the

foundation for much of the modern, cliometric analysis of the central bank – like functions of clearinghouses.

Cliometricians have expanded our understanding of clearinghouses, building on the early work of Cannon and Sprague. Timberlake (1984) and Gorton (1985) describe how private and commercial banks formed among themselves a system of clearinghouse associations, starting with the New York Clearinghouse in 1853. Timberlake (1993) notes that as early as 1857 the New York Clearing House began exerting power in ways akin to a central bank through the issuance of clearinghouse loan certificates as a means to increase liquidity in the short run. The Clearing House also developed powers of examination of its members that resemble those of the Comptroller of the Currency and the Federal Reserve today. These responses to a banking system without a central bank were limited by the legal restrictions on banks and on the limitation of legal tender status to specie alone for tax payments. Furthermore, these measures to expand liquidity were temporary and regional at best, as the New York Clearing House could do little to expand liquidity outside of New York. Richard Timberlake (1978, 1984, and 1993) is most notable for his analysis of clearinghouses. He agrees that clearinghouses were providing some central bank-like functions. But he goes beyond that result and comes to the conclusion that with some minor modifications, clearinghouses could have adequately provided central bank or lender of last resort services without the establishment of a formal central bank like the Federal Reserve.

He points out that federally chartered clearinghouses in each state could have provided emergency liquidity during panics about as well as a government run central bank. The use of clearinghouse loan certificates could have been expanded to serve as the source of emergency liquidity. During panics, clearinghouses, New York's most notably, would allow banks to present securities to serve as collateral for the issuance of a quasi-currency that would serve as a settlement medium, one that would circulate among banks in place of legal tender or specie. The released legal tender and specie could then be used to pay panicked depositors, or more likely panicked correspondent banks in the interior. Critics of loan certificates viewed them as a poor substitute for actual currency, owing to the illegality of circulating the certificates to the general public. Timberlake asserts that just the opposite was true, given how readily they circulated among the public in cities where they were allowed to circulate in 1907. New York was a key exception, letting them circulate only among the member banks of the Clearing House. Even legal authorities in the Treasury ignored the illegality, given the obvious function they were providing to keep transactions continuing. Still, as Timberlake notes, the appearance of illegality and a "hucksterish" quality fueled the desire for an official currency, and the Fed and its notes prevailed.

Gary Gorton (1985) argues that the rise of the clearinghouse was in part due to the increasing use of checks and demand deposits in the middle of the nineteenth Century. Unlike banknotes, for which the private sector had produced methods of pricing notes, such as the regular publication of bank note detectors, it was much more difficult for the private sector and depositors to value or price individual demand deposits owing to their being a claim not just on an individual bank but also

on an individual's deposit. A secondary market was much less likely to arise on deposits than on notes, and one never developed. Gorton and Mullineaux (1987) elaborate on the point that the clearinghouse arose as a means to overcome the information asymmetries involving higher monitoring costs for deposits compared to banks and banknotes. Moen and Tallman (2015) argue that the big six national banks in New York City appear to have acted less in coordination as a central bank during the Panics of 1873, 1884, 1890, 1893, and 1907 than has been popularly believed. They show that the use of clearing house loan certificates during these panics appears to be much more bank-specific, with some issuing large volumes while others issued none.

This discussion foreshadows Gorton's (2010) more recent emphasis on the role of information sensitive and insensitive assets in the financial crisis of 2008. Information insensitive assets are simply those whose value is immediately and easily recognized, like currency or federal government bonds. During a crisis or panic, investors increase their demand for information insensitive assets to shed risk. Gorton also expands the analysis of the role of shadow banking in the recent financial crisis. He argues that it was the collapse of overnight lending in the Repo market that was the ultimate source of financial collapse in 2008, not the housing bubble itself.

Moen and Tallman (1992, 2000) show that the Panic of 1907 was surprisingly severe in the eyes of New York bankers because the panic was focused on trust companies, intermediaries outside of the formal protection of the New York Clearing House. They were participating in financial markets along with Clearing House members, the national banks, a type of shadow banking described by Gorton. They also show that the more complete coverage of all intermediaries, including trust companies, by the clearinghouse in Chicago led to much less disruption in intermediation than in New York. They highlight the role that incomplete coverage by a lender of last resort, in this case the New York Clearing House, can have on the market. In 1907 the clearing house did not protect state-chartered trust companies, which operated in the call loan market alongside the national banks in New York. They became the source of panic because they were not members of the Clearing House.

Christopher Hoag (2012) looks at Clearing House behavior and the use of clearing house loan certificates in the Panic of 1893. He shows that banks did not suffer from moral hazard when borrowing with loan certificates and that pre-panic balance sheet characteristics do not explain borrowing through loan certificates. Of course, during a panic even otherwise healthy banks would want to conserve on their legal tender reserves and use loan certificates instead. He also argues that Clearing House membership in New York did not mitigate deposit contractions during the panic.

Quite recently, Vincent Bignon (2017) has shown that clearing houses can fail, calling into question the ability of these private sector institutions to provide reliable lender of last resort functions. While he analyzes a stock market clearing house, his insights point to the consideration that a private sector lender of last resort may not be able to summon the resources to provide a substantial boost to net liquidity during a crisis.

The National Banking Era and Currency Reform

In the United States, the later nineteenth century was a hotbed of proposals for currency reform in light of the perceived "inelastic supply" of national bank notes. National bank notes were backed by Federal government bonds, the supply of which had no clear relationship to economic activity. This problem became an issue during the recurring bank panics of the later nineteenth Century. The keen emphasis on currency reform, however, diverted attention from creating a new lender of last resort, at least until 1907. But examining issues around currency reform helps illuminate the role of clearinghouses and the debate over the structure of the Federal Reserve System and the means to increase liquidity during a crisis. The function of clearinghouse loan certificates is important here. Viewing the amount of federal debt outstanding as a constraint on national bank note issue, a movement began to take hold that recommended replacing the government bond-backed National Bank notes with other notes backed by short-term assets of individual banks like commercial paper or acceptances. It was thought that such asset backing would allow the currency to expand and contract with business conditions. A secondary market for commercial paper in the United States never really developed under National Banking, and such a market never really became established until the Federal Reserve System. In addition, the acyclical nature of federal debt was credited with failing to ensure adequate volume of backing for the provision of liquidity and credit to commerce. The inelasticity of national bank note issue was left unexamined until recent research cast some doubt on the allegations. Charles Calomiris and Joe Mason (2008) show that national bank notes were never issued up to the maximum legal limit and that it was bank specific reasons that kept some banks from issuing more notes, not the volume of government bonds that could be used as backing for new note issue.

"Asset-backed" currency proposals began to appear regularly in public discussions in the 1890s. Essentially, these proposals were aimed at producing a financial system akin to the monetary system in Canada, but without the branch banking powers of Canadian banks. At the American Bankers Association Convention of 1894 in Baltimore, one plan arose to amend the National Bank Acts by replacing the bond-backed notes with notes backed by the short-term assets of the issuing bank. A guarantee fund was also to be established at the Treasury to redeem the notes of a failed bank. The support for such a currency came from the Laissez Faire branch of the Democratic Party identified by Richard McCulley (1992). Robert Craig West (1974) notes that some prominent New York bankers opposed the "Baltimore Plan." These bankers feared that without restrictions or powerful monitoring, questionable bank assets could be used to secure currency issues. James Livingston (1986) notes that New York City bankers would not support an asset backed currency proposal without branch banking provisions. Politically, neither proposal had potential as legislation. An asset backed currency was a cornerstone of monetary reform proposals coming out of the Midwest, particularly those of J. Laurence Laughlin, the first head of the Economics Department at the University of Chicago, and James B. Forgan, president of the First National bank of Chicago (McCulley 1992).

Despite the lack of early success in legislative action, the Baltimore Plan and the asset-backed currency issue did, nevertheless, influence later reform measures.

Most notable among them, the plan proposed by the Indianapolis Monetary Commission in 1898, written by a key member of the executive committee of the convention, J. Laurence Laughlin. It is therefore not surprising that his ideas were central to this proposal.

Laughlin was strongly against a bimetallic standard and favored gold, but he was also a proponent of an asset-backed note-issue. The Indianapolis Monetary Commission plan included a 5-year phasing out of the bond-backed national bank notes, to be replaced with asset-backed notes issued by the individual banks. The plan included a five percent guaranty fund to redeem the notes of failed banks, and it even included some branching provisions. But, it did not significantly alter the system of reserves present under the National Bank Act. A movement towards a central bank with a compelling message had yet to acquire a politically effective constituency.

The extended discussion over currency reform tended to overwhelm any discussions, academic or political, of establishing a formal central bank until the Panic of 1907 shocked the New York banking community into pushing for a central bank. The panic in New York City had been focused on the trust companies, state-chartered intermediaries outside the payments system controlled by the Clearing House, and competitors to the national banks. In the aftermath of the panic, the large New York national banks realized that even though they may have profited during earlier panics, the scale of the recent panic had been enough to convince the bankers that a panic more severe than that of 1907 would be beyond the resources of the New York Clearing House banks. They just didn't control enough liquidity to act like a lender of last resort in a more severe panic. Private financiers like Morgan and Rockefeller had to put up their own funds to support the stock market and call loan market until gold flows from Europe reached New York banks. Hence, the New York bankers threw their support behind the effort to create a source of liquidity and reserves that was outside of the banking system, and the creation of what became the Federal Reserve System was underway (Moen and Tallman 2007). Like policy debates today, the debates over the structure of the new central bank included concern over intermediaries outside of the traditional banking system, trust companies in particular, and overnight lending on the call loan market.

The National Monetary Commission and the Fed

One legislative outcome spurred by the disruption of the Panic of 1907 was the creation of the National Monetary Commission. The National Monetary Commission included economists and bankers, and it produced the series of monumental volumes on various aspects of banking, including a large volume of data. The immediate goal of the Commission was to provide background information and data from other countries with central banks to be used in the creation of a central bank for the United States. Their use as a critical academic reference point was not initially apparent or important. The volumes nevertheless remain important, for cliometricians today still rely on this data when investigating the panic and other aspects of the National Banking System. A. Piatt Andrew and O.M.W. Sprague were

the key economists serving on the commission. Andrew is noted for his research into currency substitutes used during the panic (Andrew 1908, 1910), while Sprague produced the key historical work on panics, *A History of Crises under the National Banking System* (1910). This particular volume is still heavily referenced by cliometricians. Commission member Paul Warburg, a banker rather than an academic economist, influenced the structure of the early Federal Reserve. Even though he was well versed in European central banking, he knew that only a decentralized system of last resort reserves would have any chance of success in the United States (Warburg 1930). A centralized bank along the lines of those in Europe would have no chance of gaining political support, as Americans would likely view a centralized bank as easily captured by Wall Street banking interests at the expanse of small businesses and farmers. Mehrling (2016) provides a survey of the influence of US economists contemporary at the time, including economist Irving Fisher and Lauchlin Currie, Marriner Eccles' personal assistant at the Board of Governors of the Federal Reserve System, in the design of the Federal Reserve System.

The Federal Reserve was more closely examined at the time by economists and politicians after its apparent ineffectiveness in expanding the money supply during the early stages of the Great Depression. The most notable was Lauchlin Currie, an economist trained at the London School of Economics. In *The Supply and Control of Money in the United States* (1935), Currie outlined what he perceived as the weaknesses of the original, decentralized Federal Reserve System. The weaknesses included the lack of universal membership, decentralized reserves, and lack of central control over the system's reserves. The absence of universal membership in the original Federal Reserve System resulted in intermediaries operating in the same markets as Fed members, but having no access to the Fed's liquidity. This would return to hinder the Fed's effectiveness during the Great Depression, during which state-chartered banks failed in disproportionately larger numbers than member Fed banks (Wicker 2000). Panic at the nonmember state-chartered intermediaries could drag down Fed member national banks in a pattern similar to that seen in the Panic of 1907, in which the non-Clearing House trust companies threatened the member national banks with the near collapse of the call loan market. Working with Treasury and Federal Reserve officials, Currie wrote much of the Banking Act of 1935, which gave the Federal Reserve System the centralized structure controlled by the Board of Governors we are familiar with today. Universal membership, however, still was not part of the Act. As noted earlier, the Fed stepped in to provide lender of last resort services to non-members and even non-banks in 2008/2009; some lessons can be learned.

Extended Histories of Central Banking and the Fed and the Rise of Cliometrics

Up to this point, it is clear that early economists and bankers had been influential in the study and structuring of central banks in Europe and the United States. While the application of cliometrics to the study of central banking was not as sharp

a distinction from "old" economic history practices as that seen in the study of railroads or slavery, a shift did appear. It brought explicit use of economic theory and econometrics to the study of central banking and for academic reasons, apart from those spurred by responses to crises or calls for reforms, as in the case of the currency debate.

If identifying a starting point for the cliometric study of central banking is important, then the cliometric "revolution" could have its inception loosely linked to the publication of Milton Friedman and Anna Schwartz's widely read book *A Monetary History of the United States from 1867 to 1960* (1963). It is a detailed study of the components of the money supply, the monetary base, velocity, and the structure of the US banking system. At least in the short run, it links movements in the money supply to movements in the economy. Friedman and Schwartz also provide estimates of the various components of the money supply from 1867 to 1960 in Appendix A of *The Monetary History*. They provide a detailed explanation of how the series were calculated from a variety of sources. Their constructed series include Currency Held by the Public and Deposits at Commercial Banks: Bank Reserves, including vault cash and deposits at Federal Reserve Banks; Federal Government Balances at the Treasury, Commercial and Savings Banks, and later at Federal Reserve Banks; US International Capital Movements; and estimates of the Velocity of Money. A subsequent volume to *The Monetary History*, *Monetary Statistics of the United States* (1970) provides more discussion of the construction of the 1867 to 1960 series. It also discusses extant measures of the money stock before 1867 and provides material for the construction of monetary series extended back to 1775. In their *Monetary trends in the US and the UK* (1984), they extend their analysis to the United Kingdom. Because *The Monetary History* was so influential in the study of the Great Depression, the critical chapter, The Great Contraction, 1929–1933 was later reissued as a stand-alone book, *The Great Contraction* (2008).

The Monetary History provides a comprehensive analysis of the Federal Reserve's conduct of monetary policy during the early stages of the Great Depression and beyond. Friedman and Schwartz describe the monetary policy followed by the Fed after 1929 as nothing short of "inept" when compared to its vigorous and active policy before 1929. They lay much of the blame on the shift of power within the Fed away from the New York bank to the Board of Governors in Washington D.C. The problems incurred by this shift were compounded by the death of Benjamin Strong in 1928, who had been governor of the New York bank since the start. He understood banking during panic conditions, having aided J.P. Morgan during 1907, and he was familiar with European central banking practices. His death, famously described by Friedman and Schwartz as "untimely," deprived the Federal Reserve of the expertise needed to deal with crisis conditions, with Fed policy makers reverting to a commercial banking mentality, one fraught with conservative, private sector responses to banks in trouble. Unlike the policy proffered by Bagehot for a central bank to lend freely on good collateral during a crisis, the Federal Reserve leaders contracted lending, as a commercial bank would during a crisis. Friedman and Schwartz also castigate some Fed officials for

believing that some banks should be weeded out during the Depression as a punishment for bad or careless lending during the headier days of the Roaring Twenties. For an in depth and extensive review of all the contributions of *The Monetary History,* see Michael Bordo (1989).

After *Monetary History,* more detailed and extended histories of central banks and the Federal Reserve appeared. Richard Timberlake's *The Origins of Central Banking in the United States* (1978) is a significant contribution to our understanding of how central banking came about in the United States. He points out that most economies go through stages of development in money and banking markets that do not inevitably lead to a central bank fully capable of controlling the money supply. A specie standard usually evolved to formalize a monetary system. Such systems were sometimes supplemented by banks that acted as the fiscal agent for a government, and while they were sometimes called a central bank, that did not mean they also were able to control the money supply. By the twentieth century, central banks as we understand them had come into being, with metallic standards eventually being replaced with fiat standards. But a central point in his argument is that they did not spring immediately into being fully formed. For example, the Federal Reserve System started out as a collection of banks modelled on the clearing houses and did not really become the central bank we know until the 1935 Banking Act centralized power in the Board of Governors. A subsequent and extension of this book, *Monetary Policy in the United States* (1993), adds a chapter on the central banking role clearinghouses before the Federal Reserve. It then extends the examination of monetary policy after the Federal Reserve up through the 1980s. In the end, he discusses what the Fed can and cannot do and makes a well-argued case for allowing more private sector control of the money supply. One example being that an extended clearinghouse system could have controlled the money supply during normal and panic conditions.

Charles Goodhart produced an important book, among several, on the importance and benefits of a central bank, *The Evolution of Central Banks* (1988). He outlines the historical arguments for and against a central bank, comparing Thornton and Bagehot's arguments for and against a system of decentralized reserves. He then discusses how central bank-like functions can arise within the private banking system, an obvious example being the rise of private clearinghouses. But he argues that there are limits to how far the private sector can go in providing these services, as self-interest can run up against efforts to serve the broader good. This problem can be significant if there are several large banks also competing with each other in more normal times, weakening the ability to create a private club whose members have competing interests. A natural solution was the creation of a non-competitive central bank unhindered by a narrowly defined profit motive, the evolution of the Bank of England being a prime example, subsequently followed by other countries when establishing their central banks.

Richard Timberlake (1993) argued that a central bank as described by Goodhart was not a necessary development in a modern economy. The expansion of the postbellum clearing house system in the United States would have provided a more than adequate solution to the lack of a secondary source of liquidity. A recent

article by Tallman and Jacobson (2015) provides some support for Timberlake's view that improvements to the clearinghouse arrangements may have been enough to provide adequate liquidity during a panic. They show that the new emergency currency authorized by the Aldrich-Vreeland Act used in conjunction with clearinghouse loan certificates successfully alleviated the liquidity crisis of 1914 brought about by the prospect of World War I. The emergency currency was issued by groups of national banks to provide liquidity among national banks, while loan certificates were used to provide liquidity to state banks and trust companies, intermediaries that could not borrow the emergency currency issued by the national banks.

Elmus Wicker has also made substantial contributions to our understanding of what happened during bank panics during the national banking Era in the United States and during the Great Depression. He has also provided us with a concise analysis of the political maneuverings leading up to the founding of the Federal Reserve System. *The Banking Panics of the Gilded Age* (2000) updates Sprague's *History of Crises* by applying more modern economic analysis and data collection. In *The Banking Panics of the Great Depression* (1996), Wicker tempers the Friedman and Schwartz "inept" behavior argument. He argues that Federal Reserve officials misinterpreted the increase in the currency to deposit ratio as showing monetary expansion. Today we realize that increase reflects a curtailment in deposit expansion and the money supply, even though the increase in currency in circulation initially gives the appearance of more money. Wicker argues that the proper interpretation of changes in the ratio was not yet clearly a part of monetary theory as it related to the multiple expansions of deposits, at least among banking practitioners, if not economists. Thus, he places more blame on limited knowledge rather than outright ineptness.

Wicker's book, *The Great Debate* (2005), provides a very accessible history of the politics leading up to the founding of the Fed. While Carter Glass conventionally is given credit for the legislation creating the Fed, Wicker shows that Senator Nelson Aldrich's bill proposing the creation of the National Reserve Association was heavily relied on in the creation of Glass's legislation. Aldrich's proposal arose out of his work on The National Monetary Commission, so his bill's influence is not surprising in subsequent legislation. Wicker's book also provides a useful review of recent literature on the founding of the Federal Reserve, much of which is covered in this paper. But he also goes to some length to analyze the contributions of two historians, James Livingston (1986) and Richard McCulley (1992), and the political scientist J. Lawrence Broz (1997). Their intensive use of archival materials and the papers of those responsible for crafting the final structure of the Fed, those of Glass, Willis, Aldrich, Warburg, Vanderlip, and Laughlin, is what set their research apart from earlier work.

Wicker argues that these three books by non-economists are important, as they help understand the politics underlying the founding of the Fed rather than the narrower economics. Livingston argues that the increasing centralization of corporate business in the United States led to the demand for a central bank to aid the banking sector in funneling funds into the corporate sector. Broz emphasizes the importance of a central bank in establishing the international importance of

the United States. In particular, he argues that a central bank was necessary for the US dollar to become the dominant international currency. McCulley, in Wicker's view, adds to our understanding of the political origins of the Fed by employing a three-way classification of the main interest groups and the legislative proposals emanating from those groups: Wall Street, LaSalle Street, and Main Street. While LaSalle Street was motivated by currency reform and establishing an asset backed currency, Wall Street interests were more concerned with minimizing government control of the central bank, a motivation that fortuitously was shared by both the LaSalle and Main Street interests. Wicker is clear to maintain, however, that these three authors contributed little to the economic analysis of the banking functions of a nascent central bank.

In the 1970s and 1980s, the cliometric approach to analyzing central banking really took off, and card-carrying cliometricians began to make their presence known. Friedman and Schwartz extended their work from *Monetary History* with several other volumes of data and analysis of banking in the UK (Friedman and Schwartz 1970, 1984). Anna Schwartz also contributed more research into panics and central banking responses well after the publication of *Monetary History*. From the perspective of lender of last resort duties, her paper "Real and Pseudo-Financial Crises" (1986) is required reading for anyone studying financial crises and bank panics. She makes a clear distinction between a real and a pseudo financial crisis. A pseudo crisis is one in which the equity of a financial or business firm is at risk of collapsing. A real crisis also involves a direct threat to the payments system, and hence the banking system, from the failure of one or more financial firms at great risk. Her prescription for the lender of last resort is that no action should be taken during a pseudo crisis, other than to perhaps reassure markets that the risk is contained to the specific firm. Excessive aid to firms in a pseudo crisis may resort in inflation or increasing instances of moral hazard. In a real crisis where payments could suspend and real effects spread to otherwise unrelated firms, the lender of last resort needs to assure that deposits will be convertible to currency regardless of what happens; failure of mismanaged banks or other intermediaries still could be allowed to happen.

Michael Bordo is one of the most prolific monetary historians today. Much of his research has concentrated on the workings of the classical gold standard and other exchange rate regimes. He has also examined in detail the role of the money supply in economic activity. For the purposes of this essay, I will focus narrowly on his contributions to the study of central banking and bank policy related to panics, along with the origins of central banking.

An early foray into central banking looked at why Canada established a central bank later than most countries. Bordo and Redish (1987) ask why it took until 1935 for a central bank to appear in Canada. They examine three explanations for the late appearance. The first was that a competitive banking system needed a central bank or lender of last resort to function properly. Another was that the banking system needed a central bank to maintain stability after the gold standard had been suspended after 1914. They reject these hypotheses in favor of one arguing that political rather than economic factors were key. Creating a central bank in response

to the disruption of the Great Depression was viewed as a politically expedient move in light of the urging from other governments and international groups like the World Monetary and Economic Conference to create a central bank to help coordinate international monetary efforts and cooperation.

One of his more recent contributions to Canadian central banking, along with Angela Redish and Hugh Rockoff, examines reasons for why Canada has had few, if any, bank panics when compared to the United States (Bordo, Redish, and Rockoff 2011). A key reason is that the Canadian Federal Government had the privilege of chartering banks, while in the United States both states and the federal government had such authority. This gave rise to the dual banking system in the United States, one characterized by small banks and a unit banking system. This made the United States prone to producing a shadow banking system, one characterized by bank-like intermediaries operating alongside the traditional banks but outside of the regulatory framework. This left the shadow banks vulnerable to runs on deposits, as was revealed in the panics of 1907 and 2008. In contrast, Canada developed a more oligopolistic and heavily regulated banking system that included extensive branching. It was stable enough that a Canadian central bank didn't even arise until 1935, and then only because of political pressure (Bordo and Redish 1987).

He also has provided a useful historical survey of lenders of last resort that includes a series of lessons to be drawn from the historical experience (Bordo 1990). He looks at four views of the lender of last resort, ranging from a classical view of freely lending to stem a panic to a free-banking view that no government agency need provide lender of last resort activities. The classical view is grounded in Bagehot's prescription to lend freely at a high rate but only on good collateral as valued at pre-crisis values. Another view (Goodfriend and King 1988) argues that open market operations, especially under a regime of branch banking, would be adequate to provide liquidity during a crisis, as the lender of last resort would not have to evaluate the conditions of individual banks. A more intense approach to lender of last resort operations is found in Goodhart (1988), in which a lender of last resort should even lend to insolvent banks. This is the case because during a crisis determining illiquidity from insolvency is not really feasible. At the other extreme are the modern proponents of free banking like George Selgin (1988). Free-entry and nation-wide branching would go far in preventing systemic runs on deposits and panics, even without a formal lender of last resort. More advanced clearinghouse associations could be employed to reduce uncertainty about the values of demand deposits held across member banks. After the survey, Bordo concludes that while a public authority should provide lender of last resort functions, it need not be a formal central bank. While free-banking alone, in his view, would not eliminate the problem of bank panics, moving in that direction would certainly work, especially by allowing branch banking.

In a broad survey, Bordo and Siklos (2017) examine the changing roles of ten central banks over the course of more than 350 years, starting with the Riksbank of Sweden, founded in 1668. In this extensive survey, they conclude that in smaller economies, the central bank tends to be more willing to modify its operating procedures and banking policy than banks in larger or more systemically important

economies. While maintaining price stability, and indirectly macroeconomic stability, has emerged as a preferred goal of most central banks, the recent financial crises have forced a reconsideration of maintaining the stability of financial markets. Of course, this was in part the reason the Federal Reserve System was founded in the aftermath of the Panic of 1907.

Finally, in a recent article Bordo presents two fascinating counterfactual worlds in which other lenders of last resort developed in the United States in place of the Federal Reserve System (Bordo 2012). The first counterfactual supposes that the Second Bank of the United States had its charter renewed by Andrew Jackson and that it survived to the present. He argues that it is likely that the bank would have learned to behave like a lender of last resort, as the bank of England learned after the Overend Gurney Crisis of 1866. An especially interesting speculation is that the Bank may have been able to finance the US Civil War more effectively, obviating the need to resort to Greenbacks. The need to establish the National Banking System may have even been eliminated with a well-functioning Second Bank. He also speculates the Second Bank would have been more likely than the Fed to have not followed a tight monetary policy based on the real bills doctrine inherent in the early Federal Reserve.

The second counterfactual assumes the US banking system played out as it had until the need to replace the National banking System with a real lender of last resort had become apparent. But rather than the Federal Reserve System being created, a central bank much closer to those of Europe would have been put in place. His United Reserve Bank, while less centralized than the central banks of Europe, would nevertheless rediscount bills of exchange and other commercial paper, providing a secondary market for these instruments that had been pretty much absent in the United States, unlike Europe. This market, controlled by altering the discount rate when deemed necessary, would provide liquidity for banks and eliminate the volatile call loan market. This would sever the link between the stock market and the payments system, the dangers of which became all too apparent during the Panic of 1907. He believes that a system more along the lines of that proposed by Warburg would have been more active in providing liquidity during the Great Depression. If this had prevented most of the bank failures of the early 1930s, Glass-Steagall with the separation of commercial and investment banking and the creation of the FDIC may never have happened. He concludes that while the Fed appears to have finally learned from past mistakes by the early 2000s, his counterfactual worlds may have avoided much of the disruption caused by early Fed mistakes.

Eugene White (1983) clearly outlined the nature and economic effects of the dual banking system on the US economy. This is a fundamental work on dual-banking He carefully analyzes the effects of the dual banking system having state and federal-chartered banks operating in similar financial and banking markets while having greatly different degrees of regulation imposed on the two classes of intermediaries, a system unique to the United States that also contributed to panics and even the Great Depression. The National Bank Acts and the Federal Reserve Act, as White clearly points out, established new structures on top of the old. Rather than really reform the previous system, the new regulations left old weaknesses in place because

of the entrenched banking interests. The National Banking Acts left the state-banks in place, as did the Federal Reserve Act. When a crisis erupted, the less regulated intermediaries were usually the source of concern, as in 1907. This problem remained under the Federal Reserve System, because state banks were not covered by the Fed's lender of last resort services during the Great Depression. Moen and Tallman (1992) use White's research on the dual banking system to reveal that a subset of state-chartered intermediaries, the trust companies in New York City, were the focus of the Panic of 1907.

Allan Meltzer's (2003) monumental two volume history of the Federal Reserve provides great detail and insight into Federal Reserve policy from its very beginning up through 1986. This well-documented work focusses on the operation of the Federal Reserve System, highlighting its origins in the turmoil of 1907. The first volume covers the origins of the Fed and its operations up through World War II. The second starts with the Fed/Treasury accord and goes up through 1986 and the end of the Paul Volcker era and the start of Alan Greenspan's terms as Chair of the Board of Governors. While Friedman and Schwartz spent a good deal of time on the Federal Reserve System and its operations, their scope is broader than Meltzer's, encompassing an effort to measure the money supply and document changes in the supply on the economy. Meltzer argues, as do Friedman and Schwartz, that maintaining the gold standard and the influence of the real bills doctrine led the early Fed leaders to follow faulty policies, ones that made crises during the Great Depression more severe. Benjamin Strong is also less of a hero in Meltzer's view in comparison to that of Friedman and Schwartz. In the case of the Great Depression, Meltzer argues that Fed officials viewed low nominal interest rates and low member bank borrowings from the Fed as a sign that monetary policy was easy. After an initial loosening after the stock market crash, the Fed returned to a normal to tight policy.

Meltzer is less convinced than Friedman and Schwartz that had Strong lived the Fed would have followed a more accommodative policy. He believes that the Fed was following what he refers to as the "Burgess-Riefler" doctrine, under which banks would borrow at the discount window only in times of emergency. The Fed would use open market operations to encourage or discourage lending. Low interest rates and low bank borrowings were viewed as signs of monetary ease under this doctrine, while we know today that they could also indicate low inflation and reluctance to extend more bank credit. Such a policy was in effect throughout the 1920s in Meltzer's view, while Friedman and Schwartz view the 1930s as a departure from an otherwise successful policy.

Friedman and Schwartz (1963) brought the modern discussion of Federal Reserve policy failure during the Great Depression to the fore, and it is still discussed today. Another key contributor to this debate has been David Wheelock. In his book, *The Strategy and Consistency of Federal Reserve Monetary Policy, 1924–1933* (Wheelock 1991), he examines the evolution of Federal Reserve policy. First, he notes that by 1924, Fed policy makers were aware that the Fed could affect credit conditions in ways that did not have to follow a real bills doctrine approach. Credit did not have to directly follow the course of business and could be eased if business

activity was slow. He credits this in part to Governor of the New York Federal Reserve Bank Benjamin Strong's experience with the Panic of 1907. Unlike Friedman and Schwartz (1963), however, Wheelock does not believe that the untimely death of Governor Strong was the main reason the Fed failed during the Great Depression. First, Strong's influence over the Open Market Committee might not have been as strong during the early stage of the Depression, as there appears to have been some consensus at the Fed, even some from strong himself, that policy may have been too easy during 1924–1927. This would have made significant open market operations during the early stages of the Depression less appealing to Reserve Bank governors. Second, Wheelock argues that during the Depression, as it had in the past, the Fed watched member bank borrowings as an indicator, with low borrowings indicating looser monetary policy. Observing this in the early 1930s led Fed policy makers to conclude that easing credit conditions was not necessary. Furthermore, the decentralized structure of the Fed at this time made it difficult for the banks to achieve a consistent discount-rate policy. Wheelock also notes that defending the gold standard was also a priority for the Fed. In short, the death of Benjamin strong was not as critical a contribution to inept Federal reserve behavior as Friedman and Schwartz argue; the Fed continued the policies it had established in the early 1920s. Nevertheless, he still argues that the Fed acted in ways that likely worsened the Depression.

Wheelock elaborates on some of these arguments in a paper with Michael Bordo (Bordo and Wheelock 2013). In addition to misreading monetary conditions based on observing discount window borrowing and its commitment to maintaining the gold standard by keeping interest rates up, they argue that there was another factor to consider. They argue that the Federal Reserve Act failed to recreate for US banking some of the desirable features of British or German banking. In particular, it did not establish an active and deep secondary market for banker's acceptances, one in which the Fed could intervene and buy large volumes of acceptances to inject liquidity into a contracting banking system. The decentralized structure and divergent opinions of the various Federal Reserve governors on how much monetary ease was legitimate meant that whatever intervention the Fed practiced during the early 1930s would not be adequate to provide liquidity. So, while there may be more to the story behind the inept behavior of the Fed than just the untimely death of Benjamin Strong, the story of misguided banking policy on the part of the Fed appears to remain solid.

The Payments System and Correspondent Banking

The functioning of the payments system has also been subject to cliometric scrutiny. John James and David Weiman (2010) analyze the flow of funds in the United States during the National Banking Era, and they estimate the size and extent of the currency premium that appeared with regularity during these panics. They show how the U.S. payments system during the National banking period developed, with the large New York national banks evolving into a de facto central bank providing

clearing services for the correspondent banking system. They argue that the correspondent system that developed during the later nineteenth century worked quite well in routing check payments, with the call loan market providing services similar to the modern federal funds market. A system of domestic exchange rates had arisen to mediate shipments of currency or exchange across banks and regions of the country, a system similar to that of international exchange having shipping prices similar to the gold points found in international exchange.

While the privately created correspondent system worked well, James and Weiman note that the Federal Reserve rationalized reserve holdings across banks by allowing banks to hold fewer reserves. The centralizing of member bank reserves at the Federal Reserve banks allowed for this. Coordinating clearing standards and prices and introducing the telegraph into clearing and payments further improved the payments system under the new Federal Reserve.

James et al. (2014) extend this analysis and show that clearing house loan certificates provided liquidity locally but segregated the country into local currency areas during panics, either because they could not circulate widely or, in the case of New York, could circulate only among Clearing House member banks. To a limited extent the Treasury could rearrange its deposits to expedite funds flows that could occur with loan certificates. But its resources were limited by the fact that the Treasury could not produce a net increase of new reserves.

Anderson et al. (2018) examine how joining the Federal Reserve System affected liquidity and lending at member banks. They show that even by the mid-1920s, less than a third of state banks and trust companies had joined the Fed. They show that state banks that faced seasonality in loan demand were more likely to join, reducing the variation in lending. Larger state banks higher up in the correspondent system were more likely to join as a way to gain access to the discount window, improving their ability to provide liquidity to smaller banks. Lending increased at these banks. Smaller state banks that were part of the correspondent system were less likely to join, as they still had indirect access to the discount window while avoiding the higher reserve requirements imposed by Fed membership. Incomplete coverage of lender of last resort services introduced an element of moral hazard and shadow banking.

Mark Carlson has published a number of papers on the interbank payments system and the effects of the Federal Reserve on the stability of the payments system. In a particularly insightful paper (Carlson et al. 2011) he shows that the infusions of cash by the Atlanta Federal Reserve Bank to banks in Florida suffering runs on deposits in the summer of 1929 were effective in preventing further runs on deposits. Because Reserve banks operated much more independently then, banks in districts with little previous experience with panics were less likely to respond during the Depression in a fashion similar to that demonstrated by the Atlanta bank in 1929. The Banking Acts of 1932 and 1933 authorized the reserve banks to lend on any collateral deemed valuable by the bank's governor, and the example of Atlanta may have been part of the impetus for these Acts.

Carlson and Wheelock (2016) show that the early Fed had helped to reduce seasonality in the flow of funds between banks. The Fed's reserve requirements could no longer be satisfied by correspondent deposits, reducing contagion risk in

the early years of the Fed. The decentralized nature of decision making and the large number of state banks outside of the Federal Reserve System left the banking system vulnerable to the massive liquidity shock of the early years of the Great Depression. This resulted in the significant reorganization and centralization of the Fed in 1935. Bordo and Wheelock (2013) elaborate on the reorganization of the Fed. The original Federal Reserve Act and the reforms of the 1930s, in their view, failed to promote a deep secondary market for banker's acceptances, a market that existed in Europe, contributing to the effectiveness of those central banks. Maintaining the dual banking system further hindered the effectiveness of the Federal Reserve's lender of last resort capabilities.

Finally, Stephen Quinn and William Roberds (2006) have pushed back the timeline for central banking and present the case that the Bank of Amsterdam, founded in 1609, became the first true central bank. In the early seventeenth century the Dutch Republic faced a coinage debasement problem owing to a wide variety of foreign coin circulating in the young republic. One regulatory approach to reduce debasement was the creation of an exchange bank in Amsterdam, the Bank of Amsterdam, in which commercial debts in the form of bills of exchange had to be settled through the bank with bank money, shielding creditors from payment with debased coins. Eventually, the bank's money was decoupled from any coinage, becoming one of the first "fiat" currencies. Quinn and Roberds describe this development as resembling modern day open market operations, in this instance the sale or purchase by the bank of receipts against deposits of gold or silver coin.

Clio and Banking Databases

Cliometricians have also added new and rich data sets to the analysis of banking in the United States. While many cliometricians use data sets in their research, several large data sets stand out. Caroline Fohlin has collected a substantial set of daily observations on the call loan rate, the rate charged on overnight loans made to stockbrokers on the New York Stock Exchange. While the call loan market was most notably a national banking institution, it survived well into the Federal Reserve period. Instability transmitted from the call loan market into the banking sector in 1907 gave rise to political support from the New York bankers for a lender of last resort independent of the Clearing House. Combined with the call loan data, Fohlin et al. (2016) use an amazing data set covering transactions prices, closing bid-ask spreads, and trading volumes for all stocks traded on the New York Stock Exchange for every trading day between 1905 and 1910. This data set is a subset of a larger, NSF funded set for every trading day between 1900 and 1925 (Fohlin 2015). They examine how market and funding liquidity changed over the course of the Panic of 1907, and they show that the more opaque the market for a particular stock was, the greater the wedge between bid and ask prices. Mining stocks were the most opaque, and the source of much turmoil in the early stages of the panic. Railroad stocks were much more transparent. In some cases bid-ask spreads remained elevated for months after the panic had subsided. Cash infusions to keep the stock market and its call loan

market functioning and liquid did not have lasting effects on trading volumes. They conclude that regulatory steps put in motion by the panic eventually led to the demise of unregulated securities markets, introducing more transparency to financial and banking markets.

With the help of David Weiman, Jon Moen and Ellis Tallman have collected information from the minutes of the New York Clearing House Committee, the committee summoned to oversee the issue and withdrawal of clearing house loan certificates during panics. This set contains the volume of loan certificates issued by each Clearing House member bank and when they were withdrawn from circulation, on a daily basis for the major panics of 1873, 1884, 1890, 1893, and 1907. They also link these daily observations to bank balance sheet information collected by the Comptroller of the Currency for the call date nearest the beginning of the panic. They are currently examining this data.

Matt Jaremski has collected several large sets of balance sheet data for national and state banks. They cover both the National Banking Era and the early Federal Reserve period. One contains all of the balance sheet information collected by the Office of the Comptroller of the Currency between 1865 and 1934. A related set contains similar evidence for state banks from various state banking departments, although its coverage is not comprehensive like that for the national banks. He has also identified the correspondents of national banks, state banks, and trust companies for 1890, 1893, 1897, 1900, 1910, 1919, and 1940 recorded in various bankers' directories. He has also identified the location of US clearinghouse for 1852 through 1920 and for 1929 and 1940. This information is also taken from various bankers' directories. He has produced a number of papers from these data sets, including Jaremski and Rousseau (2013), Jaremski (2014), and Jaremski and Wheelock (2017). One important result revealed by his research has been to highlight the importance of bank networks in reducing susceptibility to bank runs in the period before the Federal Reserve. His research also emphasizes the importance of the development of the banking system to the economic development of the real sector of the United States in the nineteenth century.

Conclusions

Cliometrics has contributed to a better understanding of central banking, especially in the United States. While many "non-cliometricians" now and in the past have contributed to the understanding of central banking, cliometric analysis has contributed most clearly along three lines. First, the collection of new sources of evidence, that of Matt Jaremski in particular, is a highlight of much of this work. Recent research tends to be much more disaggregated and even granular when compared to earlier research on the payments system. While traditional aggregated monetary and financial measures are still useful, the increasing availability and exploitation of individual bank data, sometimes even down to the depositor level, will open new windows into the study of central banking and the payments system. The work of Caroline Fohlin highlights this approach.

Second, the structures and institutions of central banking in the United States are now well understood. In particular, the role of clearinghouses in the establishment of the Federal Reserve System has been greatly advanced by cliometric research. In particular, the work of Moen and Tallman shows that while clearinghouses helped mitigate runs on deposits, their effectiveness was limited given that intermediaries outside of the clearinghouses, mainly state-chartered intermediaries, could not be aided by the tools available to clearinghouses before the advent of the Federal Reserve. The limited resources of the New York Clearing House in 1907 finally made the need for an outside source of liquidity apparent to the politically influential New York bankers. Along these lines, research into the workings of the National Banking System, especially that related to the dual banking system and the rise of correspondent banking and the call loan market, have provided better insight into the structure and early operations of the Federal Reserve System. The effects of incomplete coverage of lender of last resort functions, both before and after the Federal Reserve, have been shown to be a characteristic of banking in the United States, one that distinguishes it from banking elsewhere. The research of Eugene White has been a central contribution to this perspective. This line of research also helps explain the decentralized structure of the early Federal Reserve System, a structure that was modelled on the clearinghouses familiar to US bankers at the time (Timberlake 1993).

Finally, there is consensus among cliometricians that the Fed could have done a much a much better job during the Great Depression. Friedman and Schwartz clearly conclude that Fed leaders were unable to interpret basic monetary measures after the death of Benjamin Strong, producing contractionary monetary policy during the early stages of the Depression. Others agree that the Fed did not respond appropriately, but the failure of policy resulted rather from following policies that had worked well before 1930, including supporting the interwar gold standard. Incompetence is not the leading explanation in this thread. The extended work of David Wheelock presents this interpretation most clearly.

A question that remains to be fully answered, at least in the context of the United States, is just how formally organized does a lender of last resort have to be in order to provide central bank services, especially lender of last resort services. Richard Timberlake would argue, “not very.” Improved clearinghouses would have been adequate. Michael Bordo argues for more formality, but a centralized European bank was likely not necessary. There was and is room for private sector responses. Moen and Tallman agree on this point. Coming from a European viewpoint, Charles Goodhart would argue for the most centralized bank. Nevertheless, there remains much opportunity for cliometricians to contribute to this last question, especially along banking policy lines.

Cross-References

► Cliometrics and Antebellum Banking
► Financial Markets and Cliometrics
► Origins of the U.S. Financial System

► The Cliometric Study of Financial Panics and Crashes
► The Great Depression in the United States

References

Anderson H, Calomiris C, Jaremski M, Richardson G (2018) Liquidity risk, bank networks, and the value of joining the Federal Reserve System. J Money, Credit, Bank 50(1):173–201
Andrew AP (1908) Substitutes for cash in the panic of 1907. Q J Econ 23:497–516
Andrew AP (1910) Statistics for the United States. Publications of the National Monetary Commission, Washington, DC, p 21
Antipa P (2014) Fiscal sustainability and the value of money: lessons from the British paper pound, 1797–1821. Working paper, Bank of France
Bagehot W (1873) Lombard street. King, London
Bignon V (2017) The failure of a clearinghouse: empirical evidence. Bank of France working paper no 638
Blaug M (1996) Economic theory in retrospect, 5th edn. Cambridge University Press, Cambridge
Bodenhorn H (2002) Making the little guy pay: payments-system networks, cross-subsidization, and the collapse of the Suffolk system. J Econ Hist 62(1):147–169
Bordo M (1989) The contribution of "a monetary history of the United States, 1867–1960" to monetary history. In: Bordo M (ed) Money, history, and international finance: essays in honor of Anna Schwartz. University of Chicago, Chicago
Bordo M (1990) The lender of last resort: alternative views of historical experience. Federal Reserve Bank of Richmond Review, Richmond. pp 18–29
Bordo M (2012) Could the United states have had a better central bank? An historical counterfactual speculation. J Macroecon 34:597–607
Bordo M, Redish A (1987) Why did the bank of Canada emerge in 1935? J Econ Hist 47:405–417
Bordo M, Siklos P (2017) Central banks: evolution and innovation in historical perspective. NBER working paper 23847
Bordo M, Wheelock D (2013) The Promise and performance of the federal reserve as lender of last resort 1914–1933. In: Bordo MD, Roberds W (eds) The origins, history, and future of the federal reserve: a return to Jekyll Island. Cambridge University Press, Cambridge, pp 59–98
Bordo M, Redish A, Rockoff H (2011) Why didn't Canada have a banking crisis in 2008 (or in 1930 or 1907, or . . .)? NBER working paper #17312
Broz JL (1997) The international origins of the federal reserve system. Ithaca. Cornell University Press, New York
Calomiris C, Kahn C (1996) The efficiency of self-regulated payments systems: learning from the Suffolk system. J Money, Credit, Bank 28(4):766–797
Calomiris C, Mason J (2008) Resolving the puzzle of low national bank note issuance. Explor Econ Hist 45:327–355
Cannon J (1910) Clearing houses, vol 6. Publications of the National Monetary Commission, Washington, DC
Carlson M, Wheelock D (2016) Interbank markets and banking crises: new evidence on the establishment and impact of the federal reserve. Am Econ Rev Pap Proc 106(5):533–537
Carlson M, Mitchener C, Richardson G (2011) Arresting banking panics: federal reserve liquidity provision and the forgotten panic of 1929. J Polit Econ 119(5):889–924
Currie L (1935) The supply and control of money in the United States. Harvard University Press, Cambridge
Fohlin C (2015) A new database of transactions and quotes in the NYSE, 1900–1925 with linkage to CRSP. In: Johns Hopkins Mimeo. working paper, Johns Hpokins, Baltimore
Fohlin C, Gehrig T, Haas M (2016) Rumors and runs in opaque markets: evidence from the panic of 1907, CEPR working paper DP10497

Friedman M, Schwartz A (1963) A monetary history of the united states. University Press, Princeton, pp 1867–1960
Friedman M, Schwartz A (1970) Monetary statistics of the United States: estimates, sources, methods. National Bureau of Economic Research, New York
Friedman M, Schwartz A (1984) Monetary trends in the United States and the United Kingdom. NBER, Cambridge MA
Friedman M, Schwartz A (2008) The great contraction, 1929–1933. Princeton University, Princeton
Goodfriend M, King R (1988) Financial deregulation, monetary policy, and central banking. In: Haraf W, Kushmeider R (eds) Restructuring banking and financial services in America. AEI, London
Goodhart C (1988) The evolution of central banks. MIT Press, Cambridge
Gorton G (1985) Clearinghouses and the origins of central banking in the United States. J Econ Hist 45:277–284
Gorton G (2010) Slapped by the invisible hand. Oxford University Press, Oxford
Gorton G, Mullineaux D (1987) The joint production of confidence: endogenous regulation and nineteenth century commercial-bank clearinghouses. J Money, Credit, Bank 19:457–468
Hendrickson J (2018) The bullionist controversy: theory and new evidence. J Money, Credit, Bank 50(1):203–241
Hetzel R (1987) Henry Thornton: seminal monetary theorist and father of the modern central bank. Fed Reserve Bank Ric Econ Rev 73:3–16
Hoag C (2012) Clearinghouse loan certificates during the panic of 189. Academy of. Bank Stud J 11(2):93–105
Hume D (1985) Of money, in Essays: Mortal, Political, and Literary. Liberty Fund, Indianapolis
James J, Weiman D (2010) From drafts to checks: the evolution of the correspondent banking networks and the formation of the modern U.S. payment system. J Money, Credit, Bank 42(2/3):237–265
James J, McAndrews J, Weiman D (2014) Banking panics, the "derangement of the domestic exchanges, and the origins of central banking in the United States, 1893 to 1914. Working paper
Jaremski M (2014) National banking's role in U.S. industrialization, 1850–1900. J Econ Hist 74:109–140
Jaremski M, Rousseau P (2013) Banks, free banks, and economic growth. Econ Inq 51:1603–1621
Jaremski M, Wheelock D (2017) Banker preferences, interbank connections, and the enduring structure of the federal reserve system. Explor Econ Hist 66:21–43
Knodell J (2003) Profit and duty in the exchange operations of the second bank of the United States. Financ Hist Rev 10(1):5–30
Livingston J (1986) Origins of the federal reserve system: money, class, and corporate capitalism. Cornell University Press, Ithaca, pp 1890–1913
McCulley R (1992) Banks and politics during the progressive era: the origins of the federal reserve system. Garland Publishing, New York/London, pp 1897–1913
Mehrling P (2016) Economists and the fed: beginnings. J Econ Perspect 16(4):207–218
Meltzer A (2003) A history of the federal reserve, vol 1 and 2. University of Chicago Press, Chicago
Moen J, Tallman E (1992) The bank panic of 1907: the role of the trust companies. J Econ Hist 52:611–630
Moen J, Tallman E (2000) Clearinghouse membership and deposit contraction during the panic of 1907. J Econ Hist 60:145–163
Moen J, Tallman E (2007) Why didn't the United States establish a central bank until after the panic of 1907? Federal Reserve Bank of Atlanta. Working paper
Moen J, Tallman E (2015) Close but not a central bank. In: Humpage O (ed) Current federal reserve policy under the lens of economic history. Cambridge University Press, New York
Quinn S, Roberds W (2006) An economic explanation of the early bank of Amsterdam, debasement, bills of exchange, and the emergence of a central bank. Federal Reserve Bank of Atlanta working paper 2006-13
Ricardo D. Plan for the establishment of a national bank; 1824.

Rolnick A, Smith B, Weber W (1998) Lessons from a laissez-faire payments system: the Suffolk banking system (1825–58). Federal Reserve Bank of St. Louis Review 80:105–16
Schwartz A (1986) Real and pseudo financial crises. In: Capie F, Woods GE (eds) Financial crises in the world banking system. Macmillan, London
Selgin G (1988) The theory of free banking: money supply under competitive note issue. Rowman and Littlefield, Washington, DC
Sprague OMW (1910) History of crises under the national banking system. National Monetary Commission. U.S. Government Printing Office, Washington, DC
Tallman E, Jacobson M (2015) Liquidity provision during the crisis of 1914: public and private sources. J Financ Stab 17:22–34
Taus E (1943) Central banking functions of the U.S. treasury. Columbia University Press, New York, pp 1789–1941
Thornton H (1802) An enquiry into the nature and effects of the paper credit of Great Britain. Hatchard, London
Timberlake R (1978) The origins of central banking in the United States. Harvard University Press, Cambridge
Timberlake R (1984) The central banking role of clearinghouse associations. J Money, Credit, Bank 16:1–15
Timberlake R (1993) Monetary policy in the United States: an intellectual and institutional history. The University of Chicago Press, Chicago
Tooke T (1844) An inquiry into the currency principle. Longman, Brown, Green, and Longmans, London
Warburg P (1930) The federal reserve system: its origin and growth, vol 1 and 2. The Macmillan Company, New York
West R (1974) Banking reform and the federal reserve. Cornell University Press, Ithaca, pp 1863–1923
Wheelock D (1991) The strategy and consistency of federal reserve monetary policy, 1924–1933. Cambridge University Press, Cambridge
White E (1983) The regulation and reform of the American banking system: 1900–1929. Princeton University Press, Princeton
Wicker E (1996) The banking panics of the great depression. Cambridge University Press, Cambridge
Wicker E (2000) Banking panics of the gilded age. Cambridge University Press, Cambridge
Wicker E (2005) The great debate on banking reform: Nelson Aldrich and the origins of the fed. Ohio State University Press, Columbus
Wicksell K (1934) Lectures on political economy, money, vol 2. Routledge, London

Sovereign Debt

Mauricio Drelichman

Contents

Abstract

Sovereign debt is one of the most enduring puzzles in economics. Why would anyone lend to an entity not subject to external enforcement? This chapter examines the emergence of sovereign loans, their transformation from personal to transpersonal debt, and their evolution through time. The different models that have been proposed to explain the rationale and sustainability of sovereign

M. Drelichman (✉)
The University of British Columbia, Vancouver, Canada
e-mail: mauricio.drelichman@ubc.ca

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_37

lending are discussed and evaluated in light of the historical record, together with a brief discussion of data sources and the political economy of debt.

Keywords
Sovereign debt · Sustainability · Reputation · Sanctions · Default · Bonds · Loans

Introduction

Debt and time are inextricably linked. The very existence of debt requires at least two transactions – the disbursement of a loan and its subsequent repayment – at separate points in time. The price of debt, which is simply the interest paid on a loan, is determined by the time elapsed between these basic transactions. Government debt instruments usually mature over several decades, and lending relationships between states and financial institutions can last centuries. It is no surprise, then, that the study of debt relationships is central to cliometrics, a discipline fundamentally defined by the study of economic magnitudes, problems, and behaviors through time.

The element of time intrinsic in any debt contract introduces an implicit, yet unavoidable element of risk. Once a loan has been disbursed, it is always possible that the debtor will fail to repay it, either because of insolvency or because they judge that breaking the contract will have a lower cost than complying with it. Default risk, the probability that a loan is not repaid in accordance with its original terms and conditions, is a key determinant of the structure of debt instruments, the pricing of loans, and, indeed, the very existence of debt markets. If default risk is not addressed as part of the institutional structure of a debt market – that is, if any debtor can decide not to repay a loan with impunity – it is quite possible that no loans will be made in the first place. Without transactions, there are no markets.

Among the myriad institutional arrangements and legal instruments through which different societies, polities, and individuals have conceived and implemented debt through history, sovereign debt emerges as a special category, exercising an undying appeal for theorists, applied economists, and cliometricians alike. The reason is that there is very little to mitigate default risk when the debtor is a sovereign who holds both legislative powers and a comparative advantage in the exercise of violence. In the case of a loan to an absolutist ruler, to use an obvious example, it is usually impossible to compel a reluctant borrower to repay a loan. But even when the borrower is a democratic government subject to a large array of internal checks and balances, foreign creditors have found that legal frameworks can be changed with surprising speed, often to their detriment.

The study of sovereign debt, both as a theoretical subject and in its historical expressions, almost always converges to a well-defined set of lines of inquiry. Why is sovereign debt possible in the first place? Just how much debt can a sovereign carry? What is the nature of defaults, and what are their implications for the viability of sovereign debt? The remainder of the analysis expands on these questions and examines the interplay of theory and history in the scholarly efforts to address them.

The Big Questions

What Makes Sovereign Debt Possible?

Debt agreements subscribed between private parties usually fall under the legal jurisdiction of a state entity. In case of nonperformance, the creditor can, at least notionally, seek redress in court. Enforcement may pose complications, especially when the parties reside in different jurisdictions, but these are no different than the challenges in enforcing any other type of contract. Some debt arrangements, such as those extended by criminal organizations, fall outside the legal protections of the state, but in these cases, the creditor is typically the stronger party and usually has access to a coercion or punishment technology. The Mafia does not take out loans, and it is very effective at collecting the ones it extends. Sovereign debt is unique in that creditors are typically the weaker party, while debtors control the means of coercion.

Loans to sovereigns have not always been voluntary arrangements. Medieval city-states in northern Italy used forced loans on wealthy citizens, called *prestanze prestitix*, as their primary means of financing (Pezzolo 2003). Charles I of England also levied forced loans and imprisoned reluctant noblemen as part of his escalating conflict with Parliament (Cust 1985). Forced debt was also a preferred method of financing by German occupation authorities in Greece during World War II (Christodoulakis 2014). To the extent that these forced loans can be considered debt at all, rather than the outright confiscation of resources they often resulted in, it is not hard to see what makes them possible. Debtor participation is coerced under threat of violence.

The interesting case, therefore, is the one in which a lender freely enters into a debt contract with a sovereign entity. Because the sovereign can choose not to repay the debt with apparent impunity, the question of what makes sovereign debt possible might well be reformulated as "Why do sovereigns ever repay their debts?" or, extending the logical chain one step further, "Why would a creditor ever extend a loan to a sovereign?" These questions stand at the very core of the sovereign debt literature in economic theory and history. The first mainstream answer has emphasized reputation-based lending relationships, while an alternative explanation has focused on the possibility of creditors imposing costly sanctions on borrowers. Both these schools of thought are explored in detail further below.

How Much Debt Can States Carry?

Sovereign states carry a wide range of debt loads, using different instruments and committing themselves to varying repayment schedules. Some borrow very little and only for short periods. Others take on fantastically high obligations and may carry them for very long periods, even intending never to reduce their overall indebtedness. What explains these differences, and what are their determinants?

These questions relate to what economists call the "sustainability" of debt. The term is used in a somewhat ambiguous fashion throughout the literature, as it may refer to either the fiscal ability of a sovereign to service and repay their debts or to the

incentives and willingness to do so, regardless of fiscal capacity. The section on fiscal sustainability explores the first aspect, while the discussion of each model of repayment incentives elaborates on the latter.

The Nature of Defaults

Anyone extending a loan for a profit motive clearly expects to be repaid. Lenders facing a higher risk of default demand additional compensation in the form of higher interest or other guarantees. If the risk is too high, or the potential losses too large, lenders will choose not to lend at all. In such cases, a market for sovereign debt might unravel, or not emerge in the first place.

The traditional approaches to the theory of sovereign debt usually feature rational agents in a context of complete information. In these models, borrowers want to signal their willingness and ability to repay; lenders only lend to those borrowers that are solvent and credible enough. Incentives are aligned, debt is properly priced, and loans are repaid. In such an environment, however, there is no place for default. Even in contexts with uncertainty generated by external events, insurance markets can be used to spread the risk. Default is, in economic parlance, an off-the-equilibrium-path outcome.

The problem is, of course, that sovereign default is common – in fact, far too common to be explained away by informational mistakes. The history of sovereign debt is heavily punctuated by episodes of kings reneging on their obligations, developing countries declaring bankruptcy, and revolutions disowning the financial commitments of the preceding regimes. Nor are creditors always eager to punish states that default. Castile under the Habsburg monarchs decreed general payment stops seven times in less than a century, before finally falling out of favor with international financiers. Argentina and Venezuela boast comparable records in the nineteenth and twentieth centuries, with others not far behind. For the past five centuries, up to 10% of borrowers ended up suspending payments in any given period, with default rates reaching 20% in particularly turbulent times (Reinhart and Rogoff 2009). The notion of off-the-equilibrium-path phenomena appears insufficient to explain the general ubiquity of sovereign default and the surprising resilience of "serial defaulters."

Relying on advances in contract theory and on the idea of incomplete markets, the literature has begun to interpret defaults as part and parcel of sovereign debt transactions. In a new generation of models that reflect some of the myriad complex relationships in sovereign debt markets, defaults are more fully integrated into the potential equilibrium outcomes. Defaults that are triggered by unexpected, exogenous events through no fault of the borrower can be deemed "excusable" by lenders, and not held against a borrower's record (Grossman and Van Huyck 1988). In situations where lenders have market power, defaults can be a stable feature of the market equilibrium (Kovrijnykh and Szentes 2007). And, when markets are incomplete, the impossibility of creating provisions for every possible eventuality means that lenders and borrowers both operate in the knowledge that some states of the world will bring about defaults, and price their loans accordingly (Arellano 2008).

Why Does Sovereign Debt Exist at All?

Sovereign debt repayment may be subject to the whim of a ruler, plagued with the possibility of defaults, or susceptible to political upheaval and macroeconomic instability. Why would lenders put up with the uncertainties inherent in such a market and the vagaries of whimsical kings and turncoat politicians? The answer has been the same throughout history – because "there's gold in them thar hills." While some investors certainly suffered losses in individual crises, returns to sovereign debt have on average exceeded those of comparable safer investments over the long run, even after accounting for the effect of defaults. Lindert and Morton (1989), for example, found that between 1850 and 1983, government loans generated an average excess return of 0.4% over similar American or British bonds, considered to be reference safe assets. This also proves true going further back in history; even that most iconic of serial defaulters, King Philip II of Spain, generated profits for his lenders. Drelichman and Voth (2011b) showed how bankers that lent through the last two Castilian defaults of the sixteenth century earned an excess return of approximately 2% over safer bonds. A market for sovereign debt exists because it delivers a premium that risk-neutral lenders find attractive.

On the sovereign's side, the motivation for borrowing is no different from that of any other borrower – bridging a gap between expenditures and revenues. In a context of economic growth, debt can be a powerful tool to finance public investment. As long as the social rate of return on the projects financed with debt exceeds the interest rate paid by the government, state borrowing can increase overall welfare and contribute to a sustained development path.

Of course, not all loans are motivated by the desire to finance growth-oriented projects, and in any case sovereign debt predates the onset of modern economic growth by several centuries. Debt is often used to finance current expenditure rather than investment, and, both in current and historical times, the funding of military enterprises has been a prime motivator. In cases like these, borrowing is essentially an intertemporal transfer of resources, with future taxes being committed to pay for current expenditures. It is possible that, if a war financed by debt is won, the spoils might provide the means for repayment. These instances, however, are not common; the imperial expansion of eighteenth- and nineteenth-century Britain constitutes perhaps a rare exception. A sovereign's motive for borrowing need not change a lender's disposition, as long as future funds are available. However, if borrowing is motivated by short-term political vanity or survival, and is unlikely to result in a growth payoff, access to debt markets might well result in negative welfare consequences for the country as a whole.

Sovereign Debt in History

The study of state financing has always been one of the top areas of interest in economic history research, perhaps second only to that of economic growth. Its development as a scholarly field started long before the cliometric revolution,

yielding pioneering works like those of P. G. M. Dickson (1967) for England, that of Ramón Carande for Castilian short-term debt (Carande 1987), and the collected works of Giuseppe Felloni for Genoa (Felloni 1999), among many other prominent examples. This literary tradition, which continues to the present day, is typically focused on uncovering the institutional features of sovereign lending through history and, whenever possible, on quantifying its magnitude. This invariably requires painstaking archival work, parsing centuries-old documents to carefully reconstruct the minute details of each transaction. The resulting works are an indispensable foundation for any cliometric work, providing invaluable context to frame and interpret quantitative models and results.

Instruments and Innovations: A Very Short Summary

One of the main concerns of the historical literature is documenting the evolution of contractual and institutional instruments used throughout history – the "technology" of debt, in some sense. Of note is the fact that public debt largely did not exist before the twelfth century; not even the Roman Empire, with its sophisticated tax and coinage systems, had anything resembling public debt instruments. Europe would have to wait until the onset of the Commercial Revolution and the emergence of the merchant city-state to witness the first manifestations of what would become government obligations.

Since antiquity, rulers had issued token coinage, which can be thought of as a form of debt. In a full-bodied currency system, replacing a gold coin with a convertible token requires a degree of trust in the ruler if the token is voluntarily accepted or coercion if not. Devaluations, restampings, and debasements can all be thought of as a form of default. This conception of currency as a sovereign obligation lasted well into the twentieth century. After World War I, the British government used the notion of currency as debt to justify its deflationary impetus, eventually returning the pound to its prewar gold parity rather than devaluing (Eichengreen 1996).

The consolidation of city-states first and territorial states later made it possible for rulers to count on reliable tax revenues that could be used as security for debt repayment. The first sovereign loans were not very different from tax farming. A group of merchants would advance a sum to the state in exchange for the right to collect a specific tax for a number of years. This method was introduced by the Genoese *compere*, which are documented as early as the first half of the twelfth century and would later be widely adopted throughout Western Europe, eventually becoming the favored system of issuing long-term debt (Felloni 2006). Castilian *juros*, French municipal *rentes*, Dutch *renten*, and the Florentine loans to Edward III all used specific revenue streams as the source of repayments for cash advances to their governments. In some polities, such as Genoa, the management of public debt would become part and parcel of the city's mercantile activities. The Casa di San Giorgio, the company of Genoese merchants that lent funds to the Republic and administered its debt, introduced many of the innovations that would make modern

state financing possible (Felloni 2006). Debt secured by specific tax streams is commonly referred to as "funded debt."

David Stasavage (2011) has convincingly argued that, during this formative period, the ability of polities to issue and service their debts depended crucially on the presence of representative assemblies that gave strong voices to merchants, as well as on the ease of travel and communications between different merchant centers within the polity. This gave an initial advantage to city-states and small territorial units like the Netherlands over larger territorial states like Castile and France. The advent of the fifteenth-century Military Revolution, which saw the introduction of standing armies and complex fortifications, tilted the scales in favor of territorial states with the ability to muster the enormous resources required to finance them. Thus Castile, then the Netherlands, and finally France and England all accumulated debts whose size, relative to GDP, rivaled those of contemporary states (the section on fiscal sustainability below provides a quantitative discussion). The period also saw the rise of the legal construct of "state" debt in the context of territorial states. While in the Late Middle Ages, all loans were personally underwritten by the king and guaranteed by his revenues and demesne; by the sixteenth century, debt had become clearly transpersonal; it was an obligation of the Crown, rather than of the person wearing it.

The pledging of specific tax revenues, which underpinned almost all medieval, as well as most early modern debt issues, is quite removed from modern conceptions of debt. It elegantly solved the problem of the credibility of a ruler by transferring the exploitation of a revenue stream to a lender for a specified period or in perpetuity. The debt was treated like property and, being subject to the applicable jurisprudence, provided a certain level of security against repudiation. The format of pledging revenues also allowed for the circumvention of usury laws, as the transaction was considered a sale or a rent contract, rather than a loan. One step closer to modern instruments was taken with the appearance of unsecured debt instruments, which hinged only on a ruler maintaining their word. By far the largest unsecured loans by volume in the early modern era were Castilian *asientos*, debt contracts between the Habsburg kings and bankers from all over Europe. *Asiento* lending started in earnest under Charles V, who borrowed from the Fugger and Welser families. Fueled by the silver remittances from South American mines, it reached its maximum expression under Philip II, for whom Genoese bankers, building upon centuries of financing their own Republic, designed a system that had most of the characteristics of modern sovereign debt (Drelichman and Voth 2014a). As innovative as Castilian debt was, neither the *asientos* nor the tax-backed securities used by most rulers and municipalities were tradable in secondary markets (though *juros*, the Castilian version of tax-backed securities, could be reissued in a different name upon the payment of a fee). The innovation of tradable securities would be introduced by the Dutch Republic in the following century. The creation of the Wisselbank, arguably the first fully functional central bank, completed the Dutch financial revolution and established the first truly modern state financing system (Tracy 1985).

By the eighteenth century, the locus of innovation in the arena of public debt was shifting from Amsterdam to London. The reforms of the Glorious Revolution

(discussed below) enabled a large expansion of government borrowing, while military successes in Britain's many wars and the profits generated by imperial trade and industrial development provided the resources needed to keep public finances in good health. The early eighteenth century saw some experimentation with schemes using the stock of chartered companies as a vehicle for refinancing government debt. Variations of this idea were behind both the South Sea Bubble in Britain and the Mississippi Bubble in France (Carlos and Neal 2006; Velde 2007). Eventually, virtually all British debt would be issued and held in the form of consols, perpetual annuities that could be redeemed at face value. This allowed the government to take advantage of falling interest rates by offering coupon rate reductions. Consols would be the dominant form of debt in Britain until World War I, when the enormous financing needs imposed by the war required the issuance of shorter-term debt (Ellison and Scott 2017). The last British consols were redeemed in 2014. Nowadays, states the world over typically have a debt structure of bonds of different maturities, with those above 30 years being relatively uncommon.

Currency of Issue

One key choice for any government issuing debt in modern times is whether to do so in domestic or in foreign currency. Domestically marketed debt is almost always issued in a country's own currency. As for debt contracted with foreign lenders, the trade-off is clear: debt denominated in domestic currency can lose value as a result of inflation, while debt denominated in foreign currency leaves the borrowing government exposed to exchange rate risk. Governments with little credibility or with a small domestic market tend to prefer foreign currency-denominated debt issues as a way of attracting investors that would otherwise not be willing to take on the inflationary risk, while developed economies on a solid fiscal footing can often afford to issue debt denominated in their own currency.

Before the twentieth century, cross-border debt was overwhelmingly denominated either in gold or silver, or in a currency subject to a metallic standard, thus leaving little leeway for rulers to reduce the value of their debt by tinkering with the currency. Scholars have argued that for countries that were not major financial powers – the so-called periphery – issuing debt in foreign currency was a commitment device. Under the Classical Gold Standard, perhaps the most studied international monetary system in history, peripheral countries tied the value of their currencies to gold as an instrument to increase their creditworthiness – a "good housekeeping seal of approval" (Bordo and Rockoff 1996).

Data Sources

Some of the most important contributions of cliometricians to the sovereign debt literature are data collection efforts that cover several countries over long periods, thus enabling the panel analyses that are key for validating modern theoretical

frameworks. While a thorough enumeration of such compilations is beyond the scope of this chapter, there are nonetheless some noteworthy entries. For medieval and early modern times, Stasavage's (2011) dataset offers a window into the emergence of sovereign debt across 31 European countries and city-states, together with political and demographic correlates. Likewise, Bonney's (2007) European State Finance Database, while not as homogeneous, has become a repository for a wide variety of fiscal and financial data from European countries. Neal (1991) provides series of British and Dutch joint-stock company shares and government debt prices between 1710 and 1820. Homer and Sylla (2005), a work that has grown in scope and depth through its four editions in over four decades, offers the longest-term view of the return on all sorts of assets, including government obligations, from Babylonian times to the early 2000s. More recently, databases have prioritized comprehensiveness, attempting to capture the majority of known debt issues together with a large number of covariates suitable for the increasingly demanding quantitative analysis associated with newer-generation models. In this vein, one can find the contributions of Reinhart and Rogoff (2009) and Abbas et al. (2010), among others.

The Interplay of Theory and History

The main comparative advantage of cliometrics is the harnessing of quantitative data, in combination with detailed institutional knowledge, to evaluate the predictions of economic theory. Conversely, by using the deep repository of history to uncover situations and data types that may not be available in contemporary experience, cliometrics also contributes to the development and refinement of economic theory. This section explores how such a dialogue has emerged in the context of sovereign debt, reviewing the different models proposed to explain its main stylized facts, the inconsistencies pointed by historical studies, and the revisions introduced in theoretical frameworks as a result.

Fiscal Sustainability

One of the first empirical tests to which a sovereign borrower is typically subjected – both in scholarly scrutiny and in actual practice – is whether it can possibly repay the amount of debt it has taken or intends to take on. If a borrower does not have the means to repay a debt, there is no point in going forward with the transaction. For the purposes of this first evaluation, the borrower is assumed to be willing to repay and to employ their every last penny in doing so if necessary. The ability to repay a debt in the context of sovereign lending receives the name "fiscal sustainability." It should not be confused with the willingness or strategic convenience of a borrower to repay, for which the term "sustainability" (without the "fiscal" modifier) is also sometimes loosely employed.

The resources to repay government debt must ultimately come from taxation. At this point, it is important to distinguish between debt denominated in local currency and loans issued in foreign currency and some other unit of account over whose value the government has no power (e.g., precious metal weight). In the former case, the government might reduce the value of the debt through inflation. While inflation is typically interpreted by economists as a tax, it does have a particularly large incidence on holders of debt denominated in domestic currency, who see it erode the value of their future coupon and capital repayments. Lenders do, of course, factor in their inflationary expectations when pricing domestic debt, but a government can always print more money *ex post*. While unanticipated inflation may be regarded as a partial default on domestic debt, it can also help avoid an outright suspension of payments. A government whose debt is largely denominated in domestic currency, and which has access to effective monetary policy instruments, will therefore face relatively lax fiscal sustainability constraints. If, on the other hand, government debt is largely denominated in foreign currency, or the government must adhere to a nominal anchor (as in the case of a currency board, a dollarized economy, or a full-bodied monetary medium), then the means of debt repayment must consist entirely of real resources, and the fiscal sustainability constraint will be strictest. In order to provide a streamlined analysis, the discussion in this section focuses on foreign currency-denominated debt; the section on political economy provides additional insight into currency mix decisions and their effects.

As mentioned previously, some of the earliest forms of sovereign debt directly acknowledged the centrality of fiscal sustainability by tying loan repayment to a specific revenue source. As early as the fourteenth century, Castilian *juros* – perpetual or lifetime bonds – entitled the holder to regular interest payments sourced from a specific tax stream. The bond was said to be "placed" (*situado*) on the tax stream, which was described as having been "sold" (*enajenada*), and could not be used for any other purpose. The annual payment was typically collected directly from the tax farmer or administrator. As long as the tax stream generated sufficient revenue, there was no question about the sustainability of the debt. This practice was later adopted by the French Crown when it established the *rentes sur l'Hôtel de Ville* and echoed in the practice of earmarking, one of the key financial reforms of the Glorious Revolution (North and Weingast 1989). In the case of Castilian *juros*, the arrangement effectively shifted the risk of fiscal nonperformance to the borrower; if the tax stream on which the bond was placed dried up, the Crown was not obliged to make up the shortfall and was not considered to have defaulted. After the Glorious Revolution, the British government improved on this by guaranteeing debts regardless of the health of earmarked fiscal streams; a sinking fund absorbed surpluses and covered shortfalls.

Most debt, however, is not tied to specific fiscal sources. In order to assess its sustainability, it becomes necessary to consider the overall solvency of the state. Evaluating the solvency of an individual is a straightforward exercise conducted by banks and other lenders countless times per day; one asks whether the present value of the individual's free cash flow exceeds the present value of scheduled debt repayments. The formula can, in principle, be extended to a sovereign, with some

important modifications. First, while an individual's earnings are limited by their life cycle, a state is, in principle, an infinitely lived agent. This carries a key advantage, in that the state never has to repay the debt principal if it doesn't want to; it can simply refinance its loans indefinitely and only pay interest. The second difference is that, while an individual's earnings first grow and then decline over their life cycle, a state's revenues may conceivably grow indefinitely, and hence they could support an ever-increasing debt load without risking insolvency. A financial solvency test for a state may therefore be reformulated as requiring that the present value of revenues exceed the present value of future interest payments.

In practice, estimating the future growth path of state revenues is quite challenging, as government policies and the business cycle can introduce large fluctuations over the short run. An alternative approach is to operate on the assumption that the government has the tools to claim a relatively stable portion of GDP (especially if averaged over a longer period) and require that the debt-to-GDP ratio not grow in order to consider the debt load sustainable. This criterion, developed by the International Monetary Fund, is overwhelmingly favored by international lending organizations (IMF 2003). It is noteworthy that such a rule is agnostic regarding how large the debt level can be relative to GDP; a stable debt-to-GDP ratio implies that interest payments are being met out of ordinary revenue rather than new borrowing, and hence the debt is deemed sustainable. Conversely, unsustainability can be defined as a situation where a country needs to incur new debt just to pay interest on its existing obligations.

Assessing the sustainability of a country's debt during a particular period using the IMF methodology requires fairly complete information on its national accounts. The ideal data to carry out this exercise include annual or more frequent estimates of revenue, ordinary (noninterest) expenditure, interest payments, the stock of debt, interest rates, and the growth rate of the economy. These can be plugged directly into a sustainability criterion equation, as derived, for example, in Aizenman and Pinto (2005), to obtain a straightforward answer on whether the debt-to-GDP ratio is stable.

While the data needed to assess fiscal sustainability are often readily available for developed countries in modern times, they can pose significant challenges for the cliometrician working on pre-statistical periods. Annual GDP series are available for several developed economies from the early twentieth century onward; further back in time their frequency becomes more sporadic, their reliability declines, and obtaining continuous series requires increasingly strong assumptions. Currently, the most complete set of historical GDP statistics suitable for fiscal sustainability analysis is the one provided by Bolt et al. (2018), who rebase and extend the estimates originally compiled by Angus Maddison for 14 countries starting as far back as 1820. For earlier periods, the analysis must rely on increasingly noisy income estimates calculated at much longer intervals or focus on a different measure, such as fiscal revenue.

A second empirical hurdle is the compilation of fiscal accounts –revenue, expenditure, and interest payments – as well as estimating the stock of outstanding debt. The availability and completeness of these data vary enormously across time and

states. Gaps in the data, however, can sometimes be overcome by resorting to the standard accounting identities linking the different variables. The earliest usable continuous historical series is the one compiled by Drelichman and Voth (2010) for Castile between 1566 and 1596. Other early examples include the Netherlands between 1601 and 1712 (De Vries and Van der Woude 1997), the United Kingdom between 1698 and 1793 (Crafts 1995; Mitchell 1988), and France between 1720 and 1780 (Sargent and Velde 1995; Velde 2007). Additional fiscal accounts for these and other European countries are available in the European State Finance Database (Bonney 2007). Table 1, based on Drelichman and Voth (2014a), summarizes the data for the United Kingdom, Castile, and France.

The figures in Table 1 must be used with ample caution, as GDP estimates in particular are subject to substantial noise. Taking them at face value, however, these early modern economies appear, on many dimensions, similar to contemporary ones. Their debt-to-GDP ratios are in the same range as modern developed countries, and their rate of revenue growth matches or exceeds that of twenty-first-century economies. Where they differ is in the magnitude of debt service; while a country that devoted 50% of its tax revenue to paying interest would be considered in serious financial distress today, these pre-modern polities spent money on little else than waging war and paying interest on the debt they had incurred for fighting previous wars. Without a large civil service or a social safety net, they managed to generate large enough primary surpluses to keep their debt loads manageable. In fact, of the three economies represented in Table 1, only France found itself in a manifestly unsustainable equilibrium, largely for political reasons (Sargent and Velde 1995). Castile's debts were eminently sustainable (Drelichman and Voth 2010). The United Kingdom, for its part, reached debt-to-GDP ratios as high as 260% in 1821 without sustainability being called into question (Reinhart and Rogoff 2009). Deft fiscal management and effective government bureaucracy supported one of the largest imperial expansions known to history (Bordo and White 1991; Brewer 1988). The debt burden was progressively reduced throughout the nineteenth century on the strength of Britain's economic growth and the long Pax Britannica. The United Kingdom once again experienced very high debt-to-GDP ratios after World War II, peaking at 238% in 1947, but almost half of the outstanding obligations during that period were intergovernmental loans provided by the United States at preferential terms, rather than market-traded bonds (Ellison and Scott 2017).

Table 1 Fiscal accounts for early modern economies

	United Kingdom	Castile	France
Average debt service/revenue	43% (1698–1793)	51% (1566–96)	38% (1720–80)
Growth rate of tax revenue	1.47% (1692–1794)	3.30% (1566–96)	1.26% (1661–1717)
Primary surplus/revenue	19.5% (1698–1794)	31.50% (1566–96)	14.2% (1662–1717)
Debt/GDP	74% (1698–1793)	14.7–51.4% (1566–96)	81.1% (1789)

Any analysis of fiscal sustainability is only as good as its underlying data. When trustworthy statistics are readily available, correctly performed sustainability calculations are no more than a mechanical exercise that consulting firms and international lenders can perform without too much trouble. As the focus shifts to the pre-statistical past, the craft of the cliometrician becomes increasingly important in unearthing scant sources and compiling the bits and pieces of information that may eventually offer a glimpse of whether a king could actually repay his loans. So far, the literature has found that in the vast majority of cases, debt turns out to be sustainable. This is good news for *homo economicus* – it would be worrisome if lenders offered their capital to borrowers that could not possibly pay them back. The resources used to repay loans – or, at least, to pay interest in perpetuity – are by and large present in most historical instances of sovereign debt. The next step, therefore, is to ask what incentives sovereigns have to actually keep their commitments, why defaults happen, and why financiers still want to lend to seemingly fickle rulers and states.

First-Generation Reputational Models

The seminal theoretical treatment of sovereign debt in the economics literature is due to Eaton and Gersovitz (1981). At its core, the model features a repeated game between a borrower and one or more lenders. The equilibrium involves a trigger strategy. As long as the borrower services the debt and repays the principal on schedule, the lenders keep providing new loans or rolling over old ones. Should the borrower default, however, the lenders will permanently withdraw. The borrower loses access to the income smoothing provided by debt and must self-finance. This is a costly outcome, as the entire state operations will now have to be financed out of taxation or inflation, regardless of the business cycle.

The first requirement for reputational equilibria to be possible is that lenders must be able to maintain a solid boycott on defaulting borrowers. While this can be plausibly achieved in certain situations, it can prove problematic in others. Sustaining a boycott against a defaulter is easy if there is only one lender but becomes increasingly difficult as the number of market participants grows. If there are a large number of lenders, a defaulting borrower can attempt to overcome the boycott by offering a premium to some of them. In the absence of coordination, such a strategy is likely to succeed (Greif et al. 1994). To overcome this problem, second-generation models (discussed further below) incorporate coordination devices among lenders, such as cheat-the-cheater strategies.

The second requirement is that sovereigns must not be able to replicate on their own the income-smoothing services offered by lenders. A borrower with sufficiently deep pockets, for example, could institute a self-financing scheme. This could take the form of building a sinking fund or establishing a large deposit with a trusted financial institution, commonly referred to in the literature as a "Swiss bank." Whenever income exceeds expenditure, the government would deposit the surplus in the sinking fund or the Swiss bank. Conversely, when facing a shortfall, it would

withdraw from them, thus exactly replicating the financial flows provided by sovereign debt. It is nonetheless unlikely that defaulting countries would be able to save enough so as to effectively smooth over future budgetary shortfalls. The vast majority of defaults occur in times of fiscal distress, when governments are least able to put aside substantial resources to guarantee steady financial flows (Tomz and Wright 2007). As for Swiss bankers, despite their reputation as impervious to external pressures, they turn out not to be a great option when it comes to acting as financial agents for delinquent borrowers. Historically, international financiers have sometimes gone under, taking with them the deposits of foreign governments. In modern times, defaulting countries have found it extremely costly to operate in global financial markets before reaching a settlement with their creditors. After its 2001 default, the threat of legal action by creditors meant that Argentina was unable to hold any sizable financial assets outside its own borders until it reached a final settlement in 2016.

The Eaton–Gersovitz model is important because it establishes that sovereign lending relationships can exist even when lenders do not have access to legal or military enforcement mechanisms. The mere threat of losing access to capital markets is sufficient to keep borrowers honest. This feature of the theory aligns well with the historical record, as lenders seldom have the ability to impose punishments on sovereigns other than by withdrawing their loans. The model, however, runs into trouble when contrasted with other aspects of historical experience. The first one of these is that there are no cases in which a borrower has been permanently excluded from capital markets, something that typically happens only when a state entity is itself dissolved; even then, its debts are typically inherited by the successor state (Reinhart and Rogoff 2009). Most defaulters regain access to credit within a few years, often borrowing again from the same creditors. This is a common problem for theoretical frameworks focusing on why borrowers repay; its implications are discussed in the section on the nature of defaults further below.

The toughest hurdle for the Eaton–Gersovitz model comes from quantitative calibration exercises. Even assuming that defaulters can be effectively excluded from capital markets forever, the loss in utility to the sovereign resulting from the lost smoothing services is not large relative to the onetime gains from confiscating the lenders' assets. As a result, the simple reputational mechanism can only sustain very low levels of debt, typically around 10% of GDP (Arellano and Heathcote 2010). Since the vast majority of borrowers carry much higher debt levels, sometimes up to ten times more, alternative explanations are needed.

Sanction-Based Models

In the Eaton–Gersovitz setup, the only punishment a defaulter faces is an exclusion from capital markets, which appears to be insufficient to keep borrowers from engaging in opportunistic behavior. It is also unrealistic to expect that a defaulter would be excluded in perpetuity. If the marginal utility of a sovereign's debt is really high, it would make sense to offer high interest rates or a lump-sum settlement to entice one or more lenders to return to the market. In light of these shortcomings of reputational models, an alternative literature proposes that punishments above and

beyond the mere exclusion from capital markets – known as "sanctions" – are needed to sustain debt repayment as an equilibrium.

The first sanctions model was proposed by Bulow and Rogoff (1989). They noted that, at any point in the life of a loan, the contract can be thought of as an immediate cash transfer and a schedule of similar future transactions. For example, when a loan first originates, the disbursement is the immediate transfer, while the scheduled repayments are the promised transactions in the future. However, because the sovereign can default, the future schedule is essentially "cheap talk," as the parties can agree to renegotiate it at any time. A loan can therefore be thought of as a series of renegotiations – a "constant recontracting" process. Punishing defaulters with costs above and beyond the exclusion from capital markets (sanctions) provides appropriate incentives for borrowers to honor their debt obligations. In the parlance of game theory, the imposition of sanctions supports renegotiation-proof lending equilibria.

The central difficulty for any model based on sanctions is immediately apparent. How can a private lender sanction a sovereign? The relationship is by definition lopsided – the borrower has access to enforcement mechanisms and can exert some form of legitimate violence, while lenders, in principle, cannot. Bulow and Rogoff conceptualize their model by allowing lenders to cause a defaulting country to lose a fraction of its GDP, for example, by imposing trade sanctions or freezing assets held in foreign bank accounts. But can this idea capture actual events? In modern times, governments are reluctant to impose wide-ranging trade restrictions for the benefit of bondholders; Eichengreen and Portes (1989), for example, document that it was the Foreign Office's policy in the 1930s not to intervene on behalf of British investors in international credit disputes. Private efforts to freeze a defaulter's assets are similarly unlikely to succeed, either because of sovereign immunity or because defaulters are quick to shelter any liquid holdings from the reach of international courts (as Argentina did, e.g., after its 2001 default, thought at the cost of having its ability to operate in global markets severely curtailed).

There are, however, specific historical episodes that robustly support the sanctions view. The most obvious of these is the military takeover of revenue-generating assets, such as Britain's seizure of the Khedive's financing in the 1870s, before directly incorporating Egypt into the British Empire. Another example is the "Roosevelt corollary," an addition to the Monroe Doctrine, which saw the US threaten military action in Latin America in case of default; markets responded positively, and bond prices increased substantially. These extreme examples, dubbed "supersanctions" by Mitchener and Weidenmier (2010), clearly fit the Bulow–Rogoff framework and have been shown to increase borrower compliance. Supersanctions, by their very nature, require the cooperation of a military power; it is much harder to show that private parties could impose a punishment on the same scale, and there are virtually no historical instances of such an event.

Though it is unlikely that a country that does not honor its debts will be slapped with trade sanctions, trade does nonetheless shrink significantly following a default. Typically, no private party is granted a higher credit rating than the country in which it is based, and hence exporters and importers see their ability to borrow severely

curtailed. As a consequence, GDP falls sharply (Rose 2005). The loss of output and the difficulties of operating in international markets can be interpreted as a punishment beyond the exclusion from capital markets, thus providing additional support for the sanctions view.

The sanctions literature shares the same weak spot as first-generation reputational models. If the threat of punishment is effective, defaults should not be observed in equilibrium, as no lender would want to incur the steep costs associated with them. Under this theoretical framework, defaults remain an off-the-equilibrium-path outcome, and the ones that do take place should carry severe consequences for debtors, which are seldom observed in practice. While it is possible to interpret some features of sovereign lending at specific points in history as being consistent with the sanctions theory, the sheer prevalence of defaults, as well as the many debtors that emerge from them not much worse for wear, continues to suggest that a different approach is needed.

Second-Generation Reputational Models: Cheat-the-Cheater Strategies

The Eaton–Gersovitz and Bulow–Rogoff models were both seminal contributions to the sovereign debt literature. They provided powerful explanations for several stylized facts while capturing some important institutional features of sovereign debt arrangements through history. Nevertheless, historical evidence made it clear that important gaps remained, opening the door for a wave of second-generation models, in which reputation, aided by advances in contract theory, once again took center stage.

Reputational equilibria uniformly rely on the ability of borrowers to exclude a defaulting lender from further participation in capital markets. As hinted earlier, this can be a difficult proposition in a competitive environment. If a ruler is unable to self-finance in order to mitigate the consequences of sharp income fluctuations, the marginal utility for external financing is very high, setting the stage for attempts at breaking the boycott. A government in default could pick out one or two powerful financiers and offer them sufficiently attractive interest rates to break ranks with those refusing to provide funds. In a similar setup involving rulers that confiscate the goods of merchants, Greif et al. (1994) show how, absent a coordination device, rulers will always be able to find fresh trading opportunities by enticing individual merchants with richer-than-normal rewards.

In practice, boycott breaking in sovereign debt markets is rare. When observed, it is often the result of official lending for political reasons, rather than market-mediated transactions. For example, the Soviet-era lending to Cuba and the Venezuelan government loans to Argentina in the 2000s both occurred during international boycotts. The former, at preferential rates, was never repaid, while the latter, well above market rates, was quite profitable for the Venezuelan government. Importantly, both were extended for geopolitical reasons rather than for a strictly financial profit motive.

Kletzer and Wright (2000) asked how the boycotts underpinning reputational equilibria could be maintained in the absence of effective institutional enforcement on the lenders' side, which they termed an "anarchic" environment. Their model proposed a "cheat-the-cheater" strategy. If one lender, enticed by rich promises from a defaulting sovereign, breaks the boycott, all other lenders retaliate by refusing to conduct any further business with the boycott breaker and confiscating any assets they may be able to seize. A similar behavior between traders had been documented by Greif (1993) in his famous study on the Maghribi traders. Kletzer and Wright did not anchor their theory with any historical examples, but Drelichman and Voth (2011a) showed how Genoese and German bankers lending to Philip II of Spain operated under this exact principle. In normal times, lenders participated in what was by all accounts a well-functioning, competitive market. During the 1575 and 1596 default episodes, however, all lending stopped. Genoese bankers joined forces in a coalition that offered a single negotiating front to the king. At the time, virtually all bankers had dealings with each other outside of sovereign lending, engaging in commercial ventures, holding each other's financial deposits in several European locations, and marrying into each other's families. This provided ample opportunity for retaliation against boycott breakers and ensured the stability of the coalition. Although the king repeatedly tried to reach separate agreements with some of the richest families, the bankers always refused. Remarkably, cheat-the-cheater incentives also worked to keep non-Genoese bankers from breaking the boycott. The most notable case was that of the Fugger of Augsburg, who did not have any commercial or lending relationships with the Genoese. The king also tried to entice them to resume lending by offering twice the normal interest. The head of the German house declined, and their correspondence shows they feared that the king would use any new loan to settle his debt with the Genoese and default on the Fugger instead, effectively cheating on the cheaters.

Theories of Default

A lingering shortcoming of all the models discussed thus far is their treatment of defaults. The single-minded focus on explaining why sovereigns repay their debts often leads to ignoring that quite substantial proportion of loans – up to 20% in particularly troubled years – that are not serviced or repaid according to their original terms. Any theory where defaults are an off-the-equilibrium-path outcome will struggle with such a sobering statistic. Accordingly, from the very early studies of sovereign debt in both theory and history, scholars have been proposing ways in which defaults could actually be conceptualized as equilibrium outcomes, at least some of the time.

Bubbles, Sentiment, and Irrationality

One of the first and most enduringly popular explanations for the prevalence of defaults is that lenders are myopic. Braudel (1966) argued that most early modern sovereign bankruptcies were made possible by the greed of bankers, who, blinded by

the promise of high interest rates and easy profits, were time and again driven to their ruin by calculating, unscrupulous kings. An updated version of this argument underpins the work of Reinhart and Rogoff (2009), who argue that irrational lender sentiment is the common denominator of most defaults across the five centuries of history they examine.

Explanations based on irrationality are hard to reconcile with the central tenets of economic theory; individuals that are motivated by sentiment over reason, or are fooled by smart rulers, should be driven out of the market. In order for sovereign lending not to die out, it would be necessary for an endless supply of new myopic financiers to keep stepping up, replacing those that suffered the last suspension, only to be defaulted upon in turn. The historical record shows that such dynamics are implausible. When lending resumes after a default is settled, the lenders that provide the new funds are typically the same that were affected by the suspension of payments (Drelichman and Voth 2014a). They won't have necessarily profited from the episode, but they certainly are not ruined, nor do they appear to be chastised enough to abandon the market.

A different approach considers some defaults as being enabled by debt bubbles. Flandreau and Flores (2009) show how Latin American bonds in the early 1820s became extremely popular with investors, who drove up their prices only to lose substantial amounts in a wave of defaults from 1825 onward. Bubbles, however, need not rely on irrationality; investors may be fully aware that the market prices of an asset do not reflect its fundamentals yet will still buy it with the explicit intention to sell at a profit before the trend is reversed and the bubble bursts. All buyers in a bubble are, in a sense, rational; some are just unlucky. However, despite being clearly important in a few specific episodes, bubble dynamics seem to be of little importance in the vast majority of defaults.

Excusable Defaults

In an early adaptation of the Eaton–Gersovitz framework, Grossman Grossman and Van Huyck (1988) noted that not all defaults are the result of opportunistic borrower behavior. Countries sometimes face extreme adverse events over which they have no control and which seriously strain their ability to make timely payments on financial obligations. Foreign military attacks, the sudden collapse of a key commercial or productive sector through blockade or natural disaster, and a rapid and unanticipated rise in interest rates are all factors that can undermine the fiscal capacity of a sovereign. When exogenous events like these lead to a default, borrowers typically understand it as a consequence of *force majeure* and do not impose a punishment on the borrower. Further lending is suspended only until the extenuating circumstances are reversed; the borrower is allowed to make a settlement payment, some of the debt may be forgiven or rescheduled, and new loans are offered in short order. Such defaults are termed "excusable," as borrowers are not held responsible for them to the same extent as with opportunistic behavior.

Identifying excusable defaults in the historical record involves establishing that the suspension of payments was triggered by an exogenous event, that the exclusion from capital markets was lifted as soon as a settlement became feasible, and that the

borrower regained access to credit largely on the same terms as before the default. This necessarily involves setting thresholds for what should be considered a "standard" exclusion length. Benjamin and Wright (2009) calculate that, between 1800 and the present, defaulters are excluded from capital markets for an average duration of 8 years. Drelichman and Voth (2014b) argue that Philip II of Spain's defaults in 1575 and 1596 were excusable. Both episodes were triggered when delays in precious metals shipments from the Americas coincided with Castile's military adversaries ramping up military operations and requiring an increase in outlays. Both were settled in less than 2 years, and the financing terms available to the king after each episode were essentially unchanged from those that prevailed before.

Market Power and Incomplete Contracting

One approach considers that lenders, rather than competing with each other in an atomistic way, may possess a degree of market power (Kovrijnykh and Szentes 2007). This may happen because only a handful of financiers are willing to lend to the government or because there is a large concentration of capital in the hands of a few prominent bankers. In such cases, lenders will use their leverage to increase interest rates and obtain profits above those that would prevail in a competitive market. Because of these extraordinary profits, however, they will also be reluctant to completely sever a lending relationship after a default. Borrowers, knowing this, will find it optimal to renege on their obligations every now and then. The frequency of these equilibrium defaults will be higher the more concentrated the market power of lenders. Conjunctures like the extreme concentration of capital in the hands of Florentine bankers in the Late Middle Ages, or the outsize influence of the Rothschild and Morgan banking houses during the late nineteenth and early twentieth centuries, may be good fits for this type of framework.

A second line of research starts from the observation that it would be desirable for borrowers and lenders to write contracts that account for different possible future situations, usually called "states of the world" in the specialized literature. Thus, for example, if a country were to unexpectedly experience a serious recession, lenders might extend the repayment horizon, perhaps in exchange for a slightly increased interest rate. This type of contingent contract has the potential to avoid the costs associated with a default and an exclusion from capital markets. An early example can once again be found thanks to the seemingly endless financial creativity of sixteenth-century Genoese bankers. Many of the contracts they offered to the Spanish monarchy contained clauses that explicitly modified the repayment terms in specific circumstances, such as a delay in the arrival of the Atlantic treasure fleets (Drelichman and Voth 2014b). These contingent clauses typically increased the costs of borrowing following adverse events for which the king was deemed responsible (such as a unilateral rescheduling), but left them unaltered, or even reduced them, in cases that were outside his control (such as bad weather in the Caribbean delaying the arrival of precious metals). More recently, some emerging economies have experimented with GDP-linked bonds, which increase payments to bondholders in good years and reduce them in recessions.

Despite their apparent efficiency and desirability, and the ingenuity of Genoese bankers notwithstanding, contingent sovereign debt contracts are, in practice, extremely rare. One possible explanation for their lack of popularity is that it is difficult to find objective indicators tied to the state of the economy that cannot be influenced by the sovereign. Philip II did not control the weather, and, through its influence on treasure shipments, the weather happened to be directly tied to the state of his finances. But magnitudes like GDP or inflation are a different story. Governments can certainly influence GDP growth or inflation rates, although doing so to avoid debt payments would seem much too costly a strategy. More importantly, however, a government can tinker with the measurement of macroeconomic magnitudes, biasing them in a direction contrary to the interest of creditors.

Between the moral hazard issues and the sheer number of unpredictable circumstances that may exogenously impact the fiscal position of a government, implementing fully contingent debt contracts appears to be little more than a chimera. It is simply not possible to provide for every potential state of the world; this leads to a situation known in contract theory as "incomplete contracting." Arellano (2008) pioneered a literature showing that, in such circumstances, when events take a turn not contemplated in the original agreement, defaults can arise as an equilibrium outcome.

That lenders and borrowers cannot write complete contracts, however, does not mean that they are ignorant of the nature of the game. Quite to the contrary, every actor in the market understands that, if events take an unpredictable turn for the worse, default is a likely outcome. In technical terms, defaults are part of an "implicit contract," which will be reflected in the risk premium built into the interest rate of any loans. Lenders are not naïve; they lend to sovereigns fully aware that default is a potential outcome and are compensated for that risk. This also explains why lenders may "forgive" defaulters after a partial settlement: they have been already compensated earlier through the higher rates they received throughout the course of the lending relationship.

The incomplete contracting literature jives well with the excusable defaults interpretation. Both recognize the uncertain environment in which sovereign lending takes place, and both allow for reputational equilibria that do not require sanctions, thus matching many of the empirical characteristics of lending and default that emerge from historical data.

The Preeminence of Reputation

In the horse race between reputational and sanction-based models, both have found a measure of empirical support, and both bring important elements to our understanding of sovereign lending. The sanction story is clearly superior in instances where the use of force has been applied to collect loans or to deter sovereigns from defaulting, with the gunboat diplomacy era providing a few historical junctures in support of this view. The negative impact of defaults on a country's trade can also be interpreted as being aligned with the sanctions view, but trade represents an increasingly

shrinking source of income as one looks farther into the past. Tomz (2007) examined the universe of sovereign lending on a global scale during the last three centuries, concluding that reputation, rather than sanctions, was the central factor underpinning loan repayment and default settlement in the vast majority of cases. Coupled with advances that can account for defaults as an occasional equilibrium outcome, this evidence suggests that market mechanisms and incentives are sufficient to make sovereign lending viable. The final linchpin is that sovereign lending is also profitable. As mentioned earlier, investors in government bonds received an average premium over safe assets of 0.4% in the nineteenth and twentieth centuries and up to 2% in early modern times, after accounting for the losses sustained in defaults. Sovereign finance may be complicated, kings may be whimsical, and defaults may shock investors and even ruin some. However, when viewed through the long lens of history, sovereign debt markets have been remarkably resilient and, in their best incarnations, have needed little more than the usual self-interest of borrowers and lenders to deliver gains for both.

The Political Economy of Debt

The scholarly dialogue between the theoretical and historical literatures has shown that sovereign debt markets can work, sustained by reputational equilibria. As soon as the first reputational models emerged, however, a new set of questions appeared: how do rulers build a reputation? What makes a government trustworthy? These would become an integral part of the research agenda of the then-nascent "new institutional economics" (NIE), a field itself deeply rooted in the intellectual sphere of economic history.

The question of what underpins the credibility of a ruler is at the center of North and Weingast's (1989) study of the political and financial consequences of the Glorious Revolution, undoubtedly one of the most influential pieces in the entire economic history literature. Its main argument is one and the same with the central creed of the NIE: institutions, not individual behavior, determine the long-run outcomes of state policy and, ultimately, of states themselves. The Glorious Revolution was, at its core, a radical institutional reform that enshrined merchant interests at the heart of the British system of government (O'Brien 2009). By subjecting the Crown to parliamentary supremacy, it stripped the king of the ability to expropriate subjects and foreign lenders alike, while ensuring that overall government policy would be more aligned with the preferences of the triumphant commercial elites. This policy was, in turn, crystallized in a set of institutions, among which the most prominent were the earmarking of taxes, the sinking fund, and the establishment of the Bank of England. These acted as constraints on the government's borrowing ability and as safeguards that ensured the availability of funds for repayment. The boost in credibility they generated, North and Weingast argued, enabled Britain to issue unprecedented levels of debt while drastically reducing its borrowing costs. This financial capacity would then prove crucial in defeating its enemies militarily, establishing an overseas empire, and, ultimately, creating the conditions that fostered the Industrial Revolution.

While the central argument of North and Weingast continues to inspire scholars and politicians alike, its empirical analysis has been called into question. Sussman and Yafeh (2006) showed that the interest rates in North and Weingast's article, calculated on the basis of bonds' face value rather than auction prices, did not reflect Britain's true cost of borrowing. Once corrected, the effects of the Glorious Revolution are much harder to detect. Interest rates remained at pre-revolutionary levels for several decades, before finally inching down in lockstep with those of other European countries. In fact, the pace of military events would appear to be a better predictor of rates than political reforms. More recently, Cox (2011, 2015) has argued that political institutions, such as the establishment of ministerial responsibility under Walpole, played as important a role as financial reforms in the wake of the Glorious Revolution.

Conclusion

Four decades after the development of the first reputational model, the sovereign debt literature has reached a reasonably mature state. Cliometricians, by creating comprehensive databases of debt issues and defaults, painstakingly uncovering key institutional details, and taking testable predictions to sometimes centuries-old data, have been at the very frontier of these intellectual efforts. Historical research has provided inspiration for diverse historical frameworks, supplied the data to validate them, and pointed to areas where further advances were needed. It is fair to say that few areas of economic research are so tightly integrated with economic history as the study of sovereign lending.

Cross-References

- ▶ Institutions
- ▶ Interest Rates
- ▶ Origins of the U.S. Financial System
- ▶ Political Economy

References

Abbas SA, Belhocine N, ElGanainy A, Horton M (2010) A historical public debt database. IMF Working Papers 10/2045

Aizenman J, Pinto B (2005) Managing economic volatility and crises: a practitioner's guide. Cambridge University Press, Cambridge, UK

Arellano C (2008) Default risk and income fluctuations in emerging economies. Am Econ Rev 98(3):690–712

Arellano C, Heathcote J (2010) Dollarization and financial integration. J Econ Theory 145(3):944–973

Benjamin D, Wright M (2009) Recovery before redemption: a theory of delays in sovereign debt renegotiations. UCLA Working Paper
Bolt J, Inklaar R, de Jong H, Van Zanden JL (2018) Rebasing Maddison: new income comparisons and the shape of long-run economic development. Maddison Project Working Paper 10
Bonney R (2007) European state finance database. 23 Aug 2007. http://www.le.ac.uk/hi/bon/ESFDB/
Bordo MD, Rockoff H (1996) The gold standard as a 'good housekeeping seal of approval'. J Econ Hist 56(2):389–428
Bordo MD, White EN (1991) A tale of two currencies: British and French finance during the Napoleonic wars. J Econ Hist 51(2):303–316
Braudel F (1966) The Mediterranean and the Mediterranean world in the age of Philip II. William Colins & Sons, Glasgow
Brewer JS (1988) The sinews of power. Harvard University Press, Cambridge, MA
Bulow J, Rogoff K (1989) A constant recontracting model of sovereign debt. J Polit Econ 97(1):155–178
Carande R (1987) Carlos V y Sus Banqueros. Crítica, Barcelona
Carlos AM, Neal L (2006) The micro-foundations of the early London capital market: Bank of England shareholders during and after the South Sea Bubble, 1720–25. Econ Hist Rev 59(3):498–538
Christodoulakis N (2014) Germany's war debt to Greece: a burden unsettled. Springer, Berlin
Cox G (2011) War, moral hazard and ministerial responsibility: England after the glorious revolution. J Econ Hist 71(1):133–161
Cox GW (2015) Marketing sovereign promises: the English model. J Econ Hist 75(1):190–218. https://doi.org/10.1017/S0022050715000078
Crafts NFR (1995) Exogenous or endogenous growth? The industrial revolution reconsidered. J Econ Hist 55(4):745–772
Cust R (1985) Charles I, the privy council, and the forced loan. J Br Stud 24(2):208–235
De Vries J, Van der Woude A (1997) The first modern economy. Cambridge University Press, Cambridge, UK
Dickson PGM (1967) The financial revolution in England. A study in the development of public credit. Macmillan, London
Drelichman M, Voth H-J (2010) The sustainable debts of Philip II: a reconstruction of Castile's fiscal position, 1566–1596. J Econ Hist 70(4):813–842
Drelichman M, Voth H-J (2011a) Lending to the borrower from hell: debt and default in the age of Philip II. Econ J 121(557):1205–1227
Drelichman M, Voth H-J (2011b) Serial defaults, serial profits: returns to sovereign lending in Habsburg Spain, 1566–1600. Explor Econ Hist 48(1):1–19
Drelichman M, Voth H-J (2014a) Lending to the borrower from hell: debt, taxes, and default in the age of Philip II. Princeton University Press, Princeton
Drelichman M, Voth H-J (2014b) Risk sharing with the monarch: excusable defaults and contingent debt in the age of Philip II, 1556–1598. Cliometrica 9(1):49–75
Eaton J, Gersovitz M (1981) Debt with potential repudiation: theoretical and empirical analysis. Rev Econ Stud 48(2):289–309
Eichengreen B (1996) Globalizing capital: a history of the international monetary system. Princeton University Press, Princeton
Eichengreen B, Portes R (1989) After the deluge: default, negotiation and readjustment of foreign loans during the interwar years. In: Eichengreen B, Lindert P (eds) The international debt crisis in historical perspective. MIT Press, Cambridge, MA
Ellison M, Scott A (2017) Managing the UK national debt 1694–2017. CEPR Discussion Paper 12304
Felloni G (1999) Scritti Di Storia Economica. Università di Genova, Genova
Felloni G (2006) La Casa Di San Giorgio: Il Potere Del Credito. Brigati Glauco, Genova
Flandreau M, Flores JH (2009) Bonds and brands: foundations of sovereign debt markets, 1820–1830. J Econ Hist 69(3):646–684. https://doi.org/10.1017/S0022050709001089

Greif A (1993) Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. Am Econ Rev 83(3):525–554
Greif A, Milgrom P, Weingast BR (1994) Coordination, commitment, and enforcement: the case of the merchant guild. J Polit Econ 102(4):745–776. https://doi.org/10.1086/261953
Grossman HI, Van Huyck JB (1988) Sovereign debt as a contingent claim: excusable default, repudiation, and reputation. Am Econ Rev 78:1088–1097
Homer S, Sylla R (2005) A history of interest rates, vol Fourth. Wiley, Hoboken
IMF (2003) World economic outlook. IMF, Washington, DC
Kletzer KM, Wright BD (2000) Sovereign debt as intertemporal barter. Am Econ Rev 90(3):621–639
Kovrijnykh N, Szentes B (2007) Equilibrium default cycles. J Polit Econ 115(3):403–446
Lindert P, Morton PJ (1989) How sovereign debt has worked. In: Sachs J (ed) Developing country debt and economic performance. Chicago University Press, Chicago
Mitchell BR (1988) British historical statistics. Cambridge University Press, Cambridge, UK
Mitchener KJ, Weidenmier MD (2010) Supersanctions and sovereign debt repayment. J Int Money Financ 29(1):19–36
Neal L (1991) The rise of financial capitalism: international capital markets in the age of reason. Cambridge University Press, Cambridge, UK
North DC, Weingast BR (1989) Constitutions and commitment: the evolution of institutional governing public choice in seventeenth-century England. J Econ Hist 49(4):803–832
O'Brien PK (2009) Mercantilist institutions for the pursuit of power with profit. The management of Britain's national debt, 1756–1815. In: The fiscal military state in eighteenth century Europe: essays in honour of P.G.M. Ashgate, Dickson
Pezzolo L (2003) Government borrowing before 1500. In: Mokyr J (ed) The Oxford encyclopedia of economic history. Oxford University Press, New York
Reinhart CM, Rogoff K (2009) This time is different: eight centuries of financial folly. Princeton University Press, Princeton
Rose A (2005) One reason countries pay their debts: renegotiation and international trade. J Dev Econ 77:189–206
Sargent TJ, Velde FR (1995) Macroeconomic features of the French revolution. J Polit Econ 103(3):474–518
Stasavage D (2011) States of credit: size, power, and the development of European polities. Princeton University Press, Princeton
Sussman N, Yafeh Y (2006) Institutional reforms, financial development and sovereign debt: Britain 1690–1790. J Econ Hist 66(4):906–935. https://doi.org/10.1017/S0022050706000374
Tomz M (2007) Reputation and international cooperation: sovereign debt across three centuries. Princeton University Press, Princeton
Tomz M, Wright M (2007) Do countries default in 'Bad Times'? J Eur Econ Assoc 5(2/3):352–360
Tracy JD (1985) A financial revolution in the Habsburg Netherlands: Renten and Renteniers in the county of Holland, 1515–1565. University of California Press, Berkeley
Velde FR (2007) John Law's system. Am Econ Rev 97(2):276–279

Corporate Governance

Carsten Burhop

Contents

Abstract

This article provides information on how financiers of public limited companies in the past ensured that they achieved a return on and a return of their investment. I am investigating this with examples from the USA, Great Britain and Germany. First, I examine when and to what extent public limited companies were formed and whether and to what extent ownership and control were separate. Moreover, control and incentive mechanisms, e.g. monitoring by supervisory board or incentive pay to managers, are investigated.

Keywords

Corporate governance · Joint-stock companies · United States · Britain · Germany · Executive compensation

C. Burhop (✉)
University of Bonn, Bonn, Germany
e-mail: burhop@uni-bonn.de

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_32

Introduction

Problems of corporate governance arise if principals and agents have different targets, if information is asymmetrically distributed between the two parties, and if the writing or monitoring of fully specified contracts is too costly. In such situations, the governance structure of a corporation allocates the residual rights to certain actors by using contracts, charters, or the law. In the end, "Corporate governance deals with the way in which suppliers of finance to corporations assure themselves of getting a return on their investment" (Shleifer and Vishny 1997: 737).

The starting point of any analysis is thus to provide evidence regarding the existence of corporations and the separation of ownership and control. I deal with those questions in sections "The Emergence and Relevance of Corporations" and "The Separation of Ownership and Control" of this entry. One important way to assess the economic relevance of agency problems is, at least according to Shleifer and Vishny, the pay-performance sensitivity of management compensation. Are managers and rent-seekers extracting big pay-checks from corporations or do they receive adequate incentive pay? Therefore, I deal with management compensation in section "Managerial Incentives." Other potential ways to solve agency problems identified by Shleifer and Vishny are legal institutions (see section "The Emergence and Relevance of Corporations"), ownership by large investors (see section "The Separation of Ownership and Control"), dividend payments (see section "Dividends"), board structure (see section "Boards"), and product market competition (see section "Product Market Competition").

Morck (2007) and Herrigel (2007a, b) provide useful reviews of the older literature. Thus, I focus on work published during the last decade and on contributions using quantitative data. The papers published in the volume edited by Randall K. Morck focus on ownership structures around the world, whereas the article written by Gary Herrigel is broader in scope. He distinguishes "old" and "new" historical corporate governance research. According to him, old research focuses strongly on questions regarding ownership dispersion, legal protection of minority owners, the role of banks, and the depth of financial markets. More recently, the evolution of national systems of corporate governance became an important line of research. Moreover, research based on archival material became prominent, whereas the evaluation of printed or published sources was used by earlier researchers.

In the end, Herrigel (2007a) distinguishes four explanatory camps. According to the Chandlerian view, professional managers were hired by owners of firms once these firms have been grown in scale and scope. This growth is considered as a natural process, but it can be constrained by political, legal, or cultural obstacles. However, it is not clear why firm growth leads to a separation of ownership and control since, historically, most growth has been financed by internal resources, especially retained earnings, and not by issuing shares. According to the political view, ownership remains, or becomes, concentrated to counterbalance political power concentration (in the hand of left wing parties). Third, an interest group view of governance systems has been put forward. According to this view, interest groups in closed economies wanted to set up a political and legal system protecting

their own interests, and not the interests of a broad group of shareholders. Thus, ownership concentration varies negatively with the openness of countries. Finally, the law-and-finance-school put forward the legal history of countries as an explanatory force. In particular, the common law legal system protects minority owners better than the civil law legal system. Thus, dispersed ownership emerges in the former, but not in the latter system. Obviously, this rather static view does not help very much in explaining variations of corporate governance systems over time.

Over the last decade, the focus of analysis shifted from the macro explanations summarized by Herrigel (2007a) toward micro explanations based on firm-level data. Moreover, the recent literature goes beyond correlations and associations. Causal identification of governance relations became popular. Ownership structure is still an important area of research, whereas the law-and-finance hypothesis has been rejected by many scholars.

The Emergence and Relevance of Corporations

What is a corporation? This question is not easy to answer, since definitions vary by country and by period. Being a separate legal "person" and having entity shielding are common among definitions used in the literature (see Hannah 2014, 2015; Sylla and Wright 2013, for a debate). Without a clear definition, counting corporations is not straightforward. In the 2010s, approximately 50 million corporations existed around the world, but less than 50,000 were listed on stock markets (Hannah 2015). By comparison, in 1910 only about 460,000 corporations existed; most of them from North America (62%) and Europe (30%). The most important countries were the USA, Britain, and Germany (Hannah 2015). Thus, this entry focuses on these three countries. In addition, one should note that most corporations were not listed on stock markets, and when they were listed, the liquidity of their shares was rather low. For example, around 10,000 British corporations appeared in stock market manuals as being listed, but only 600 were actively traded (Hannah 2015). Stock market listing and a certain market liquidity seem, however, necessary conditions for the separation of ownership and control.

Locking in large amounts of capital with limited liability differentiates corporations from other forms of business organizations, such as partnerships. In particular, the survival of corporations does not depend on the survival of their owners and corporations make risky investments in large projects more likely. Nevertheless, corporations usually account for only a small fraction of enterprises and the relevance of corporations for the economy depends, among many other factors, on the organizational menu available to entrepreneurs. The corporation seems to be important for large scale and risky investments, but for small- and medium-scale enterprises, other types of legal organization are often superior (Guinnane et al. 2007). Furthermore, the size of corporations was skewed, i.e., relatively few firms account for most of the capital raised (Hannah 2015). Therefore, corporate governance is usually investigated by looking at the largest firms.

The menu of organizational choices available in the USA, the UK, France, and Germany reveals that ordinary partnerships were available in all four countries, whereas other types were not. For example, limited partnerships with tradeable shares were available in France and Germany, but not in the USA and the UK. Moreover, private limited companies, i.e., companies with limited liability but not tradeable shares, were available in Germany from 1892, in the UK from 1907, and in France from 1925. Corporations with tradeable shares and limited liability became available in Germany in 1870, in France in 1867, and in the UK in between 1855 and 1862. Thus, by 1870 free incorporation was possible in the major European economies – several decades later than in most states in the USA (Guinnane et al. 2007). Differences in the organizational menu lead to differences in the frequency of incorporation. On the eve of World War I, for example, 2.5 corporations for 1,000 people existed in the USA, but less than 0.1 in Germany (Guinnane et al. 2007).

In Germany, the majority of firms had been set up as ordinary partnerships until the interwar period. From the 1890s onwards, private limited liability companies became the second most important legal type of newly established firms. Corporations account for less than 5% of the newly established firms between the 1860s and the 1930s (Guinnane et al. 2007). In France, partnerships account for more than 60% of newly established firms from the 1850s to the 1920s. Right after the introduction of private limited liability firms, partnerships became unimportant and the new type of enterprise accounts for more than 60% of new firms. Corporations were more important than in Germany, accounting for around one in five of the newly established firms (Guinnane et al. 2007).

Until the second half of the seventeenth century, large commercial organizations needed support by the Crown to incorporate in England. After the Glorious Revolution, this power shifted to parliament (Freeman et al. 2013). The situation was different in Scotland, since bottom-up political organization implied bottom-up ways to incorporate. From 1641 to 1693, the Scottish parliament enacted and modified rules for general incorporation (Freeman et al. 2013). After the union in 1707, the Westminster parliament became responsible for incorporation in England and Scotland and the parliament became more reluctant to grant incorporations rights, in particular, after the 1720 Bubble Act. With the repeal of the Act in 1825, incorporation became easier in Britain. Free incorporation of unlimited liability companies was established in Britain in 1844, followed by limited liability for most companies in 1855/56. Limited liability was extended to banking in 1858 and to insurance companies in 1862. In general, the 1862 Companies Act consolidated former acts and established free incorporation for limited liability companies.

Sylla and Wright (2013) present evidence regarding the formation of corporations in the early USA. Almost no corporations were formed during the colonial period, but incorporation took-off after the turn of the century. Between 1790 and 1861, more than 22,400 corporations were chartered by various state parliaments, and at least 4,000 were set up under the reign of general incorporation laws. From the 1790s to the 1830s, New England and the mid-Atlantic states were hot spots of incorporation, with the Southern states subsequently catching up. In most states,

incorporation required a special legislative act by the state parliament. General incorporation laws came in force, starting with New York in 1811. Most states and territories moved to this system in the antebellum period, in particular during the 1840s and 1850s. From the 1830s through the 1850s, the number of newly incorporated firms grew substantially. More specifically, many banks and railroads were formed during the 1830s, whereas railway companies dominated during the 1850s.

By the 1870s, free incorporation was possible in most advanced countries. Nevertheless, important differences among countries remained. In a seminal paper, La Porta et al. (1998) measure to what extent legal rules of individual countries offer protection for shareholders ("anti-director rights"). They evaluate legal codes of 49 countries as they existed during the 1990s. It turns out that, on average, countries having a common law legal system offer the best protection to shareholders, whereas countries having a French-style civil law legal system offer the worst protection. Britain, the USA, and Canada offer the highest level of protection – the same level as India or Pakistan. La Porta et al. (1998) point out that the effectiveness of the rules depends on law enforcement. This variable is strongly correlated with economic development – law is much better enforced in richer countries, like the USA and the UK, than in poorer countries like India or Pakistan. Moreover, they show that countries with weak rules and enforcement tend to substitute legal rules with other mechanisms, e.g., mandatory dividend payments. Nevertheless, substitutes are not perfect. Thus, the level of legal protection of shareholders is positively associated with stock market development: the widely held listed corporation tends to be more common in countries offering good protection for minority shareholders.

This law-and-finance-story has been controversially debated. Spamann (2010), for example, recalculated the anti-director index and demonstrates that the legal protection of shareholders was comparatively weak in the USA at the turn of the millennium, but relatively strong in Germany, France, and Japan – three of the leading civil law countries. In contrast, just before World War I, protection of shareholders was strongest in the USA and weak to nonexistent in Britain, France, and Germany (Musacchio and Turner 2013).

Franks et al. (2009) demonstrate that legal protection of investors was basically nonexistent in Britain until World War II. The UK anti-director rights index was one out of six points until the 1948 Companies Act. From 1948 to 1980, the score has been three, subsequently rising to five. Moreover, the public enforcement index was zero until 1986. Only the private enforcement index, which can take values between 0 and 1, indicates some protection for shareholders. This index was zero in 1900, 0.5 in 1929, and 0.67 between 1948 and 1967. Yet, common law systems might offer protection to shareholders not via government action, but via court cases. The evidence is, however, unfavorable. Prominent court cases in 1843 and 1883 confirmed the absence of shareholder rights (Franks et al. 2009).

In the USA, corporate law is primarily determined by the individual states (see Cheffins et al. 2013a for a short review). During the late nineteenth century, a "race to the bottom" began when states attempted to attract corporations with low taxes and few constraints on managerial behavior. In the end, Delaware won the race, but before World War I, New Jersey has been the place to incorporate. Legislators in the

USA knew that New Jersey attracted many corporations. Thus, the government of Delaware decided to copy this successful corporate law in 1899. A few decades later, in 1933, almost 30% of the NYSE-listed companies were incorporated in Delaware – followed by firms incorporated in New York (18%) and New Jersey (12%). In 1899, legal standards were fairly high in New Jersey and in Delaware. Shareholders were allowed to vote by proxy, shares were not blocked before the general meeting, minority shareholders could mobilize judicial powers against unfair managerial behavior, and shareholders had pre-preemptive rights. Thus, both states scored four out of six in the anti-director rights index, which was very high when compared to shareholders in late Victorian Britain. Moreover, the score remained at that level until it was reduced by one point due to a legal reform in 1967 (Cheffins et al. 2013a). The race to the bottom seemed to take-off during the late twentieth, not the late nineteenth century.

Franks et al. (2006) provide a review of German corporate law, starting with the 1843 Prussian corporate law. Until World War II, major reforms occurred in 1861, 1870, 1884, 1897, and 1937, but only the 1861 reform had an impact according to the anti-director rights index. This index increased from zero to one, since 10% of the shareholders could now ask for an extraordinary shareholder meeting. Surprisingly, the index remained at this level at least until the early 2000s. The other indices of the law-and-finance school also showed a low degree of legal protection in Germany around 1900. For example, the private enforcement index stood at zero in 1900 and remained at this level until 1987. Thus, shareholders in Germany were basically unprotected over the course of the twentieth century resulting in a high ownership concentration.

The cases of Delaware and the UK demonstrate that legal rules were not fixed centuries ago, but were determined in the political sphere, and are thus subject to change. Indeed, Pagano and Volpin (2005) demonstrate that democracies with a system of proportional representation tend to offer better protection to managers and employees, whereas democracies with a majoritarian political system (e.g., Britain and the USA) tend to offer better protection for shareholders. In a pure cross-sectional econometric investigation, the legal origin still has an impact on the level of shareholder protection. However, once we consider the dynamics of political processes, i.e., when we account for the fact that political decisions change the legal rules over time, the legal system loses its significance for the level of shareholder protection (Pagano and Volpin 2005). Moreover, Pagano and Volpin (2006) suggest the existence of a positive feedback loop: countries with good shareholder protection see a rising number of shareholders and these shareholders vote for parties fostering shareholder rights. The reverse holds for countries with weak shareholder protection. Thus, a bimodal world emerges: one part is populated by market-oriented, shareholder friendly countries while the other is not.

Foreman-Peck and Hannah (2015) investigate the impact of law, politics, and culture on the number of corporations, the impact of corporations on economic growth and exports, and the feedback of economic growth and exports on the number of incorporations. An economy dominated by a narrow elite should, according to Rajan and Zingales (2003), be a closed economy, i.e., an economy

with a low level of foreign trade. Furthermore, according to Pagano and Volpin (2005), more inclusive political institutions, in particular parliamentary systems with proportional representation, should lead to more inclusive economic institutions, i.e., a large number of welfare-enhancing corporations. Foreman-Peck and Hannah (2015) mobilize new data about the number of corporations in more than 80 countries in 1910 and combine them with existing data on GDP, trade, legal origin, and various other explanatory variables. They end up with a sample of 51 countries from around the world. Using OLS, 2SLS, and 3SLS regressions, they demonstrate (i) that richer countries which are open to international trade and having a common law legal system do have a relatively large number of corporations; (ii) that the number of corporations is not associated with the intensity of foreign trade; and (iii) that the number of corporations is positively related to the level of GDP.

Within-country variations of legal rules have been used to refute the law-and-finance thesis. For example, Freeman et al. (2013) compare developments in common law England with those in civil law Scotland between 1600 and 1850. In particular, they draw on a large dataset of 452 corporate charters from the period between the 1720 Bubble Act and the 1844 Companies Registration Act. In contrast to expectations from the law and finance literature, Scottish firms offered better protection to minority owners than English firms. A comparison of civil law and common law systems within one country is also possible for the USA, since Louisiana operated a French-style incorporation system until 1845. The difference in outcomes when compared to similar US states was rather small. The number of incorporations was slightly lower in Louisiana, but the average size of a corporation slightly larger (Sylla and Wright 2013).

Compliance of firms is also an important point recently evaluated by Guinnane et al. (2017). The 1856 British corporate law did not contain detailed governance rules. Instead, the parliament provided a model set of governance rules to make life easier, in particular for small corporations (Guinnane et al. 2017). To what extent did firms follow the model and to what extent did they change the rules? Were the rules chosen by firms better or worse for (minority) shareholders? Do we observe differences between large and small firms or between listed and non-listed firms? In general, corporate statutes shifted power from shareholders to directors and this pattern became more pronounced over time. Guinnane et al. (2017) investigate the articles of associations for three randomly selected cross sections of firms newly established in 1892, 1912, and 1927. The 1892 sample, for example, includes 54 firms. Average share capital was about 40,000 pounds and nine of the firms had 50 or more shareholders. Less than 10% of the firms accepted the governance rules as provided by parliament (Guinnane et al. 2017).

Governance rules are often complicated and sometimes several rules interact to get a specific result. Thus, looking at individual rules often does not help to understand the governance system of a corporation. Keeping this in mind, we may look at individual rules. For example, about 80% of the 1892 sample adjusted shareholder voting rights by moving from the graduated voting scheme proposed by parliament to a one-share, one-vote rule. Election of directors and restricting the power of directors is also one of the key tasks of shareholders. In 74% of the firms,

shareholders never got to elect a full board and in 64% of articles, directors could name one of themselves managing director (Guinnane et al. 2017). Guinnane et al. (2017) do not use advanced statistical methods to disentangle potential correlations or causations in their sample. Instead, they compare mean values of some relevant variables. It seems that firm size did not play a significant role. Large and small firms adjusted statutes in the same way. Listed and nonlisted firms were also quite similar, and there is no time trend visible in the data.

Firm-level governance and performance data were also mobilized by Burhop (2009) to evaluate the impact of governance rules of German banks during the 1870s on their valuation and performance. More specifically, he looks at 202 corporate charters and analyses how these charters regulate 28 governance issues. For example, the 1870 German corporate law contains a one-share, one-vote rule as a default with the possibility to change this rule in the corporate charter. Indeed, more than 90% of banks changed the rule, usually by restricting the voting rights of small shareholders (e.g., by requiring multiple shares to cast one vote) and block holders (e.g., by restricting the maximum number of votes an individual shareholder can cast). In addition, Burhop illustrates how certain rules interact to generate specific outcomes. For example, one of Germany's largest banks stated in its charter that the general meeting had to be announced 14 days before it would take place. In another rule, the charter prescribes that shareholders had to register shares 28 days in advance. Thus, only insiders could register shares right in time to attend the annual meeting.

The association of 28 governance rules with firm performance (measured by Tobin's Q, the change in the market to book ratio, and survival) is evaluated using a set of univariate and multivariate regressions. Unsurprisingly, most governance rules had no systematic association with firm performance. For relatively large banks, use of a one-share, one-vote rule has been associated with higher firm valuation (Burhop 2009). In principle – and in accordance with the theory of information efficient stock markets – no variable had a statistically and economically significant association with changes of firm valuation. Furthermore, voting rights for small shareholders as well as an independent and relatively large supervisory board enhanced the probability of firm survival during and after the financial panic of 1873 (Burhop 2009). In general, voting and monitoring rights seem to improve governance outcomes.

The Separation of Ownership and Control

In global perspective, transparency of corporate ownership structures is rather low. Faccio and Lang (2002), for instance, present one cross section of ownership data for about 5,200 European listed corporations. Widely held and family controlled firms dominate among European corporations at the end of the last century. Families are more important for relatively small firms and for nonfinancial corporations. In nearly all countries, family-dominated firms are more frequent than widely held firms; Britain and Ireland are the exception to this rule. Outside of Europe, widely held firms are most important in Australia, Canada, Japan, New Zealand, the USA, and

South Korea (La Porta et al. 1999). Thus, "One of the best-established stylized facts about corporate ownership is that ownership of large listed companies is dispersed in the UK and the USA and concentrated in most other countries" (Franks et al. 2009: 4009).

Comprehensive historical corporate ownership data for the nineteenth and twentieth centuries are in short supply. Britain and, to a certain extent, the USA are the exceptions. Firms established under the 1856/1862 UK corporate law were required to submit shareholder registers to the Registrar of Companies. Using this source, Acheson et al. (2015) collected ownership data for five cross sections for the period 1865 to 1900. In total, they collect 890 records for 488 unique firms. On average, companies had 410 shareholders. The number of shareholders tended to increase over time from about 300 to about 600. Insiders (i.e., directors plus shareholders owning more than 10%) held about 18.5% of capital and 16.1% of votes. The five largest investors represented 26.6% of capital and 22.2% of votes. Breweries tend to have a high ownership concentration, whereas banks tend to have a low concentration.

Using a multivariate regression analysis, Acheson et al. (2015) demonstrate that ownership tends to disperse over time, and that firms headquartered in London and firms having their shares listed on multiple stock exchanges had more widely dispersed ownership. Mergers tend to increase ownership concentration. The size of the company, the par value of each share, and being listed at the official market of the London stock exchange did not systematically affect ownership dispersion. The existence of nonlinear voting schemes, which became less popular over time, did not affect ownership concentration, but it did affect the concentration of voting rights. Voting concentration was lower in firms operating nonlinear voting schemes, i.e., these schemes tend to disfavor large owners.

In general, ownership concentration was lower in Victorian Britain than in modern Britain. In Victorian Britain, only one in three firms had a single shareholder owning more than 10%. This is similar to S&P-500 firms today. Yet, directors owned a much larger share of capital in the nineteenth century than during later periods. From the 1880s to the turn of the century, the median ownership share of directors was 9%. This declined to less than 3% during the interwar period and less than 1% today (Acheson et al. 2015).

Post-1900 data are also available. Foreman-Peck and Hannah (2012) provide a comprehensive reassessment of the divorce of ownership and control on pre-World War I Britain by looking at the ownership structure of the 337 largest listed companies. In particular, they look at the number of shareholders per firm and they calculate the ownership share of directors in these companies. Moreover, they evaluate if large discrepancies between ownership and voting rights existed. The median firm had 3,000 shareholders, while the mean number of shareholders per firm was 6,177 (Foreman-Peck and Hannah 2012). The median number of shareholders varies substantially among sectors, with railways (4,750 shareholders) at the top and breweries (1,600 shareholders) at the bottom of the ranking. In total, the 1911 data reveal more than two million shareholders. Thus, Britain was a "shareholder society" already before the Great War. In nearly all companies, they observe a divorce of

ownership and control. Only 5–10% of the firms evaluated were perhaps controlled by manager-owners. On average, board members controlled 2.5% of the capital and 2.8% of the votes. Shares were higher in manufacturing and breweries. Taking the voting power of board members as an indicator of the divorce of ownership and control, they demonstrate that this divorce was stronger in the UK in 1911 than in the USA in 1935 or 1995.

Franks et al. (2009) track the development of ownership structure of British companies over the twentieth century by drawing several random samples. Moreover, they develop indices of ownership dispersion and ownership mutation to assess the dynamics of ownership structures beyond simple measures, i.e., ownership shares of the largest one, three, or five shareholders. It turns out that the ownership structure has been quite similar at different moments in time during the twentieth century. Yet insiders (i.e., board members) were replaced by outsiders (i.e., institutional investors) over the course of the twentieth century as dominant shareholders. From 1900 to 1960, they observe dispersion of ownership, and since 1960 a concentration of ownership. Driving forces behind ownership dispersion have been acquisitions by way of equity exchanges, increases of firm size, and directors selling out. Remarkably, investor protection indices are uncorrelated with ownership dispersion. Trust and the existence of provincial stock exchanges are put forward to explain ownership dispersion in times of weak shareholder protection. Shareholders tend to live close to corporate headquarters and close to the board members managing the firms. In 1910, more than half of the shareholders lived less than six miles away from the corporate headquarter. Mutation of ownership becomes faster over time. Thus, early in the century controlling shareholders held the shares, whereas controlling stakes changed hands with increasing speed later in the century. This mutation of ownership is significantly and positively related to investor protection indices. Better protection of outside shareholders improved stock market liquidity and facilitated the transfer of large ownership stakes from insiders to outsiders.

Another explanation for the high degree of ownership dispersion in Britain around 1900 has been put forward by Hannah (2007). He suggests that the two-thirds rule of the London Stock Exchange contributed to the divorce of ownership and control in Britain before World War I. According to this rule, at least two thirds of the capital of listed firms has to be distributed to outside owners during the IPO. Cheffins et al. (2013b) evaluate if this has indeed been the case. Based on a sample of IPOs from the period 1900 to 1913, they demonstrate that many firms retained an inside ownership of more than one third of the capital. Moreover, they show a stronger separation of ownership and control for firms listed in the special settlement section of the stock exchange. This is surprising, since the two-thirds rule did not apply for firms opting for special settlement – the rule only applied to firms opting for an official quotation. Thus, Cheffins et al. neither supported the result nor the reason given by Hannah. Responding to this critique, Hannah and Foreman-Peck (2014) make clear that they focused on relatively large and established firms, whereas Cheffins et al. focused on relatively small firms shortly after their IPO. Moreover, the sample size of Cheffins et al. is much smaller and thus less

representative. However, when Hannah and Foreman-Peck restrict their sample to relatively small firms, they find the same result as Cheffins et al. Thus, the two-thirds rule operated by the London Stock Exchange most likely did not contribute to the separation of ownership and control in British corporations before World War I.

To what extent ownership and control of US corporations has been separated is still debated (see Cheffins and Bank 2009, for a review). Hilt (2008) documents and analyses the separation of ownership and management in early US corporations. In contrast to hypotheses put forward by classical writers, he presents evidence for a divorce of ownership and management of New York corporations already in the 1820s. On average, firms had 74 shareholders and the Gini-coefficient of ownership stakes was 0.57. Moreover, ownership was more concentrated than voting rights since nearly half of the firms had graduated voting schemes. Yet, this deviation from the one-share, one-vote rule did not have a systematic impact on the distribution of shares. Furthermore, managers owned 28% of the firms, a relatively high share when compared to data from the 1930s or 1990s. In addition, Hilt observes a correlation between managerial ownership and the absence of graduated voting schemes. Firms with a high managerial ownership tend to follow a one-share, one-vote rule. This was not good news for investors: share prices tend to be higher when managerial ownership stakes were low and when voting rights were restricted.

Holderness et al. (1999) look at managerial stock ownership of large US corporations over the twentieth century. More specifically, they compare the ownership structures of 1,419 listed firms in 1935 with ownership structures of 4,202 listed firms in 1995. Their key finding is surprising: managerial ownership is higher in 1995 than in 1935. Mean and median managerial ownership shares of all companies and of NYSE-listed companies were higher in 1995 than in 1935. In addition, this finding also holds for all 13 branches and for old as well as young firms. Moreover, the inflation-adjusted average holding of insiders is four times larger in 1995. In a next step, Holderness et al. (1999) look at the holdings of the top officer, making their results comparable to data presented by Jensen and Murphy (1990) for the years 1938, 1974, and 1984.The mean ownership share in 1935 as well as in 1995 was 1.25%, and the median ownership declined slightly from 0.09% to 0.06%. Jensen and Murphy report mean (median) holdings of the top officer of 1.7 (0.3%) in 1938 and 1.0 (0.03)% in 1984. In 1935, managerial ownership stakes tended to be lower in regulated industries (e.g., public utilities), in larger firms, and in firms with volatile share prices. In 1995, regulation had no effect, but size and share price volatility still had the same effect. Moreover, older firms tend to have smaller managerial ownership stakes (Holderness et al. 1999).

Ownership data for Germany are in short supply since most firms used bearer shares and only shareholders attending the general meeting had to register their shares. Searching shareholder lists in the archives of individual companies is, obviously, quite time-consuming and it is thus difficult to sample shareholder lists. Some systematic evidence is available in public archives since firms had to submit attendance lists of a general meeting to stock exchange officials in case of seasoned equity offerings or similar events. However, this rule has been in force only since 1897 and it results, perhaps, in a biased sample.

Using this source, Franks et al. (2006) base their analysis of ownership structures on 156 shareholder lists from 55 companies. Before World War I, the median annual meeting was attended by 14 to 18 shareholders, with a maximum of 259 shareholders attending a meeting of Deutsche Bank. The numbers do not change much during the interwar period. The five largest shareholders present at annual meetings during the first half of the twentieth century represented about 80–90% of the share capital attending the meeting. On average, two-thirds of the total share capital was present at general meetings. Thus, ownership concentration was quite high in Germany and it tended to increase slightly over time. Moreover, large inside owners (i.e., members of the executive and supervisory board) were more important than large outside owners until the 1920s. In addition, we observe a pronounced move of voting power away from individuals toward banks and other firms. Banks did not invest their own money into shares. They mostly cast proxy votes for other investors.

Managerial Incentives

Executive compensation has been a hot topic in corporate governance research during the last three decades (see Edmans and Gabaix 2016 for a review). In particular, the steeply rising compensation of CEOs has been – and is – controversially debated. Economic theory puts forward four theories to explain rising compensation: managerial rent extraction, the provision of incentives, the scale of the firms, and increasing returns to general rather than specific skills. The rent extraction and the incentive explanation are obviously related to corporate governance.

Bayer and Burhop (2009) provide evidence for the incentive hypothesis by evaluating the impact of the 1884 German corporate law reform on executive compensation. This reform strengthened monitoring incentives and possibilities for shareholders and reduced the problems of asymmetric information between principals and agents. Thus, managerial incentives can be reduced. Indeed, using data for a sample of 37 manufacturing firms, they show that the sensitivity of executive pay and accounting profits declined by 50–75% after the implementation of better corporate governance rules. Robustness checks for banks and firms employing only one executive point in the same direction. Moreover, they show that total bonus payments increase after the reform and that the profit share of executives is broadly constant. The effect of the reform has thus not been the level of compensation, but the association of bonus payments with firm performance.

Impressive research has been conducted using data from the USA. The main reason is most likely the relatively easy access to large data sets. When the US government strengthened shareholder rights after the Great Depression, more financial information became available. Since 1938, firms had to disclose information regarding executive pay. Disclosure of pensions and deferred compensation followed in 1952 (Dew-Becker 2009). Using data available since the 1930s, Frydman and Saks (2010) document executive compensation in the USA between 1936 and 2005. Regarding the level of pretax compensation, they distinguish three phases: slightly declining real compensation from the mid-1930s to the early 1950s,

slowly increasing compensation from the early 1950s to the mid-1970s, and strongly increasing compensation subsequently. Moreover, they also document shifts in the composition of compensation. In the beginning, executives earned salaries and current bonuses. Long-term bonuses became more important from the 1960s onwards and stock options took-off in the 1980s.

While pretax compensation measures the costs of executives for the firms, it does not reflect incentives for executives, since incentives result from after-tax compensation. In the beginning, after-tax compensation fell more than pretax compensation since tax rates were high. Since the mid-1960s declining tax rates imply a stronger growth of after-tax compensation when compared to pretax compensation. Moreover, compensation grew relatively equally for all executives until the early 1990s. Afterwards, CEO compensation grew much faster than compensation of other top-level executives (Frydman and Saks 2010).

Over time, the association between executive compensation and firm size (i.e., the market value) changed substantially. The correlation between the growth of executive compensation and the growth of market value of firms has been 0.1 from the mid-1930s to the mid-1970s and about one afterwards). Quite similarly, the correlation between executive compensation and firm performance was also quite stable form the 1950s to the 1980s. Subsequently, the size of the correlation doubled (Frydman and Saks 2010. This changing correlation can neither be explained by firm size nor by underlying firm characteristics (like growth opportunities, return on assets, or industry regulation). Moreover, the pay-performance sensitivity is not affected by the size, the composition of the board, or the existence of a large outside shareholder. This latter finding leads Frydman and Saks (2010: 2128) to the conclusion that "corporate governance did not play a significant role in the rise of executive pay."

This, obviously, raises the question: What does play a role? Frydman and Molloy (2012) put forth the idea that unions affected executive compensation. Starting from the observation of declining executive compensation during the 1940s, they investigate the cause driving the decline. To do this, they mobilize four cross-sectional data sets covering the compensation of the three best paid executives from about 250 firms for the 1940s. Comparing the median compensation of executives with the average annual earnings in the economy, they calculate that executives earned 24 times the average annual pay in 1940, but only 17 times in 1949. Moreover, mean and median executive compensation not only declined when compared to average annual earnings in the economy, but also when compared to the inflation rate. Executives lost purchasing power during the 1940s. One reason for this extraordinary development might have been wartime regulation of pay. For example, increases of high salaries were prohibited between October 1942 and November 1946. Nevertheless, one quarter of executives received a salary increase. Remarkably, we do not observe huge differences between either war and nonwar industries or between high- and low-income industries. Furthermore, income tax regimes did also not affect executive compensation. Finally, they use a multivariate regression analysis to uncover factors associated with executive pay. The only factor depressing the relative – but not the absolute – level of executive compensation after the war is

the degree of unionization. The governance structure of firms, e.g., board size and board structure, did not affect executive compensation.

Managerial ownership can also mitigate agency conflicts between shareholders and executives. Calomiris and Carlson (2016) investigate this relationship by using data from the US banks from the late nineteenth century. In particular, they evaluate the endogenous emergence of governance mechanisms to limit rent-seeking and to keep risks in line with shareholder preferences. They employ examination reports for 206 national banks from 37 larger cities as a data source to test their hypothesis that corporate governance becomes less formal when managers (i.e., the president, the vice-president, and the cashier) own a relatively large share of the bank. Managerial ownership is, however, endogenous and they use unforeseen managerial turnover events (like death or illness of the manager) as an instrument to identify causal effects of ownership on governance. Governance is reflected in a five-component index (board meets at least monthly, high percentage of outsiders on board, active discount committee, president bonded, and cashier bonded). These components are positively correlated with each other and each is negatively correlated with managerial ownership.

On average (mean), the three managers owned collectively 24% of the share capital, outside directors on the corporate board owned another 15%, and 61% of shares were owned by outsiders (Calomoris and Carlson, 2016). By using a multivariate regression, Calomiris and Carlson demonstrate a negative and statistically significant association between formal governance and managerial ownership. Moreover, managerial ownership is negatively associated with managerial turnover. Using an instrumental variables approach, they demonstrate a causal impact of managerial ownership on risk taking – higher ownership results in a less risky loan portfolio and a lower probability of bankruptcy during the 1893 financial panic. Looking at the interrelationship between managerial ownership, formal governance, and rent seeking, it turns out that higher ownership shares are associated with higher managerial salaries, whereas more formal governance is associated with lower salaries. The level of insider loans depended neither on the size of managerial ownership nor on formal governance rules. However, the distribution of loans among insiders depends on these variables. Calomoris and Carlson observe a positive correlation between managerial ownership and loans to managers. Taken together, the evidence shows a positive association between managerial ownership and managerial rent-seeking, and a negative association between ownership and risk. Formal governance instruments were thus used to reduce rent-seeking.

Dividends

In perfect capital markets, shareholders should not care about dividend payments since they should not affect the value of a firm. When capital markets are imperfect, e.g., due to agency problems, dividend policy becomes relevant for firm value. For example, announcing a higher dividend may result in a higher stock price if money is

transferred from greedy managers to owners. La Porta et al. (2000) explain and test two agency models of dividends. According to the outcome model, dividends are paid because minority owners force insiders to do so. According to the substitution model, firms pay dividends to build up a reputation with minority shareholders, and this helps to issue additional equity. Empirical evidence from the 1990s tends to support the outcome model.

Historical data have also been used to test theories of dividend policy and to derive conclusions regarding the corporate governance system. In a seminal paper, Braggion and Moore (2011) use data from London around 1900 to evaluate the predictive power of various theories explaining the impact of dividend policy on stock prices. They show that dividend announcements disclose private information held by firm insiders to shareholders. A positive dividend announcement yields abnormally positive returns at the stock market, whereas a negative dividend announcement yields a negative return. Thus, investors take an announcement of a dividend change as information about the future cash flow of the firm. If this is indeed the case, a positive dividend announcement should be followed by higher earnings – and this is indeed the case. Moreover, Braggion and Moore investigate whether agency considerations are also relevant to understand dividend payments in Victorian Britain. Using the cross-sectional variation in the data, they find no effect. Dividend announcements have the same effect for all kinds of firms, leaving no room for competing explanations.

Campbell and Turner (2011) investigate whether dividend payments were used as a substitute for weak legal protection of shareholders in Victorian Britain. Using a sample of 823 companies listed in 1883, they demonstrate that pay-out ratios were similar in the cross section of firms. The only variable significantly related to pay-out ratios is firm size – large firms tend to pay relatively generous dividends. In addition, dividend policy (pay-out ratio or dividend-par value ratio) are positively associated with Tobin's Q.

Deloof et al. (2010) investigate dividend policy of Belgian firms between 1905 and 1909, with a special focus on the role of universal banks. Firms having a banker on board may not need to pay high dividends since the bank affiliation is already a positive signal to investors and since banks may provide capital. Yet, it might also be the case that bankers on board pressure firms to pay abnormally low dividends and to keep cash holdings, since this makes credit default less likely. At first glance, firms affiliated with a bank were more likely to pay dividends, their dividend yield was higher on average, they were more likely to continue dividend payments on an established level, and they had a higher pay-out ratio. In addition, bank-affiliated companies were more reluctant to cut dividend payments. In a multivariate setting, they provide evidence for the certification hypothesis. Bankers on board certify the good investment opportunities of firms (reflected in a high Tobin Q) having a higher pay-out ratio. But another explanation is possible: firms may need to pay higher dividends to compensate shareholders for the risk that bankers on board channel funds from owners to creditors. Yet, Deloof et al. demonstrate that this is most likely not the case, since a more intense relationship with a bank (i.e., reflected in several board members) does not result in higher dividend payments.

Guenther (2017) provides some evidence for Germany by looking at dividend announcements in 1895. For a sample of 166 firms, he finds statistically and economically significant positive abnormal stock market returns after a positive dividend announcement. Moreover, he finds that this effect is larger for relatively small firms and for firms providing relatively little financial information. Thus, we have some evidence for the dividend signaling hypothesis.

Boards

An extensive literature deals with the relationship of boards and corporate governance (Adams et al. 2010; Fear and Kobrak 2010). The literature based on historical data typically focuses on short periods of time and specific firms or events. In contrast, Lehn et al. (2009) has a broader scope by looking at the composition of corporate boards for 82 US companies during the period 1935–2000. He reports that board size is positively related to firm size (captured by market value) and negatively related to growth opportunities (reflected in the ratio of market-to-book value of assets). Large firms need more monitoring capacity and need to draw expertise from many different board members. Fast growing firms need smaller boards ready for quick action. Boards get larger, at least temporally, after mergers and acquisitions and during geographic expansion of corporations. Furthermore, the number of insiders declined by about two-thirds between 1950 and 2000 (Lehn et al. 2009). Using a 2SLS regression, Lehn et al. demonstrate that board size and board composition did not affect corporate performance once the factors driving board size and composition are accounted for.

The impact of bankers as board members on corporate performance has been investigated. In a classic paper, Ramirez (1995) looked at the function of J.P. Morgan for corporate investment in the USA around 1900. He compares 16 firms linked to J.P. Morgan to 30 unlinked firms between 1908 and 1912. The key question is whether being connected to J.P. Morgan mitigates cash flow restrictions. Ramirez estimates a standard cash flow to investment regression including firm fixed effects. It turns out that investment of firms not related to J.P. Morgan reacts negatively and statistically significantly to changes in cash flow, whereas cash flow played no role for investment of firms related to Morgan. Thus, monitoring by a banker alleviated capital market imperfections by reducing asymmetric information between investors and managers. Moreover, Simon (1998) shows that market valuation fell by 7% when bankers left corporate boards after passage of the Clayton Act in 1914. According to Simon, however, this decline is not an indication of monitoring services provided by banks, but an indication of declining market power. Bankers on boards of various firms from the same industry helped to establish collusion among firms. This coordination activity ended in 1914. The decline in stock market value thus reflects an increase in consumer surplus.

Recently, Frydman and Hilt (2017) used the Clayton Act to provide a causal interpretation of a classical question: did investment banks on corporate boards mitigate the problems of asymmetric information and help all shareholders or do

they exploit the additional information they get on the board to exploit rents from firms? If the monitoring role dominates, restricting board membership should lead to higher interest payments and less investment. If rent-seeking dominates, shareholder value should increase once bankers are removed from boards. In principle, the Clayton Act prohibited underwriters of railroad securities to sit on the board of the same railroad company. Thus, investment banks had to choose between underwriting a business and membership on its board. However, the relevant part of the Clayton Act was not implemented in 1914 and Congress postponed it until it went into effect in 1921. Nevertheless, the partners of J.P. Morgan & Co. announced immediately after passage of the Clayton Act that they were going to leave the boards of 30 companies.

Frydman and Hilt's empirical analysis was based on accounting, governance, and stock market data for 71 NYSE-listed railroads and 79 NYSE-listed industrials and utilities. Usually, railroads were linked to high-ranking investment banks, they had many interlocking directorates with other railroads, and they had a large number of shareholders. When it became clear that the act came into full force, the stock prices of the affected railroad companies reacted negatively, whereas returns of utilities and industrials did not decline. Looking beyond short-run returns, they show that Tobin's Q, investment rates, and leverage of affected railroad companies declined, whereas average interest payments increased. Thus, all proxy variables indicate a value to having bankers on board, a value that has been destroyed by policy intervention.

This difference-in-difference approach is supplemented by an instrumental variables regression. The strength of the railroad underwriter relationship in 1913 – before passage of the Clayton Act – is used as an instrument to causally explain changes taking place after 1920. Reassuringly, the instrumental variables approach confirms the earlier results (Frydman and Hilt 2017). The monitoring role of bankers on board has been especially relevant for railroads heavily using debt finance. Moreover, monitors from the leading investment banks created more value for railroads than did monitors from less prestigious underwriters.

A causal interpretation of board connections is also provided by Deloof and Vermoesen (2016). More specifically, they investigate the value of corporate boards for Belgian firms during the Great Depression, using it as an external shock allowing the identification of the causal relationship between board structure (as reflected in board size, presence of bankers on board, and the busyness of board members) and firm performance. During normal times, board structure is endogenous, since firms chose their board to maximize firm value. A larger board may provide more expertise, but the risk of free-riding is also larger. While bankers may provide debt finance, they may also ask for a low-risk firm strategy to make loan repayment more likely. Busy boards may offer many contacts to other firms, but busy officers may become too busy – in particular, during a crisis.

Using data for 150 firms for the period 1928–193,1 Deloof and Vermoesen (2016) estimate a random effects model to explain the market-to-book valuation of the firms. In principle, and with varying degrees of statistical and economic significance, all three indicators of board structure are positively related to firm valuation until the onset of the Great Depression, but these relationships turn negative during the Depression. The main driver during the Depression is board busyness. While the

effect of bankers and of board size is often insignificant, the effect of board busyness remains significant. In normal times, a busy board is value enhancing, but the effect turns negative during an extreme downturn.

Beyond bankers on board, politicians on board have become a focus of recent research. For example, Grossman and Imai (2016) provide evidence for a weakly negative association between board membership of members of parliament and valuation of banks. They collected data for all English and Welsh banks for the period 1879–1909 on biannual frequency. In particular, they collect balance sheet, stock market, and board composition data as well as firm fixed data, e.g., distance to London and year of foundation. One reason for the weak effect might be the stability of English and Welsh banks during this period. Managers simply did not need political connections when business was calm and successful. Board members are allocated to three groups: noble members, members of parliament (together, they form the group of politically connected board members), and other members. Between 1879 and the turn of the century, about 20% of the banks had a member of parliament on the board. Subsequently, the share increased to 35%. The share of noble directors increased steadily from about 5% in 1879 to 35% in 1909.

To assess the impact of board structure on stock market returns, Grossman and Imai (2016) calculate a fixed effects panel regression of all observations and they use a propensity matching estimator based on total bank assets to neutralize size effect. The latter method yields basically no significant results, whereas a cross-sectional regression approach provides some (rather weak) evidence of a positive association between board connections and stock market returns. Using a dynamic fixed effect regression, there is no effect of noble directors on stock market returns and even negative effects of MPs board presence on returns. This result obviously raises the question why banks put MPs on board. It turns out that some banks – namely small banks and those located fairly far away from London – profited from politically connected board members.

To account for potential selection bias, Grossman and Imai (2016) perform event studies for six general elections. They focus on election results where board members won or lost with a small margin. This indicates that the election outcome was not clear from the beginning and one may thus expect a financial market reaction to the outcome. However, no event had a significant effect on stock returns. Thus, they conclude that politically connected directors had no equity value for English and Welsh banks between 1879 and 1909.

Braggion and Moore (2013) search for value-enhancing effects of politicians and they provide evidence of a positive impact of political board members on stock market valuations for new-tech firms in late Victorian Britain. Election or reelection of a board member to the parliament results in an increase of stock market value of about 2%. Moreover, firms having a politician on board were more likely to make a seasoned equity offering. They use standard event-study methodology to uncover potential effects of MPs on stock market values. In particular, they look at the elections of 1895, 1900, and 1906 to measure potential effects. However, in most constituencies, election results were pretty predictable and should have been factored into stock prices a long time before the election. To detect potential effects, they focus

on those districts where results were rather unpredictable. While they still find no effect for all firms, they do find an effect for new-tech firms. In addition, they investigate whether politically connected firms raised more money by way of seasoned equity offerings or bond offerings. When all firms are considered, this is not the case. Yet, a positive impact of political connections for new-tech firms is detected.

Product Market Competition

Evidence regarding the interaction of product market competition, corporate governance, and firm performance is reviewed by Crafts (2012). Product market competition makes it easier to assess managerial performance since price signals from the market are more precise. Moreover, product market competition acts as a disciplinary device, fostering a decent performance of lazy managers. According to Crafts, dispersed ownership and restricted voting rights contributed to the weak productivity performance of UK railroads before World War I. In addition, the increasing tendency toward a separation of ownership and control after World War II and the absence of strong shareholders fostered managerial ineffectiveness in an environment with weak product market competition from the 1940s to the 1970s. The deregulation of financial markets and stronger competition on product markets changes this from the 1970s onwards.

The potential managerial failure at railroad companies during the nineteenth century has been investigated in detail by Campbell and Turner (2015). They look at the impact of competition on firm performance in the railway sector, controlling for ownership structure. Railways were incorporated with limited liability by special parliamentary acts, started in 1826. Frequency of travel and price regulation started with the 1844 Railway Act. This act also allowed the nationalization of lines authorized after 1844 and it allowed concessions for competing lines. The mid-1840s were also characterized by a boom and bust cycle in the railway sector, with railroad companies established after 1843 typically serving relatively sparsely populated areas. Campbell and Turner look at the ability and cost to passengers of choosing between competing lines and they investigate whether competition increased during the 1840s. In 1843, no substitute was available for two-thirds of the lines, whereas more than 80% of railroad lines could be substituted by travelling on other lines in 1850. In particular, competition increased on main lines connecting important cities. As a consequence, prices declined.

Building lines in sparsely populated areas and increased competition between lines might be reasons for declining financial performance of railroad companies during the 1840s. Managerial failure might be another reason. In a counterfactual analysis, Campbell and Turner (2015) evaluate the potential impact of three managerial strategies for established lines (i.e., railroad lines already in operation before 1844). The strategies are: no expansion and no mergers; no expansion, but mergers with other lines; expansion and mergers. The first scenario is correlated with the worst outcome, the last scenario with the best outcome. Since managers of established lines indeed followed an expansion and merger strategy, there is no

sign of managerial failure. Moreover, a regression analysis shows that "those firms which had the greatest potential exposure to competition were those which expanded the most" (Campbell and Turner 2015: 1267). Thus, there is little evidence for managerial failure of British railroad executives during the 1840s.

The relationship between product market competition and corporate governance has also been investigated by Burhop and Lübbers (2009). In particular, they evaluate the impact of the formation of the RWKS, a large cartel dominating the German coal market from 1893 until World War II, on productive efficiency of the cartelized firms. Reducing incentives from the product market should lead to declining efficiency of firms. However, shareholders might increase incentives by offering generous bonus payments to managers. Indeed, the executive board of cartelized firms received 3.6% of profits as bonus payments, whereas executives of non-cartelized firms received only 2.5%. Moreover, an econometric model demonstrates that an increase of bonus payments by 1% (730 Mark) reduced inefficiency by about 790 Mark. Thus, increasing managerial incentives is a good answer in face of declining product market competition. However, the numbers also suggest that executives received nearly the entire efficiency gain as bonus payment.

Conclusion

Our knowledge about historical corporate governance has substantially improved over the last decade. Theory-based data collection efforts and the application of modern econometric methods markedly contributed to this progress. Today, we can be quite sure that separation of ownership and control has been a feature of many – perhaps most – listed UK and US corporations before World War I. With respect to Germany and other countries, much more research is necessary. Moreover, ownership structures of US corporations before the 1930s are a somewhat under-researched area. Furthermore, the separation of ownership and control and its evolution over time is unrelated to the legal origins of countries. The contributions of other macro-theories explaining ownership structures seem also not to explain very much. Consequently, historical corporate governance research turned from cross-country toward cross-firm evidence. In addition, institutional differences across time and space have been used to go beyond associations and correlations and to uncover causal factors driving governance structures and outcomes. Future research may combine mirco-data from several countries to assess the potential impact of cross-country differences of corporate governance.

Cross-References

- ▶ Early Capital Markets
- ▶ Financial Markets and Cliometrics
- ▶ Financial Systems
- ▶ Institutions
- ▶ Origins of the U.S. Financial System

References

Acheson GG, Campbell G, Turner JD, Vanteeva N (2015) Corporate ownership and control in Victorian Britain. Econ Hist Rev 68(3):911–936

Adams RB, Hermalin BE, Weisbach MS (2010) The role of boards of directors in corporate governance: a conceptual framework and a survey. J Econ Lit 48(1):58–107

Bayer C, Burhop C (2009) Corporate governance and incentive contracts: historical evidence from a legal reform. Explor Econ Hist 46(4):464–481

Braggion F, Moore L (2011) Dividend policies in an unregulated market: the London stock exchange, 1895–1905. Rev Financ Stud 24(9):2935–2973

Braggion F, Moore L (2013) The economic benefits of political connections in late Victorian Britain. J Econ Hist 73(1):142–176

Burhop C (2009) No need for governance? The impact of corporate governance on valuation, performance, and survival of German banks during the 1870s. Bus Hist 51(4):569–601

Burhop C, Lübbers T (2009) Cartels, managerial incentives, and productive efficiency in German coal mining, 1881–1913. J Econ Hist 69(2):500–527

Calomiris CW, Carlson M (2016) Corporate governance and risk management at unprotected banks: national banks in the 1890s. J Financ Econ 119:512–532

Campbell G, Turner JD (2011) Substitutes for legal protection: corporate governance and dividends in Victorian Britain. Econ Hist Rev 64(2):571–597

Campbell G, Turner JD (2015) Managerial failure in mid-Victorian Britain? Corporate expansion during a promotion boom. Bus Hist 57(8):1248–1276

Cheffins BR, Bank S (2009) Is Berle and means really a myth? Bus Hist Rev 83(3):443–474

Cheffins BR, Bank SA, Wells H (2013a) Questioning law and finance. US stock market development, 1930–70. Bus Hist 55(4):601–619

Cheffins BR, Koustas DK, Chambers D (2013b) Ownership dispersion and the London stock Exchange's 'two-thirds-rule': an empirical test. Bus Hist 55(4):670–693

Crafts N (2012) British relative economic decline revisited: the role of competition. Explor Econ Hist 49(1):17–29

Deloof M, Vermoesen V (2016) The value of corporate boards during the great depression in Belgium. Explor Econ Hist 62:108–123

Deloof M, Roggemann A, van Overfeld W (2010) Bank affiliations and corporate dividend policy in pre-World War I Belgium. Bus Hist 52(4):590–616

Dew-Becker I (2009) How much sunlight does it take to disinfect a boardroom? A short history of executive compensation regulation in America. CESifo Econ Stud 55(3–4):434–457

Edmans A, Gabaix X (2016) Executive compensation: a modern primer. J Econ Lit 54(4):1232–1287

Faccio M, Lang LHP (2002) The ultimate ownership of Western European corporations. J Financ Econ 65:365–395

Fear J, Kobrak C (2010) Banks on board: German and American corporate governance, 1870–1914. Bus Hist Rev 84(4):703–736

Foreman-Peck J, Hannah L (2012) Extreme divorce: the managerial revolution in UK companies before 1914. Econ Hist Rev 65(4):1217–1238

Foreman-Peck J, Hannah L (2015) The diffusion and impact of the corporation in 1910. Econ Hist Rev 68(3):962–984

Franks J, Mayer C, Wagner HF (2006) The origins of the German corporation – finance, ownership, and control. Rev Financ 10(3):537–585

Franks J, Mayer C, Rossi S (2009) Ownership: evolution and regulation. Rev Financ Stud 22(10):4009–4056

Freeman M, Pearson R, Taylor J (2013) Law, politics, and the governance of English and Scottish joint-stock companies, 1600–1850. Bus Hist 55(4):636–652

Frydman C, Hilt E (2017) Investment banks as corporate monitors in the early twentieth century United States. Am Econ Rev 107(7):1938–1970

Frydman C, Molloy R (2012) Pay cuts for the boss: executive compensation in the 1940s. J Econ Hist 72(1):225–251

Frydman C, Saks RE (2010) Executive compensation: a new view from a long-term perspective, 1936–2005. Rev Financ Stud 23(5):2099–2138
Grossman RS, Imai M (2016) Taking the lord's name in vain: the impact of connected directors on 19th century British banks. Explor Econ Hist 59(1):75–93
Guenther J (2017) Capital market effects around dividend announcements: an analysis of the Berlin stock exchange in 1895. Account Hist Rev 27(3):249–278
Guinnane T, Harris R, Lamoreaux NR, Rosenthal J-L (2007) Putting the corporation in its place. Enterp Soc 8(3):687–729
Guinnane TW, Harris R, Lamoreaux NR (2017) Contractual freedom and corporate governance in Britain in the late nineteenth and early twentieth centuries. Bus Hist Rev 91(2):227–277
Hannah L (2007) The divorce of ownership from control from 1900: re-calibrating imagined global historical trends. Bus Hist 49(4):404–438
Hannah L (2014) Corporations in the U.S. and Europe 1790–1860. Bus Hist 56(6):865–899
Hannah L (2015) A global corporate census: publicly traded and close companies in 1910. Econ Hist Rev 68(2):548–573
Hannah L, Foreman-Peck J (2014) Ownership dispersion and listing rules in companies large and small: a reply. Bus Hist 56(3):509–516
Herrigel G (2007a) Guest editor's introduction: a new wave in the history of corporate governance. Enterp Soc 8(3):475–488
Herrigel G (2007b) Corporate governance. In: Jones G, Zeitlin J (eds) The Oxford handbook of business history. Oxford University Press, Oxford, pp 470–497
Hilt E (2008) When did ownership separate from control? Corporate governance in the early nineteenth century. J Econ Hist 68(3):645–685
Holderness CG, Kroszner RS, Sheehan DP (1999) Were the good old days that good? Changes in managerial stock ownership since the great depression. J Financ 54(2):435–469
Jensen MC, Murphy KJ (1990) Performance pay and top management incentives. J Polit Econ 98(2):225–264
La Porta R, Lopez-de-Silanes F, Shleifer A, Vishny RW (1998) Law and finance. J Polit Econ 106(6):1113–1155
La Porta R, Lopez-de-Silanes F, Shleifer A (1999) Corporate ownership around the world. J Financ 54(2):471–517
La Porta R, Lopez-de-Silanes F, Shleifer A, Vishny RW (2000) Agency problems and dividend policy around the world. J Financ 55(1):1–33
Lehn KM, Patro S, Zhao M (2009) Determinants of the size and composition of U.S. corporate boards: 1935–2000. Financ Manag 38(4):747–780
Morck RK (2007) A history of corporate governance around the world. Family business groups to professional managers. University of Chicago Press, Chicago
Musacchio A, Turner JD (2013) Does the law and finance hypothesis pass the test of history? Bus Hist 55(4):524–542
Pagano M, Volpin PF (2005) The political economy of corporate governance. Am Econ Rev 95(4):1005–1030
Pagano M, Volpin PF (2006) Shareholder protection, stock market development, and politics. J Eur Econ Assoc 4(2/3):315–341
Rajan RG, Zingales L (2003) The great reversals: the politics of financial development in the twentieth century. J Financ Econ 69:5–50
Ramirez CD (1995) Did J.P. Morgan's men add liquidity? Corporate investment, cash flow, and financial structure at the turn of the twentieth century. J Financ 50(2):661–678
Shleifer A, Vishny RW (1997) A survey of corporate governance. J Financ 52(2):737–783
Simon MC (1998) The rise and fall of bank control in the United States: 1890–1939. Am Econ Rev 88(5):1077–1093
Spamann H (2010) The "antidirector rights index" revisited. Rev Financ Stud 23(2):467–486
Sylla R, Wright R (2013) Corporation formation in the antebellum United States in comparative context. Bus Hist 55(4):653–669

Part VI

Government, Health, and Welfare

Anthropometrics

Richard H. Steckel

Contents

Abstract

This chapter describes the rise, evolution, and durability of the new anthropometric history, an interdisciplinary approach to assessing biological aspects of the standard of living in the past using measurements such as height and weight, which are abundant from numerous historical sources such as military muster rolls, identification systems, slave manifests, and prison records. By comparing anthropometric data with traditional sources used by economists, such as wages, income, and occupation, scholars have found intriguing similarities and differences in these approaches to assessing the standard of living or human welfare

R. H. Steckel (✉)
Ohio State University, Columbus, OH, USA
e-mail: Steckel.1@osu.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_61

broadly defined. The chapter provides a brief history of anthropometric measures as used by anthropologists and human biologists, and gives an overview of the methodology that is useful for interpreting anthropometric data used by social scientists. Much of the discussion considers applications of anthropometric evidence to topics such as slavery, inequality, industrialization, mortality, colonization, and economic development. It concludes with notes on possible pitfalls of using anthropometric data that are important to practitioners in the area and with identification of research frontiers.

Keywords
Health · Height · Weight · Living standards

Introduction

The word "anthropometrics" descends from the Greek *anthropos* ("man") and *metron* ("measure"), and so the term refers to the measurement of man, traditionally in a physical sense. Anthropologists take hundreds of measurements, and in the nineteenth century, they were consumed by dimensions of the head and their relationship to "race" (Gould 1996). Anthropologists have long since abandoned this approach to race or ethnic identity but the term endures.

In cliometrics, the term refers to the study of height and weight, mainly because these measures have meaning in human biology and they are readily available (particularly stature) in the historical record. One can find reports on shoe size, chest expansion, arm length, and so forth, but these are difficult to interpret in relation to biological aspects of the standard of living.

This chapter provides a brief account of anthropometrics in the study of economic history beginning with its earliest applications to the analysis of important historical questions, including the nature of slavery, dimensions of inequality, the consequences of industrialization, and the effects of colonization, among others. The list of applications is long and growing, in part because height records number in the tens of millions around the globe. Femur length and various lesions assessed from skeletal remains recently entered the laboratory of economic history.

Origins

Surprisingly, interest in the human body and its dimensions did not begin with anthropology or human biology but with art (Tanner 1981). Artists of the Renaissance sought to recover ancient knowledge of human form and function by taking detailed measurements of Greek and Roman statues with a goal of learning ideal body proportions. The first medical person to take an interest in measuring the human body may have been Johann Elsholtz (1623–1688) who originated the term "anthropometry."

The intellectual history of the two major approaches to measuring human welfare (national income accounts) and auxology (the study of human growth) have two things in common: the first substantial efforts occurred in the seventeenth and eighteenth centuries and early studies were sporadic, imprecise attempts made by individuals. Unlike national income, however, investigators could make useful anthropometric measurements on a small scale. Systematic national income data awaited government involvement and support in the twentieth century while important progress in auxology was made by the middle of the nineteenth century (Studenski 1958; Tanner 1981)

Table 1 charts milestones in anthropometry from the perspective of human biology. Researchers took initial steps in the seventeenth and eighteenth centuries, but progress was slow until the second quarter of the nineteenth century. The realization that environmental conditions systematically influenced growth stimulated interest in growth studies in the 1820s. Auxological epidemiology arose in France, where Villermé studied the stature of soldiers; in Belgium, where Quetelet measured children and formulated mathematical representations of the human growth curve; and in England, where Edwin Chadwick inquired into the health of factory children (Villermé and Golfin 1829; Quetelet 1835; Chadwick 1842). After examining the heights of soldiers in France and Holland and studying the economic conditions in their places of origin, Villermé concluded in 1829 that poverty was

Table 1 Milestones in auxology

Place	Investigator	Year	Events or developments
Germany	Elsholtz	1654	Graduation thesis on Anthropometria
Germany	Jampert	1754	Cross-section measurements of stature by age
Germany	Roederer	1754	Measures and weights of newborns
France	Montbeillard	1777	First longitudinal study from birth to adult
France	Villermé	1829	Studied environmental influences on growth
England	Chadwick	1833	First survey of factory children
Brussels	Quetelet	1842	First mathematical formulation of growth
England	Roberts	1876	Used frequency distributions to assess fitness; studied growth by social class
U.S.	Bowditch	1877	School surveys; analyzed velocity of growth
Italy	Pagliani	1879	Longitudinal studies; school surveys
England	Galton	1889	Studied inheritance of height; introduced regression coefficient
France	Budin	1892	First infant welfare clinic established
U.S.	Boas	1891–1932	Tempo of growth; concept of developmental age; growth studies in anthropology; standards for height and weight
France	Godin	1903	Detailed growth surveillance
U.S.	Baldwin	1921	Supervised the first large longitudinal study
England	Douglas	1946	First national survey of health and development
England	Tanner	1952	Models underlying clinical standards

Source: Extracted from Tanner (1981)

much more important than climate in influencing growth. In 1833, the English Parliament put these ideas into action in legislation using stature as a criterion in evaluating minimum standards of health for child employment.

The greatest strides in the modern study of human growth occurred beginning in the mid-nineteenth century with the work of Charles Roberts, Henry Bowditch, and especially, Franz Boas (Boaz 1898; Roberts 1876, 1878; Bowdotch 1877). Roberts raised the level of sophistication in judging fitness for factory employment by using frequency distributions of stature and other measurements, such as weight-for-height and chest circumference. Bowditch assembled longitudinal data on stature to establish the prominent gender differences in growth. In 1875, he supervised the collection and analysis of heights from Boston school children, a data set on which he later used Galton's method of percentiles to create growth standards. In a career that spanned several decades, Boas identified salient relationships between the tempo of growth and height distributions and in 1891 coordinated a national growth study, which he used to develop national standards for height and weight. Later he pioneered the use of statistical methods in analyzing anthropometric measurements and investigated the effects of environment and heredity on growth. The volumes by Phyllis Eveleth and Tanner (1976, 1990), *Worldwide Variation in Human Growth*, summarize the explosion of growth studies in the twentieth century.

Methodology

Generations of observers have sensed that stature reflects biological well-being because casual empiricism identified factors behind stunting. Farmers knew, for example, that animals required certain foods to thrive and to work, and that plants were most productive if they received adequate amounts of water and sunlight. Trial and error could readily identify these forces but careful scientific observation and measurement were required to sponsor the rise of epidemiological auxology, clinical practice, and human biology.

In recent decades, the study of human growth has become heavily interdisciplinary, delving into the sciences of nutrition, genetics, biochemistry, and epigenetics, the latter being the study of ways that gene expression (whether active or dormant) is influenced by environmental and biological factors, such as diet, exercise, disease, and age (Francis 2012; Launer 2016). Researchers now recognize critical periods of human development, the most sensitive to environmental conditions being the period in utero and especially the first trimester, when rapid cell division creates specialized organs and structures. This new material may be difficult reading for social scientists, but several publications provide accessible introductions (Tanner and Preece 1989; Ulijaszek et al. 1998; Cameron and Bogin 2012).

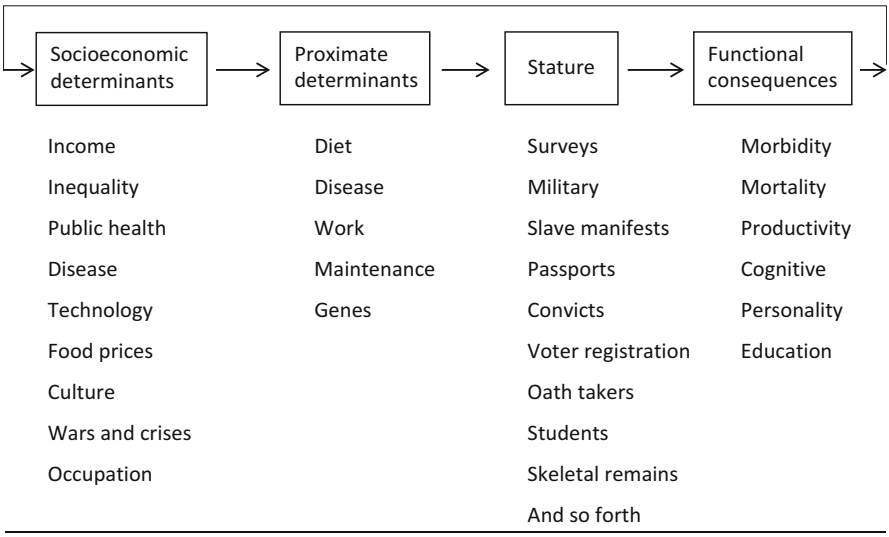
Although human biologists and physical anthropologists have known for some time that socioeconomic factors such as social class impinge on child growth and therefore adult height, a richer understanding of the relationship began to emerge when economists, historians, and other social scientists joined the conversation in the 1970s (Steckel 1998). Economic historians introduced new sources of data,

added several useful concepts, and discovered numerous puzzles or apparent anomalies in the past that elucidated the contribution of socioeconomic factors to growth. Because the historical record encompasses a rich variety of human experiences, their efforts helped to illuminate intergenerational influences on body size, measure the human capacity for growth following extreme deprivation, and expand the knowledge of cultural conditions that are ultimately expressed through proximate influences on growth.

In the late 1970s, economic historians formulated the concept of net nutrition (similar in meaning to nutritional status as used by nutritionists), which can be explained metaphorically by viewing the human body as a biological machine. Our machine operates on food as fuel (composed of protein, fat, micronutrients, and so forth), which it expends at idle (resting in bed), replacing depleted cells, and while fighting infection or engaging in physical activity. Diseases may stunt growth by diverting nutritional intake to mobilize the immune system that combatted infection or by causing incomplete absorption of food that is eaten. Similarly, arduous physical activity or work places a substantial claim on the diet, which makes it possible to lose weight or even starve on 3000 calories per day. For these reasons, average height reflects a population's history of *net* nutrition; growth occurs only if enough fuel is available after other expenditures, needed for survival, have been met. This is an adaptive mechanism that humans found beneficial for the survival of children, who have small stores of fat and muscle that can be converted into energy during dietary crises. If better times follow a period of deprivation, growth may exceed that ordinarily found under good conditions. Catch-up (or compensatory) growth is an adaptive biological mechanism that complicates the study of child health using adult height, because it can partially or completely erase the effects of deprivation. Between birth and maturity, a person may undergo several episodes of deprivation and recovery, thereby obscuring important fluctuations in the quality of life. Chronically poor net nutrition inevitably results in slow growth and stunting, which the National Center for Health Statistics formally defines as falling below the 5th percentile of modern height standards.

Figure 1 illustrates relationships involving stature. The arrows show that human growth is directly affected by diet and work, disease, and nutritional resources required for basal metabolism and replacement of worn-out cells. In turn, these variables respond to a host of environmental conditions, such as income, inequality, public health measures, food prices, technology, and the like, as shown in the figure. Social scientists are also interested in ways that stature influences demographic and socioeconomic outcomes, such as mortality, morbidity, labor productivity, cognitive development, and educational attainment. The figure shows that these variables shape various social outcomes, such as income, inequality, and technology used in the production process. This chapter explains and illustrates the links from socioeconomic conditions to stature via proximate determinants of growth and discusses ways that adult stature affects socioeconomic outcomes such as productivity and longevity.

Diet, disease, and work or physical activity are proximate or immediate determinants of growth that have been widely studied by human biologists. They also study genetic influences on growth, but the distinction between individual- and



Source: Adapted from Steckel (1995)

Fig. 1 A flow diagram showing determinants and consequences of stature. (Source: Adapted from Steckel (1995))

population-level outcomes is crucial. The vast majority of individual differences in stature are genetically or biologically regulated (Silventoinen 2003). Whether growth occurs under poor or good environmental conditions, how tall we become depends heavily on our inheritance from biological parents. At the population level, these individual differences tend to cancel such that average heights reflect primarily environmental conditions. Of course, this raises the question of whether systematic differences in growth potential exist around the world; that is, if populations of different continental ancestry lived under the same environmental conditions for several generations, would their average heights differ? DNA analysis may ultimately prove the most effective way to address the question, but the scientific community has yet to provide this answer.

We do know, however, that considerable convergence in average height has occurred around the globe where economic development has unfolded. In 1950, Japan had the shortest men of any industrial nation (about 160 cm) but young adult men today are approximately 173 cm, or roughly 5 cm below modern American height standards, and may well catch up within a generation (Honda 1997; Mosk 1996; Health Service Bureau 2009). While modest ethnic or ancestral differences in the potential for growth may exist, perhaps as much as a few centimeters, it is safe to say that environmental factors are known to have a much larger effect on average height differences around the world over the past two centuries. For example, in the early mid-nineteenth century, the average heights of young Dutch men were moderately small for Europe (about 164–165 cm), but today they reach 183 cm and are the tallest in the world (Drukker and Tassenaar 1997; Statline 2011).

Early Applications

Economists are familiar with the study of technological diffusion and most of them know the early work in the area by Zvi Griliches, who studied hybrid corn in American agriculture (Griliches 1957). He divided the process of technological change into three parts: (1) discovery of basic knowledge (also called invention); (2) application of this knowledge to the production process (also called innovation); and (3) diffusion of the new production process through an industry (also called spread). His work on hybrid corn mainly concerned diffusion, which he divided into three parts: origin, spread, and ceiling.

We can ask whether Griliches' framework may be useful for understanding the process of technological change in the industry of "research." If so, what are the parallels? In hybrid corn, the invention or discovery of basic knowledge was the method of hybridization, which crossed two inbred (self-pollinated) lines to create a new line that was more productive than either inbred line. In the rise of the new anthropometric history, there was no single, dramatic realization of the biology of human growth, parallel to hybridization, but rather an accumulation of knowledge over many decades, which economic historians realized with the guidance of the auxologist J.M. Tanner.

American Slavery

The first important application of this knowledge by economic historians concerned American slavery, and specifically the question of the average age at which young slave women could have had children, which Richard Steckel addressed in a paper he gave at the April 1977 St. Louis meetings of the Population Association of America (Steckel 1977a) and addressed again in his dissertation on slave fertility completed in the summer of 1977 (Steckel 1977b). The issue had lingered since the antebellum days of the slavery debates, which included claims by abolitionists of "breeding," that slave owners commonly forced women to have children at young ages. Based on plantation records (documents maintained by slave owners) and probate records (documents that often listed plantation assets such as slaves in family groups), and a procedure called the singulate mean, which is calculated using the share of women by age who had ever had children. By this method, the age of slave women at first birth was between 19.8 and 21.6 years, depending upon location and plantation size (Steckel 1977b, Table 33). If the maturation of slaves was substantially retarded due to malnutrition, however, it was possible that the age of menarche was high enough to prevent conception before the late teenage years, a result still consistent with widespread breeding.

Research in human biology shows that physical development occurs in a well-defined order, one feature of which is that on average, menarche occurs 1.0–1.5 years following the peak of the adolescent growth spurt (Frisch and Revelle 1969; Tanner 1966). Fortunately, there is a body of data known as slave manifests from which the

peak of the adolescent growth spurt can be estimated (Wesley 1942). These documents were required by law, beginning in 1808, to identify slaves transported in the coastwise trade to prove that they were not smuggled from Africa. Ship captains prepared duplicate manifests that described each slave by name, age, sex, height, color, and name of the owner or shipper. Steckel assembled a large sample of these records, which shows that the peak was about 13.3 years (Steckel 1977a). Therefore, menarche would have occurred by age 15 (2–3 years earlier than estimates for European populations of the era). Allowing a conservative estimate for the delay from menarche to first birth of 3 years would place the age at first birth of less than 18 years, if slave owners forced slave women to have children as soon as possible. It seems clear that owners in general did not manipulate fertility to this end. Possibly, owners tried to prevent early conceptions because infant mortality rates are relatively high if born to young women, or that unions of young slaves were unstable, resulting in costly labor force discontent. Whatever the explanation, to the extent that slave breeding may have existed, it was atypical.

Steckel (1979) reported that adult slave men were approximately 67.2 in. tall, which was about half an inch smaller than northern whites of the antebellum period as measured by Civil War muster rolls of the Union army. Thus, slaves did experience some nutritional deprivation, but on balance, their diet was reasonably adequate for the work effort and disease load they faced.

Subsequent research showed that children were remarkably malnourished – comparable in stature to the most disadvantaged groups of today's poor developing countries – a phenomenon that can be linked to seasonal growth retardation in utero, early interruption of breastfeeding, and a low protein diet (Steckel 1986a, b). In fact, young slave children were so malnourished that the typical child would have been cause for alarm in a modern pediatrician's office. Such deprivation would have created significant and permanent cognitive deficits among children and adults. Slave owners, however, were more interested in raw physical labor than developing skilled workers. Possibly, slaves with greater cognitive abilities would have been difficult to manage.

To economists, it may seem puzzling that slave children were greatly malnourished, despite the fact that planters owned all of their future expected net earnings. If owners invested in good nutrition for slave children, they would reap the benefits of taller and stronger slaves. On the other hand, there would have been costs in the form of meat protein (pork) and higher child-care costs. Well-nourished children are very active and would have required adult supervision, taking valuable labor away from field work. In sum, it was unprofitable to do so and therefore the children were not well-fed until they began working in the fields around age 10. This preadolescent boost in nutrition was enough to propel the slaves upward through the percentiles of modern height standards, in a process called catch-up or compensatory growth.

Thus, the nutritional circumstances of slaves were more varied and complex than anticipated by historians of the subject. Children had extraordinarily poor health, but adults were reasonably well-fed. The central role of stature in measuring and unraveling one of the important aspects of slave life was important in establishing a niche for anthropometric history.

Diffusion

Mortality

Diffusion of anthropometrics occurred in several directions. Cited by the Nobel Prize Committee in 1993, one of the most influential applications in the diffusion process was research by Robert Fogel on mortality (Fogel 1986; Fogel et al. 1983). Anthropometrics helped clarify explanations for the substantial long-term improvement in life expectancy that began in Europe after the middle of the nineteenth century (Mitchell 1978). The trends have inspired numerous hypotheses about the contributions of man and nature (McKeown 1976). Some scholars have advocated advances in medical technology while others pointed to the possibility of less virulent pathogens or greater immunity through natural selection. It seems clear that personal hygiene, public health, and improved diet were relevant. Through a process of elimination, McKeown assigned the most important role to an improved diet. He argued that changes in virulence or greater immunity through selection were implausible, that improved medical technology was largely irrelevant before the 1930s, and that advances in public health or personal hygiene were trivial before the end of the nineteenth century.

Fogel (1986) investigated McKeown's claims by using height data to measure the independent contribution of nutrition to the mortality decline. Knowing that ·heights are a proxy for net nutrition, he regressed mortality rates on heights and other factors to estimate the strength of the relationship. By applying changes in height observed over the period of declining mortality, he calculated that improved nutrition accounted for approximately 40% of the English mortality decline between 1800 and 1980.

Industrialization

The quality of life during industrialization has long been a staple of debate among scholars. Despite this interest, traditional sources, such as per capita income and real wages, have failed to resolve the debate. Often the raw data series are meager in quality, especially for the crucial early and middle phases of change, and traditional measures also fail to capture several important aspects of the quality of life. They ignore inequality, hours of work, work effort, health and safety while at work, and psychological adjustments to an urban-industrial way of life. In addition, the type and quality of goods available for purchase changed radically from the pre- to the postindustrial period, and it is hard to accurately measure changes in the cost of living over long periods of time.

Of course, stature is far from comprehensive, but everyone would agree that nutritional status is an important aspect of the standard of living. Moreover, the data are available for many countries as far back as the preindustrial period, providing a longer time series than is available for many other types of evidence. In addition, as a measure of nutritional status, stature avoids the consistency problems that distort traditional measures, such as real wages, compared over long periods of time.

The report on nutritional status and industrialization for Sweden was rather optimistic (Sandberg and Steckel 1980). Heights in that country increased with brief interruptions during the nineteenth century, providing a good nutritional foundation for worker productivity that was to follow, and moreover, heights continued to increase during the industrial phase of the late nineteenth century, contradicting claims by Marx and Engles.

Results for England and the United States were more pessimistic. In important work on the UK, Floud and Wachter (1982, 1990) found that the stature of London poor boys declined during the early and middle nineteenth century, suggesting that the poor became worse off during the heart of the Industrial Revolution. In the United States, a half-century height decline began for those born in the early industrial period (c.1830) and coincided with growing occupational disparities in average heights (Steckel and Haurin 1994; Margo and Steckel 1983; Costa and Steckel 1997). Verified in other studies and different data sources, the American pattern has given rise to what John Komlos has called the "early industrial growth puzzle" or in the American context the "antebellum puzzle" (Komlos and Coclanis 1997).

While height data did not resolve debates over the standard of living, they helped to focus thought on its meaning and they did identify episodes of deteriorating quality of life that were missed by conventional measures. Indeed, comparisons of this sort have become a small industry (see Steckel and Floud 1997). For example, the American heights challenged firm beliefs that the quality of life was improving unambiguously after 1830, which has sparked a debate over the aspects of life that were deteriorating, such as greater exposure to disease, higher food prices, and perhaps additional work effort. In England, the economic prosperity of the 1820s to the 1850s must be weighed against urbanization's negative impact on health.

Bodenhorn, Guinnane, and Mroz (2017) challenged the extent and possibly the existence of the antebellum puzzle, claiming it could have been an artifact of sample selection bias created by market wages that were rising relative to military compensation. Because wages increased with stature, in this line of reasoning, taller individuals opted out of the military in favor of civilian employment. Komlos and A'Hearn (2017) and Zimran (forthcoming) dispute the analysis and conclusions by Bodenhorn et al., the former on grounds that military pay during the Civil War rose relative to that in civilian labor markets. They also note that the Roy model they employ does not apply to occupational choice during a conflict in which patriotism rather than wages were paramount. Zimran reaches a similar conclusion by using a two-step semi-parametric sample-selection model to correct patterns in average stature for selection into military service on observable and unobservable characteristics.

These exchanges do not end the debate over the standard of living during industrialization as a general phenomenon, in part because the experiences of other countries must be considered. In this regard, it is useful to consult *Health and Welfare during Industrialization* (Steckel and Floud 1997) which compares the experiences of several countries. In Sweden, for example, industrialization had no adverse impact on health during the late nineteenth century (Sandberg and Steckel 1997), whereas health

declined during industrialization in Japan, Gail Honda argued, because the military absorbed resources that could have gone into public health (Honda 1997).

Inequality

Economists and historians have long-established interests in the fate of different social and economic groups in society. A high, persistent degree of inequality suggests that opportunities for social and economic advancement were limited, which in turn affected the willingness of individuals to save and invest in human or physical capital.

Despite great interest in inequality, estimates are often difficult to acquire and they often pose problems of interpretation. Unlike aggregate data on income, which might be estimated from census data on production, measures of income or wealth inequality must be estimated from data recorded for individuals (or families). Tax records, probate records, and census manuscript schedules have proven useful, but it is difficult to assemble consistent time series from these sources. In addition, cross-section measures of inequality, such as the Gini coefficient, do not measure variation in individual opportunities over time. The latter requires longitudinal data, which are even more difficult to obtain from historical sources.

Of course, height data do not resolve all the problems of measuring inequality in historical settings; they merely add new, useful information. The idea that average height reflects inequality was pursued first by Steckel in the 1970s. His regressions on American slave heights showed that biological inequality varied systematically by region of residence, gender, and color (Steckel 1979). In work eventually published a few years later (Steckel 1983), he demonstrated that a country's average height in the mid-twentieth century was a nonlinear function of average income and a linear function of inequality (measured by the Gini coefficient of household income). Average height is a sensitive barometer of the share of the population that lacks the basic necessities of life and therefore is a useful measure of human welfare. In very poor countries, the health of a large share of the population is significantly constrained by inadequate diet, poor housing, and meager medical care. As income increases, a growing share of the population acquires the basic necessities of life and average height increases. In rich countries, nearly everyone attains their genetic potential for growth unless there is considerable inequality, which limits opportunities for growth among the poor.

Fortified with expectations that height and per capita income were positively related, economic historians were surprised to discover that heights were higher in lower income regions and they declined during the early phase of industrialization of some early industrializers such as the UK and the USA small industry has emerged to explain the pattern using knowledge that height and income measure different aspects of the standard of living (Steckel 1995; Komlos and Coclanis 1997).

Differences in average heights by occupation, region, ethnicity, etc. are a measure of inequality in biological aspects of the standard of living. For example, class differences in stature exceeded 10 cm in the eighteenth century England (Floud

and Wachter 1982), more than twice the difference ever observed for the USA. In the late Colonial Period, heights were virtually identical across occupations (Sokoloff and Villaflor 1982), and in the midnineteenth century, they differed by less than 4 cm (Margo and Steckel 1983). This relationship to inequality measures can be explained by the declining additional nutritional intake as income rises at high levels – the Engel curve and the declining marginal product of nutrition on human growth (Komlos 1989).

Native Americans

Taken as a whole, the depictions of Native American life are arguably the most distorted and contradictory of any large ethnic group in American history. Several centuries ago, accounts of early explorers and missionaries merged with the primitivistic tradition of Western civilization to create an image of the noble savage, who inhabited an ideal landscape and lived in harmony with nature and reason (Berkhofer 1988). During the westward movement of the mid-nineteenth century, whites commonly viewed the Plains tribes as bad Indians – warlike savages who terrorized settlers, stole horses, and practiced barbaric rituals. Near the end of the century, these natives morphed into entertainers who parodied their former lives in Buffalo Bill Cody's wild-west shows. Early in the twentieth century stories in the *Saturday Evening Post* and later, western movies, caricatured their habits, customs, and superstitions (Marsden and Nachbar 1988). Beginning in the 1960s, a wave of publications viewed Native Americans sympathetically, as victims whose peaceful civilizations were decimated by Euro-American aggression and disease. With the rise of the environmental movement, some researchers proclaimed that natives were ecologically sensitive caretakers of their surroundings and therefore models to emulate for our time.

Faced with these diverse images and meager evidence, how can one discern the truth? How did these Native Americans actually conduct their lives and what was their standard of living? Three recent papers probe the latter issue for the equestrian tribes of the Great Plains using height data originally collected by Franz Boas near the end of the nineteenth century (Steckel and Prince 2001; Komlos 2003; Steckel 2010). Using data on individuals born primarily between 1830 and 1872, these papers establish that the equestrian nomads were taller than any national population living in the mid-nineteenth century – exceeding by roughly 1 cm the next tallest groups, which were European descendants living in America or Australia. Explanations center on a protein-rich diet based on bison, abundant micronutrients from various plant sources and from trade with agricultural tribes, relative equality in access to resources within tribes, and low population density accompanied by frequent movements that reduced exposure to parasites and pathogens.

These papers noted large height differences across the tribes. Cheyenne men reached 176.7 cm, about 1 cm below modern American height standards and nearly 9 cm more than the shortest tribe, the Comanche. A difference of this magnitude is large and functionally important, exceeding the increase in average height of native-born American men from the early 1700s to the present.

Fates of Children During Crises

Two strands of literature have emerged on the health of children during crises, the most numerous focusing on developing countries, where considerable height data have been collected since the mid-1980s. In this vein, papers have appeared on Cameroon (Pongou et al. 2006), Kazakhstan (Dangour et al. 2003), North Korea (Schwekendiek 2008b), Novi Sad (Bozić-Krstić et al. 2004), South Africa (Hendriks 2005), and Zimbabwe (Hoddinott 2006; Alderman et al. 2006). The growth of adolescent boys fluctuated seasonally with the supply of food in Czechoslovakia (Cvrcek 2006).

One might expect that seismic shifts or differences in political systems adversely affect the health and welfare of children. The stature of children in Nazi Germany stagnated from 1933 to 1938 following autarchy and market disintegration (Baten and Wagner 2003). North Koreans declined in health relative to the South in the late twentieth century (Pak 2004) and were poorly equipped to deal with a famine that appeared in the 1990s (Schwekendiek 2008b). Delivery of United Nations food aid, however, seems to have improved anthropometric outcomes (Schwekendiek 2008a). Eastern Europe and Russia during the transition in the 1990s were different, whereby life expectancy deteriorated, especially in Russia, but anthropometric data suggest children were protected (Stillman 2006). Apparently adults bore the brunt of declining socioeconomic conditions during and immediately after these regime changes.

Fetal Origins Hypothesis

Careful sifting of epidemiological evidence assembled for the UK in the 1980s led David Barker and colleagues to hypothesize that several adult diseases, such as hypertension, Type 2 diabetes, and some cancers, could be traced to health conditions in a critical formative period of early life, from conception through early childhood (Barker 1994). Although medical experts disagree on the biological mechanism by which health in early life extends into adulthood, the empirical regularity is well established and is sometimes called the fetal origins hypothesis, or the Barker hypothesis (Blackwell et al. 2001). Because early childhood conditions are important for adult height, several studies use stature as a marker to predict longevity and/or adult disease patterns (Christensen et al. 2007; Costa 2004; Jousilahti et al. 2000; Riley 1994; Murray 1997; Harris 1997). Of course, other measures of early childhood health have been used in this line of research, including body mass index (Henriksson et al. 2001; Linares and Su 2005), child mortality rates (Bozzoli et al. 2007), exposure to epidemic disease (Almond 2006; Almond and Mazumder 2005), and skeletal markers of physiological stress (Steckel 2005). Recently economists and other social scientists have investigated ways that early childhood health affects cognition and skill formation, which sheds light on the optimal design of public policy (Heckman 2006). Such policies should note the intergenerational consequences of a mother's reproductive fitness because healthier female children later have healthier babies (Osmani and Sen 2003).

Some economic historians are already putting the health-cognition ideas to good use by linking proxies for physiological stress in early childhood with numeracy, or the ability to recall one's precise age, in this case for census enumerators in nineteenth century Britain (Baten et al. 2008). Although the evidence is circumstantial, it is plausible to believe that grain prices and welfare subsidies under the Poor Law impacted early childhood net nutrition and cognitive development, and thus numeracy rates found among these children as adults. This type of study opens an entirely new dimension to height research for understanding the nutritional underpinnings of human capital acquisition, technological change, innovation, and economic growth.

Colonial Rule

An interesting application of height research concerns the welfare of populations that lived under colonial rule. Over the past decade, papers have appeared on India, North America, and Burma under the British, and Taiwan under Japan (Brennan et al. 1997; Komlos 2001; Morgan and Liu 2007; Olds 2003; Bassino and Coclanis 2008). Various working papers in this general area are underway for other countries. The results suggest that colonialism, often maligned as exploitative, had some positive welfare benefit, at least in these countries. Brennan et al. report a modest increase in the average heights in India under the British but suggest this might be attributed to greater engagement in world markets and expansion of the transportation system. With regard to Taiwan, Olds finds that heights of children improved under Japanese rule up to 1930, which squares with results of Morgan and Liu. Unlike Olds, the latter have evidence through 1945, which indicates that heights after 1930 were static, as they were in Japan. In Burma, the picture is more pessimistic; heights did increase after World War II, but this was apparently a recovery from an earlier height decline induced by land-clearing for rice cultivation in malaria-infested areas.

Komlos assembled evidence on heights from newspaper advertisements on American-born soldiers who deserted and on run-away apprentices (Komlos 2001), reporting a 0.5 in. decline in the first half of the eighteenth century followed by a substantial rise. By the end of the century, American men were as much as 6.6 cm taller than the English, and at age 16, the American apprentices were about 12 cm taller than poor children of London. If the British exploited the American colony, then other factors, such as abundant cheap land and low population density, more than compensated.

Research Frontiers

One important extension of anthropometric methods involves collaboration with bioarchaeologists, who excavate and analyze skeletal remains to gain insights into the quality of life in the past. With training in both medicine and archaeology, these scientists seek to collaborate with economists and historians because they gain knowledge of historical events and processes that are potentially useful in interpreting

health conditions founds on skeletal remains. Through this source, measures of health conditions may extend far into the past, going back to the origins of agriculture and creation of permanent settlements. Moreover, the measures available include not only stature (from long bone lengths) but interruptions of the childhood growth process, trauma, oral health, and infections such as tuberculosis and syphilis.

The Western Hemisphere was the focus of a project completed in 2002, based on the remains of 12,520 individuals, some of whom lived as long ago as 5000 BC (Steckel and Rose 2002). About one-half of the sample lived in pre-Columbian America and approximately one-half lived in North America. One of the most interesting findings in this project was the long-term decline in health prior to the arrival of Columbus, possibly explained by movement of the population to less healthy environments such as urban areas.

A second project of this type focuses on health in Europe since the late Roman era (Steckel et al. 2019). The results are based on the skeletal remains of 15,119 individuals who lived in what are now 17 modern European countries. Surprisingly, there was no substantial trend in health but rather a sequence of changes that were somewhat offsetting. Trauma, for example, declined following the late Middle Ages but oral health deteriorated. Climate (in the form of temperature) and settlement size (i.e., urbanization) had a large impact on health in Europe.

Conclusions

Once a niche subject widely disparaged by skeptics, over the past 40 years anthropometric history has shed new light on important subjects in the social sciences. Its success hinged on blending human biology with historical data to shed new light on questions in which there was already an established interest. Though not without challenges, prospects for the field are bright because historical records are superabundant and many scholars now believe the methodology is credible. Moreover, universities and many individual departments increasingly support interdisciplinary research, of which anthropometric history is a good example.

Cross-References

- ▶ Historical Measures of Economic Output
- ▶ Human Capital
- ▶ Nutrition, the Biological Standard of Living, and Cliometrics

References

Alderman H, Hoddinott J, Kinsey B (2006) Long term consequences of early childhood malnutrition. Oxf Econ Pap 58(3):450–474

Almond D (2006) Is the 1918 influenza pandemic over? Long-term effects of in utero influenza exposure in the post-1940 U.S. population. J Polit Econ 114(4):672–712

Almond D, Mazumder B (2005) The 1918 influenza pandemic and subsequent health outcomes: an analysis of SIPP data. Am Econ Rev 95(2):258–262
Barker DJP (1994) Mothers, babies and disease in later life. BMJ Publishing Group, London
Bassino J-P, Coclanis PA (2008) Economic transformation and biological welfare in colonial Burma: regional differentiation in the evolution of average height. Econ Hum Biol 6(2):212–227
Baten J, Wagner A (2003) Autarky, market disintegration, and health: the mortality and nutritional crisis in Nazi Germany, 1933–37. Econ Hum Biol 1(1):1–28
Baten J, Crayen D, Voth H-J (2008) Poor, Hungary and stupid: numeracy and the poor law in Britain. The Review of Economics and Statistics 96(3):418–430
Berkhofer RF Jr (1988) White conceptions of Indians. In: Washburn WE (ed) History of Indian-white relations, Vol. 4 of Handbook of North American Indians. Smithsonian, Washington, DC, pp 522–547
Blackwell DL, Hayward MD, Crimmins EM (2001) Does childhood health affect chronic morbidity in later life? Soc Sci Med 52(8):1269–1284
Boaz F (1898) The growth of Toronto children, ed U.S. Commissioner of Education. Government Printing Office, Washington, DC
Bodenhorn H, Guinnane T, Mroz T (2017) Sample-selection biases and the industrialization puzzle. J Econ Hist 77(1):171–207. https://doi.org/10.1017/S0022050717000031
Bowdotch HP (1877) The growth of children. Albert J. Wright, Boston
Bozić-Krstić VS, Pavlica TM, Rakić RS (2004) Body height and weight of children in Novi Sad. Ann Hum Biol 31(3):356–363
Bozzoli C, Deaton AS, Quintana-Domeque C (2007) Child mortality, income and adult height. In NBER working paper No. 12966, Cambridge
Brennan L, McDonald J, Shlomowitz R (1997) Towards an anthropometric history of Indians under British rule. Res Econ Hist 17:185–246
Cameron NP, Bogin B (2012) Human growth and development. Elsevier Science, Burlington
Chadwick E (1842) Report to Her Majesty's principal secretary of state for the Home Department, from the Poor Law Commissioners, on an inquiry into the sanitary condition of the labouring population of Great Britain. Craig Thornber, London
Christensen TL, Djurhuus CB, Clayton P, Christiansen JS (2007) An evaluation of the relationship between adult height and health-related quality of life in the general UK population. Clin Endocrinol 67(3):407–412
Costa DL (2004) The measure of man and older age mortality: evidence from the Gould sample. J Econ Hist 64(1):1–23
Costa DL, Steckel RH (1997) Long-term trends in health, welfare, and economic growth in the United States. In: Steckel RH, Floud R (eds) Health and welfare during industrialization. University of Chicago Press, Chicago, pp 47–89
Cvrcek T (2006) Seasonal anthropometric cycles in a command economy: the case of Czechoslovakia, 1946–1966. Econ Hum Biol 4(3):317–341
Dangour AD, Farmer A, Hill HL, S Ismail J (2003) Anthropometric status of Kazakh children in the 1990s. Econ Hum Biol 1(1):43–53
Drukker JW, Tassenaar V (1997) Paradoxes of modernization and material well-being in the Netherlands during the nineteenth century. In: Steckel RH, Floud R (eds) Health and welfare during industrialization. University of Chicago Press, Chicago, pp 331–377
Eveleth PB, Tanner JM (1976, 1990) Worldwide variation in human growth. Cambridge University Press, Cambridge
Floud R, Wachter KW (1982) Poverty and physical stature: evidence on the standard of living of London boys 1770–1870. Soc Sci Hist 6(4):422–452. https://doi.org/10.2307/1170971
Floud R, Wachter KW (1990) Height, health and history: nutritional status in the United Kingdom, 1750–1980, Cambridge studies in population, economy, and society in past time. Cambridge University Press, Cambridge/New York
Fogel RW (1986) Nutrition and the decline in mortality since 1700: some preliminary findings. In: Engerman SL, Gallman RE (eds) Long-term factors in American economic growth. University of Chicago Press, Chicago, pp 439–527

Fogel RW, Engerman SL, Floud R, Friedman G, Margo RA, Sokoloff K, Steckel RH, James Trussell T, Villaflor G, Wachter KW (1983) Secular changes in American and British stature and nutrition. J Interdiscip Hist 14(2):445–481. https://doi.org/10.2307/203716
Francis RC (2012) Epigenetics: how environment shapes our genes. W. W. Norton, New York
Frisch RE, Revelle R (1969) The height and weight of adolescent boys and girls at the time of the peak velocity of growth in height and weight: longitudinal data. Hum Biol 41(4):536–569
Gould SJ (1996) The mismeasure of man, Rev. and expanded edn. Norton, New York
Griliches Z (1957) Hybrid corn: an exploration in the economics of technological change. Econometrica 25(4):501–522
Harris B (1997) Growing taller, living longer? Anthropometric history and the future of old age. Ageing Soc 17(5):491–512
Health Service Bureau (2009) National Health and Nutrition Survey in Japan. Nutrition survey report 2007. Ministry of Health, Labour and Welfare, Tokyo
Heckman JJ (2006) Skill formation and the economics of investing in disadvantaged children. Science 312(30):1900–1902
Hendriks S (2005) The challenges facing empirical estimation of household food (in)security in South Africa. Dev South Afr 22(1):103–123
Henriksson KM, Lindblad U, Agren B, Nilsson-Ehle P, Rastam L (2001) Associations between body height, body composition and cholesterol levels in middle-aged men. The coronary risk factor study in southern Sweden (CRISS). Eur J Epidemiol 17(6):521–526
Hoddinott J (2006) Shocks and their consequences across and within households in rural Zimbabwe. J Dev Stud 42(2):301–321
Honda G (1997) Differential structure, differential health: industrialization in Japan, 1868–1940. In: Steckel RH, Floud R (eds) Health and welfare during industrialization. University of Chicago Press, Chicago, pp 251–284
Jousilahti P, Tuomilehto J, Vartiainen E, Eriksson J, Puska P (2000) Relation of adult height to cause-specific and total mortality: a prospective follow-up study of 31, 199 middle-aged men and women in Finland. Am J Epidemiol 151(11):1112–1120
Komlos J (1989) Nutrition and economic development in the eighteenth-century Habsburg monarchy: an anthropometric history. Princeton University Press, Princeton
Komlos J (2001) On the biological standard of living of eighteenth-century Americans: taller, richer, healthier. Res Econ Hist 20:223–248
Komlos J (2003) Access to food and the biological standard of living: perspectives on the nutritional status of native Americans. Am Econ Rev 93(1):252–255
Komlos and A'Hearn (2017) Clarifications on a Puzzle: the Decline in Nutritional Status at the Onset of Modern Economic Growth in the U.S.A., University of Munich working paper Munich, Germany
Komlos J, Coclanis P (1997) On the puzzling cycle in the biological standard of living: the case of antebellum Georgia. Explor Econ Hist 34(4):433–459
Launer J (2016) Epigenetics for dummies. Postgrad Med J 92(1085):183–184. https://doi.org/10.1136/postgradmedj-2016-133993
Linares C, Su D (2005) Body mass index and health among union Army veterans: 1891–1905. Econ Hum Biol 3(3):367–387
Margo RA, Steckel RH (1983) Heights of native-born whites during the antebellum period. J Econ Hist 43(1):167–174
Marsden MT, Nachbar J (1988) The Indian in the movies. In: Washburn WE (ed) History of Indian-white relations, Vol. 4 of Handbook of North American Indians. Smithsonian, Washington, DC, pp 607–616
McKeown T (1976) The modern rise of population. Arnold, London
Mitchell BR (1978) European historical statistics 1750–1970. Macmillan, London
Morgan SL, Liu S (2007) Was Japanese colonialism good for the welfare of Taiwanese? Stature and the standard of living. China Q 192:990–1013
Mosk C (1996) Making health work: human growth in modern Japan. University of California Press, Berkeley

Murray JE (1997) Standards of the present for people of the past: height, weight, and mortality among men of Amherst College, 1834–1949. J Econ Hist 57(3):585–606
Olds KB (2003) The biological standard of living in Taiwan under Japanese occupation. Econ Hum Biol 1(2):187–206
Osmani S, Sen A (2003) The hidden penalties of gender inequality: fetal origins of ill-health. Econ Hum Biol 1(1):105–121
Pak S (2004) The biological standard of living in the two Koreas. Econ Hum Biol 2(3):511–521
Pongou R, Salomon J, Majid E (2006) Health impacts of macroeconomic crises and policies: determinants of variation in childhood malnutrition trends in Cameroon. Int J Epidemiol 35(3):648–656
Quetelet A (1835) Sur l'homme et le développement de ses facultés: ou, Essai de physique sociale, Monograph, vol 1. Bachelier, Paris
Riley JC (1994) Height, nutrition, and mortality risk reconsidered. J Interdiscip Hist 24(3):465–492
Roberts C (1876) The physical development and proportions of the human body. St George's Hosp Rep 8:1–48
Roberts C (1878) A manuel of anthropometry. J. and A. Churchill, London
Sandberg LG, Steckel RH (1980) Soldier, soldier, what made you grow so tall? A study of height, health and nutrition in Sweden, 1720–1881. Econ Hist (Sweden) 23(2):91–105
Sandberg L, Steckel RH (1997) Was industrialization hazardous to your health? Not in Sweden! In: Steckel RH, Floud R (eds) Health and welfare during industrialization. University of Chicago Press, Chicago, pp 127–160
Schwekendiek D (2008a) Determinants of well-being in North Korea: evidence from the post-famine period. Econ Hum Biol 6(3):446–454
Schwekendiek D (2008b) The North Korean standard of living during the famine. Soc Sci Med 66 (3):596–608
Silventoinen K (2003) Determinants of variation in adult body height. J Biosoc Sci 35(2):263–285
Sokoloff KL, Villaflor GC (1982) The early achievement of modern stature in America. Soc Sci Hist 6(4):453–481. https://doi.org/10.2307/1170972
Statline (2011) Reported height. In: Statistiek CBvd, editor. Health, Lifestyle, Use of Medical Facilities. The Hague
Steckel RH (1977a) The estimation of the mean age of female slaves at the time of menarche and their first birth. The 1977 meeting of the Population Association of America, St. Louis, 23 April 1977
Steckel RH (1977b) The economics of U.S. slave and southern white fertility. PhD Economics, University of Chicago
Steckel RH (1979) Slave height profiles from coastwise manifests. Explor Econ Hist 16(4):363–380
Steckel RH (1983) Height and per capita income. Hist Methods 16:1–7
Steckel RH (1986a) A peculiar population: the nutrition, health, and mortality of American slaves from childhood to maturity. J Econ Hist 46:721–741
Steckel RH (1986b) A dreadful childhood: the excess mortality of American slaves. Soc Sci Hist 10:427–465
Steckel RH (1995) Stature and the standard of living. J Econ Lit 33(4):1903–1940
Steckel RH (1998) Strategic ideas in the rise of the new anthropometric history and their implications for interdisciplinary research. J Econ Hist 58(3):803–821
Steckel RH (2005) Young adult mortality following severe physiological stress in childhood: skeletal evidence. Econ Hum Biol 3(2):314–328
Steckel RH (2010) Inequality amidst nutritional abundance: native Americans on the Great Plains. J Econ Hist 70(2):265–286
Steckel RH, Floud R (eds) (1997) Health and welfare during industrialization, a National Bureau of Economic Research project report. University of Chicago Press, Chicago
Steckel RH, Haurin DR (1994) Clarifications on a Puzzle: the Decline in Nutritional Status at the Onset of Modern Economic Growth in the U.S.A. In John Komlos (ed.), Stature, Living Standards, and Economic Development (Chicago: University of Chicago Press, 1994). pp. 117–128

Steckel RH, Prince J (2001) Tallest in the world: Native Americans of the Great Plains in the nineteenth century. Am Econ Rev 91(1):287–294
Steckel RH, Rose JC (eds) (2002) The backbone of history: health and nutrition in the Western hemisphere. Cambridge University Press, New York
Steckel RH, Larsen CS, Roberts C, Baten J (eds) (2019) The backbone of Europe: health, nutrition, work and violence over two millennia. Cambridge University Press, Cambridge
Stillman S (2006) Health and nutrition in Eastern Europe and the former Soviet Union during the decade of transition: a review of the literature. Econ Hum Biol 4(1):104–146
Studenski P (1958) The income of nations; theory, measurement, and analysis: past and present; a study in applied economics and statistics. New York University Press, New York
Tanner JM (1966) Growth at adolescence: with a general consideration of the effects of hereditary and environmental factors upon growth and maturation from birth to maturity, 2nd edn. Blackwell Scientific Publications, Oxford
Tanner JM (1981) A history of the study of human growth. Cambridge University Press, Cambridge
Tanner JM, Preece MA (eds) (1989) The physiology of human growth. Cambridge University Press, Cambridge
Ulijaszek SJ, Johnston FE, Preece MA (1998) The Cambridge encyclopedia of human growth and development. Cambridge University Press, Cambridge/New York
Villermé LR, Golfin H (1829) Mémoire sur la taille de l'homme en France. Martel, Montpellier
Wesley CH (1942) Manifests of slave shipments along the waterways, 1808–1864. J Negro Hist 27(2):155–174
Zimran A (forthcoming) Does sample-selection bias explain the antebellum puzzle? Evidence from military enlistment in the nineteenth-century United States. J Econ Hist

Wealth and Income Inequality in the Long Run of History

Guido Alfani

Contents

Abstract

This article provides an overview of current knowledge about economic inequality, of both income and wealth, in the very long run of history focusing on Western Europe and North America. While most of the data provided by recent research cover the period from the late Middle Ages until today, some insights are also possible into even earlier epochs. Based on these recent findings, economic inequality seems to have been growing over centuries, with phases of clear and marked inequality reduction being relatively rare and usually associated with catastrophic events, such as the Black Death during the fifteenth century or the World Wars in the twentieth. Traditional explanations of long-term inequality growth are found to be unsatisfying, and a range of other possible causal factors are explored (demographic, social-economic, and institutional). Placing today's situation in a very long-run perspective not only leads us to question old

G. Alfani (✉)
Dondena Centre and IGIER, Bocconi University, Milan, Italy
e-mail: guido.alfani@unibocconi.it

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_29

assumptions about the future of inequality (think of current criticism of Kuznets's hypotheses) but also changes how we perceive inequality in the modern world.

Keywords
Wealth inequality · Income inequality · Distribution · Long run · Western history

Introduction

Recent years have seen a flourishing of studies on long-term trends in inequality.[1] Much of this new research has focused on the nineteenth and twentieth centuries, continuing a tendency that had been originally triggered by Simon Kuznets's seminal article published in 1955. However, significant advances have also been made regarding preindustrial times, which for a long period had remained basically uncharted territory for inequality studies. The first article covering long-run inequality trends for a European region (Holland) was published 40 years after Kuznets's opening salvo (Van Zanden 1995). It is only during the last 5 to 10 years, though, that research on preindustrial inequality has really intensified, thanks to a large degree to the activities of the ERC-funded project *EINITE-Economic Inequality across Italy and Europe 1300–1800*.[2] This research has added considerably to the amount of information available to explore the dynamics and the underlying causes of inequality change in the very long run – so much so that today, in some respects and at least for some European areas, we might know more about preindustrial inequality than about the changes in distribution of the last 50 or 60 years. Another important development is that inequality research shifted from an income focus to more attention to the inequality of wealth.

The recent findings about long-run inequality trends, of both income and wealth, have significantly changed how we look both at historical distributive dynamics and current inequality levels and tendencies. Placing today's situation in a very long-run perspective leads us to question old assumptions about the future of inequality – indeed, Kuznets's hypothesis about the existence of an innate tendency for inequality to decline after the achievement of a certain level of development has now been explicitly rejected by many scholars. But more generally, all traditional explanations of long-term inequality growth have been found to be unsatisfying (as they are unable to account for the new *facts* unearthed by economic historians), and consequently, a range of other possible causal factors have been explored. The debate over the causes and to some degree also over the consequences of long-term inequality growth is today particularly intense and sees the active contribution of many economic historians and cliometricians.

This chapter sets out to provide an overview of recent acquisitions on wealth and income inequality in history, focusing on the period for which we have at least some

[1]I would like to thank Peter Lindert for many helpful comments.

[2]www.dondena.unibocconi.it/EINITE

good-quality data (approximately from 1300 until today) but also providing glimpses into inequality levels and dynamics in even more ancient periods (from prehistory to the Classical Age). The second part of the chapter proposes an overview of the debate over different factors that might have contributed to shaping inequality trends in the long run, taking into account economic, social-demographic, and institutional factors. In the conclusion, some reflections are provided about what history can teach us about the future of inequality. For reasons of space, the focus will mostly be on Europe and North America.

How Economic Inequality Has Changed over the Centuries

As Kuznets himself had underlined, progress in our understanding of long-run inequality trends depends strictly upon our ability to collect good-quality data about income and wealth distributions in time. In fact, as the recent wave of research on inequality trends has demonstrated beyond doubt, distributional dynamics are so complex that they could not be simply inferred from other variables (say, trends in per capita GDP) but need to be measured as directly as possible. The nature of such direct evidence changes over time and across space, although usually it comes from fiscal assessments of various kinds. In section "The Medieval and Early Modern Period (from ca. 1300 to 1800)," the sources available to study inequality in preindustrial times will be discussed briefly, while for those that can be used for the nineteenth and twentieth century, I will refer to the excellent recent synthesis by Roine and Waldenström (2015).

The available studies provide us with standard measures of inequality, which almost invariably include the Gini index of concentration. The Gini is calculated using the following formula:

$$G = \frac{2}{n-1} \sum_{i=1}^{n-1} (F_i - Q_i)$$

where n is the number of individuals/households, i is the position of each individual in the ranking sorted by increasing income/wealth, Fi is equal to i/n, and Qi is the sum of the income/wealth of all individuals comprised between position 1 and i divided by the total wealth of all individuals (in other words, $F_i - Q_i$ is the difference between the share of the population up to position i in the wealth distribution moving from bottom to top and their share of the overall wealth). In this formula, the Gini index is standardized to vary between the value of 0, which corresponds to perfect equality (when each individual/household has the same income/wealth, $F_i - Q_i$ is equal to 0 for every i), and 1, which corresponds to perfect inequality (one individual/household earns/owns everything).

The Gini index is the best instrument we have to "summarize" the level of inequality in a given society. However, the same index value can correspond to different distributions. For this reason, it is important to couple it with other

measures that allow us to keep in check important changes in specific parts of the distribution. Usually this is achieved by providing information about specific percentiles of the distribution, for example, the poorest 10%, the richest 10%, and so on. In recent years, the share of the richest part of the population – the top 1%, 5%, or 10% – has become a very popular indicator on its own, partly because it can also be easily understood by those who do not have specific knowledge of inequality statistics and partly because what happens at the top of the distribution seems to be particularly relevant in explaining the overall tendency toward greater or lower concentration of income or wealth (see discussion in Atkinson et al. 2011; Alvaredo et al. 2013). For the sake of simplicity, in the following only the two most popular inequality measures will be used: the Gini index and the share of the richest 10%. Trends in both wealth and income inequality will be explored. For the preindustrial period, this will allow the comparison of more areas, as usually only one of the two has been reconstructed (note that, for most preindustrial societies and lacking more complete information, wealth inequality can also be considered a rough proxy of income inequality, as land was the main source of income for most of the population. On this point, see Lindert 2014; Alfani 2015; Alfani and Ammannati 2017). For the more recent period (1800–today), it will be possible to compare the dynamics in income and wealth inequality for each area covered.

The Medieval and Early Modern Period (From ca. 1300 to 1800)

While our knowledge of levels and trends in preindustrial inequality was extremely limited until recently, we now have good-quality, data-rich reconstructions of long-term trends in (mostly wealth, sometimes income) inequality for many parts of Europe before 1800 (e.g., for Italy, Alfani 2015, 2017; Alfani and Di Tullio 2019; for Spain, Santiago-Caballero 2011; for Portugal, Reis 2017; for the Low Countries, Van Zanden 1995; Ryckbosch 2016; Alfani and Ryckbosch 2016). Some of these reconstructions cover many centuries, with possibly the best case so far being the Florentine State (Tuscany) in Italy, where it has been possible to reconstruct the general inequality trend in the entire period from 1300 to 1800 (Alfani and Ammannati 2017).[3] This broad research campaign reached beyond Europe, as long-term inequality trends in preindustrial times were also explored for Anatolia under the Ottoman Empire (Canbakal 2013), for the prerevolutionary USA (Lindert and Williamson 2016), and for Japan in the late Tokugawa period (Saito 2015).

Most of these studies make use of fiscal sources to reconstruct wealth distributions. In particular for Southern Europe, the sources more commonly used are the property tax records – usually called *estimi* in Italy, *cadastres* in France, and similarly elsewhere – that contain information about the taxable wealth owned by

[3]Other recent research focused on single years when exceptional sources were available for specific areas, for example, Spain in 1759 (Nicolini and Ramos Palencia 2016) or Poland in 1578 (Malinowski and Van Zanden 2017).

each household. This always includes real estate (lands and buildings), which was by far the main component of wealth in preindustrial rural societies, and sometimes other items as well, like capital invested in trade. A limitation of these sources is that they only rarely include the propertyless, which are those households that had no taxable wealth. However, such households are usually very few (3–7% of the total); hence although their exclusion from inequality measurement leads to systematic underestimation of inequality levels, the distortion is very limited,[4] and more importantly, empirically we find that including the propertyless or not does not change the direction of the trend (see Alfani 2015, 2017 and Alfani and Di Tullio 2019 for further discussion).

Overall, the new research allowed the establishment of two fundamental "stylized facts" about preindustrial inequality in Europe in the period 1300 to 1800:

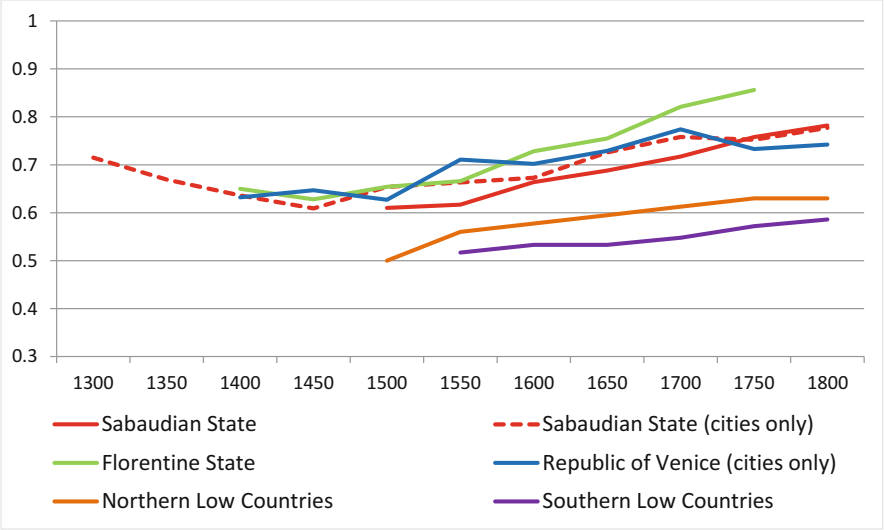
1. During the entire period, the only phase of sustained inequality decline was triggered by the Black Death epidemic, which spread through Europe from 1347 to 1351.
2. After this phase of decline, and beginning from ca. 1450 (with some regional variation), both income and wealth inequality tended to increase almost monotonically in almost all the areas for which we have evidence.

These stylized facts are clearly visible in Fig. 1a and b, which report the Gini index and the share of the richest 10% for some Italian states, the southern Low Countries (current day Belgium) and the northern Low Countries (current day the Netherlands). The measures refer to wealth inequality for Italy and to income inequality for the Low Countries – hence the trends, not the levels, should be compared (as wealth tends to be more concentrated than income, in the past as with today).

The longest series available for an entire state or region refers to the Sabaudian State (Piedmont) in northwestern Italy (cities only). There, before the Black Death the Gini index of wealth concentration equalled 0.715. By 1350, in the immediate aftermath of the Black Death, it had declined to 0.669. Decline continued in the following years, and the absolute minimum value reported for this area in the entire period 1300 to 1800 was reached around 1450, with a Gini of 0.609. After that, inequality growth resumed, continuing without interruption for about two and a half centuries. Indeed, only by the mid-seventeenth century, the pre-plague inequality levels were finally exceeded. Inequality growth stalled in the cities of Piedmont during the first half of the eighteenth century but became intense again in the second

[4]For example, in the city of Bergamo in the Republic of Venice, possibly the Italian community for which we have the most complete information about the prevalence of the propertyless over time, from 1537 to 1702, the distortion to the level of the Gini index varied from a minimum of 0.006 Gini points in 1640 (from 0.715 excluding the propertyless to 0.721 including them) to a maximum of 0.03 Gini points in 1610 (from 0.723 to 0.753) (Alfani and Di Tullio 2019).

a. Gini indexes



b. Share of the top 10%

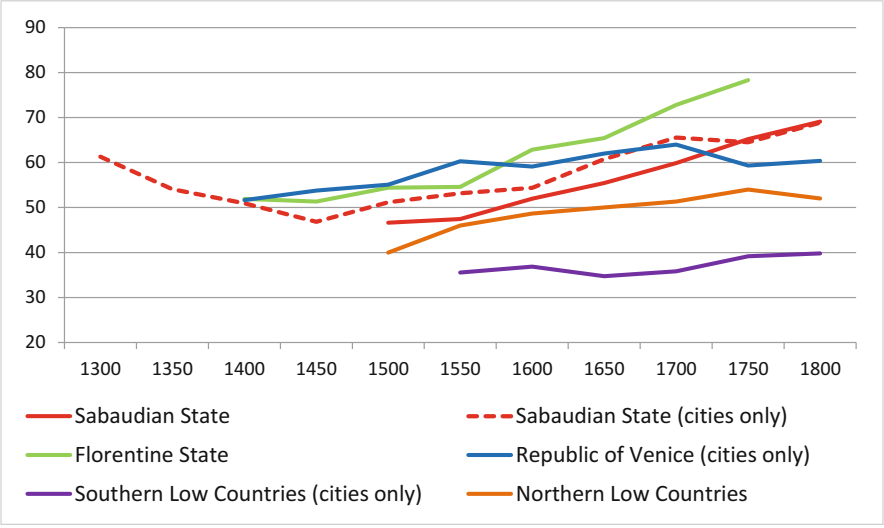


Fig. 1 Long-term trends in economic inequality in Italy and the Low Countries, 1300–1800. Notes: The series refer to wealth inequality for the Sabaudian State, the Florentine State, and the Republic of Venice (excluding those with no property) and to income inequality for the southern and northern Low Countries. (Sources: Alfani (2015) for the Sabaudian State; Alfani and Ammannati (2017) and Alfani and Ryckbosch (2016) for the Florentine State; Alfani and Di Tullio (2019) for the Republic of Venice; Van Zanden (1995) for the Northern Low Countries (Holland); Alfani and Ryckbosch (2016) for the Southern Low Countries)

half of the century, peaking at 0.777 by 1800 (if we look at the entire region, not just cities, inequality growth continued throughout the century). The same path is found looking at the share of the richest 10%, who owned 61.3% of all wealth in 1300, 46.8% in 1450, and 68.9% in 1800, as well as at the other Italian states (Alfani 2015, 2017). During the early modern period, (income) inequality growth is also found in the northern and southern Low Countries.

The distributive consequences of the Black Death are worthy of specific attention. For such an early period, evidence is relatively scarce, and to date it involves mostly the Sabaudian State (Alfani 2015), the Florentine State (Alfani and Ammannati 2017), and the southern Low Countries (Ryckbosch 2016). Although only for the Sabaudian State we have an aggregate series covering the pre- and post-Black Death, for each of these areas, we can observe some specific communities before and after this terrible mortality crisis. For all available cases, inequality declined immediately after the Black Death, with a tendency to continue for about 50 to 100 years, depending on the area. For example, in the city of Prato in Tuscany, the Gini index of wealth inequality was 0.703 in 1325, but by 1372 it had fallen to an all-time low of 0.591 (between the two dates, the share of the richest 10% declined from 65.7% to 48.1%, to the advantage of all other segments of the wealth distribution). Again in Tuscany, in the rural community of Poggibonsi, the Gini index was 0.550 in 1338 but only 0.474 in 1357, after the Black Death (Alfani and Ammannati 2017). Indeed, in the period following the terrible plague, we find the lowest levels of wealth inequality reported for preindustrial Europe – levels that, as will be seen, are not far from those found today. It should also be noted that inequality decline after this, which was probably the most terrible mortality crisis to have ever affected Europe (it killed up to 50% of the population on the continent), is the outcome that should be expected as it goes hand-in-hand with increasing real wages following the sharp reduction in the offer of labor, which contributed to reducing income inequality. Reduction in wealth inequality (and consequently, in capital income inequality) is also to be expected, as higher real wages provided a larger part of the population with the means to acquire property – in a context in which much more real estate than usual was being offered on the market, leading to cheaper prices (Alfani and Murphy 2017, pp. 332–334).

Regarding the dynamics characteristic of the early modern period, the almost monotonic inequality growth reported by almost all available cases continued in the following century, as will be seen in section "The Modern Period (From ca. 1800 Until Today)." Until now, Portugal is the only case for which has been found some evidence of (income) inequality decline during the early modern period, possibly as the consequence of "a long wave of agriculture-based economic expansion during which the demand for labour mostly ran ahead of that for land" (Reis 2017, p. 21). For the rest of Europe, however, a growing body of literature has been discussing the factors leading to inequality *growth*, as will be seen in the second part of this chapter. Before exploring the inequality trends characteristic of the last two centuries, it is useful to underline another stylized fact about preindustrial inequality: the ability of tendencies affecting the share of the top of the distribution (here, the richest 10%) to shape the overall inequality trend as

measured by Gini indexes (just compare Fig. 1a and b). This stylized fact is perfectly reproduced in contemporary societies (Atkinson et al. 2011; Alvaredo et al. 2013), as can be easily seen looking at the evidence provided in the following.

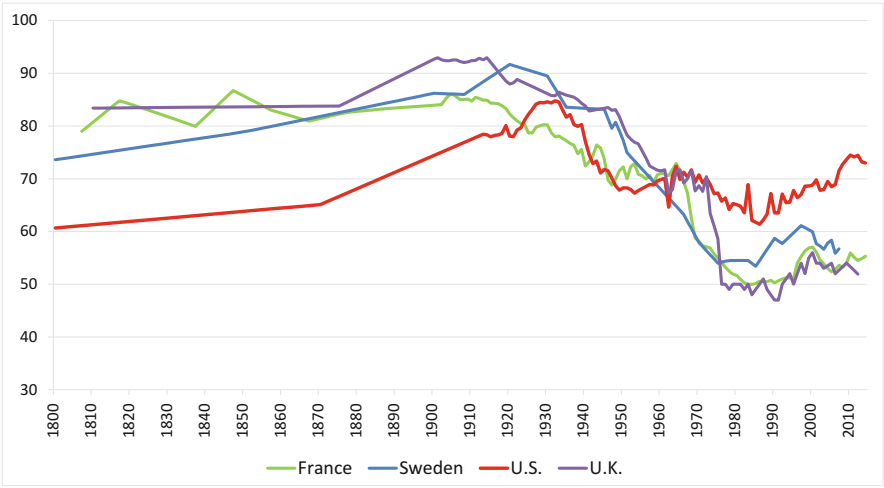
The Modern Period (From ca. 1800 Until Today)

The tendency for an ever-increasing income and wealth concentration that has been reported for the early modern period continued during the nineteenth century, with some signs of growing intensity over time. For this period, too, we have better information about wealth than income. According to all accounts and based on the data currently available, wealth inequality reached its historical maximum on the eve of World War I. For example, in France – probably the country in the world whose wealth distribution during the nineteenth century has been researched most thoroughly (Piketty et al. 2006, 2014; Piketty 2014) – in the immediate prewar years, the share of the richest 10% was about 85%, quite a bit higher (by 5–6 percentage points) than that found in 1807, which is the first year for which we have information. In the same time span, the share of the richest 1% grew even more, by more than 10 percentage points (from 44% to 54–56%).

In the early twentieth century, a continental European country like France was still characterized by a considerably more uneven distribution of wealth than the USA. In 1913, the richest American 10% owned 78.4% of all wealth, slightly less than the richest Frenchmen had owned a century earlier (the same is true looking at the richest 1%, whose share was 45.1% of all wealth in 1913 in the USA). Indeed, in that year the USA was characterized by less uneven wealth distribution than almost all European countries as, for example, the share of the richest 10% was 92.6% in the UK in 1913 and 86% in Sweden in 1908. This relatively egalitarian character of the USA – or more generally, of North America given that in Canada[5] in 1902, the richest 1% had 36.4% of all wealth compared to 53.8% in 1908 Sweden and 66.6% in 1913 UK – has been reported already for the eve of the Revolution. In fact, in 1774 the richest 10% owned "just" 59% of the overall wealth (measured as "net worth"), while the richest 1% had 16.5% (Lindert 2000, p. 188). The share of the richest tended to grow in nineteenth-century America, as the top 10% owned 65.1% of all wealth in 1870 and 78.4% in 1913. Unfortunately no in-between estimates are available, but it seems safe to presume that the overall tendency in the nineteenth-century USA was toward growing wealth concentration (also see Lindert and Williamson 2016, pp. 121–122). If we connect linearly the three dates, we get the tendency shown in Fig. 2a, which also makes it clear how the inequality gap between the USA and France tended to decline over time as a consequence of slower inequality growth in the second. A similar tendency for inequality to grow was found in the UK, where the richest 10% owned 83.4% of all wealth in 1810, 83.8%

[5]The estimate for Canada refers to the Ontario area only. Davis and Di Matteo 2018, p. 34.

a. Wealth share of the top 10%



b. Income share of the top 10%

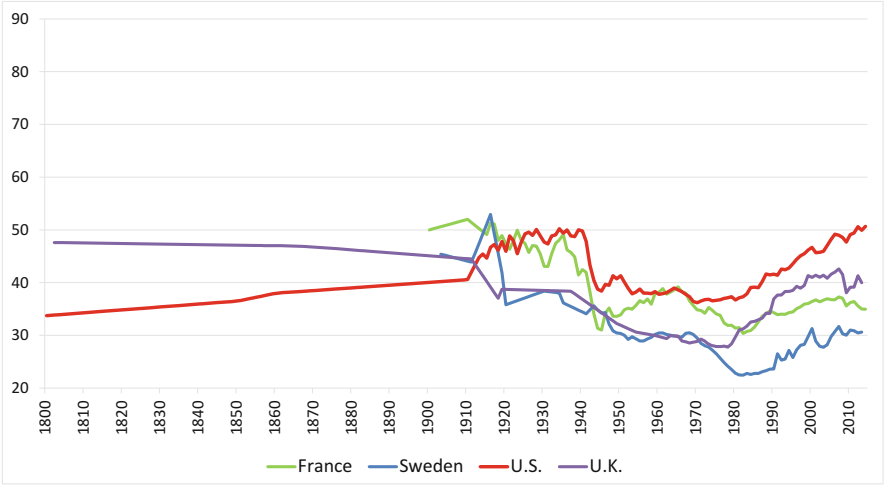


Fig. 2 Long-term trends in economic inequality in Europe and North America, 1800–2014. Notes: Data for wealth refers to households. Data for income refers to fiscal units and is pre-tax. (Sources: Wealth shares: for France, *World Wealth and Income Database* (WID); for Sweden, Bengtsson et al. (2018) for wealth shares in 1800, 1850 and 1900, and Roine and Waldenström (2015) for 1908–2007; for the U.S., Lindert (2000) for 1774, Sutch (2016) for 1870, and WID for 1913–2014 with some integrations from Roine and Waldenström (2015); for the U.K., Lindert (1986) for 1810 and 1875, and WID for 1900–1912 with some integration from Roine and Waldenström (2015). Income shares: for France, WID; for Sweden, WID; for the U.S., Lindert and Williamson (2016) for 1774, 1850, 1860, 1870, and 1910, WID for 1913–2014; for the U.K, Lindert (2000) for 1802, 1867 and 1911 (with some original elaboration), WID for 1918–2013)

in 1875, and 92.6% in 1913. These figures characterize the UK as the country where nineteenth-century wealth concentration was the highest, although admittedly the only other two cases for which we have some information since the early nineteenth century, Sweden and France, lagged just behind. These relative trends can be seen in Fig. 2a, which includes the four countries for which we have some indication of the share of wealth and/or income for the top 10% since the beginning of the nineteenth century (the notes on the figure provide the sources for the estimates discussed above and in the following).

We do not have estimates of the Gini index of wealth inequality for nineteenth-century France and the UK, and more generally, research on the last two centuries has focused more on measuring top wealth and income shares than on reconstructing complete distributions (which are required to calculate Gini indexes). For the USA we have an estimate for 1774, when the Gini equalled 0.694 (free households only). For Sweden, however, a recent study placed the wealth Gini index at 0.79 in 1750, 0.84 in 1800, 0.87 in 1850, and 0.91 in 1900 (Bengtsson et al. 2018). This trend matches perfectly that of the share of the richest 10% reported in Fig. 2a, which across the four dates grew from 68.7% to 73.6%, 78.9%, and finally 86.2% in 1900 – further confirmation of the fact that in most situations, the trend in the top share can be considered indicative of the overall direction of inequality change.

If we look at income, of the four countries included in Fig. 2, only for the USA and the UK do the data available give an impression of the overall trend in the century or so preceding World War I. In the UK, the top 10% concentrated 47.6% of the national income in 1802, 46.9% in 1867, and 44.5% in 1911 (Lindert 2000). In the USA, the share of the top 10% was 32.3% in 1774 (thirteen colonies), 36.5% in 1850, 38% in 1860, 38.5% in 1870, and 40.6% in 1910 (Lindert and Williamson 2016). For a few dates, we also have an income Gini index, which would have equalled 0.459 in the UK in 1759, 0.515 in 1801, and 0.530 in 1867 (Milanovic 2013, p. 12), while in the USA it would have been 0.441 in 1774 (13 colonies) and 0.511 in 1860 and 1870 (Lindert and Williamson 2016). Hence while in the USA the trend was clearly oriented toward inequality growth in income as well as in wealth, in the UK the picture is more mixed – which seems to fit the "Kuznetsian" idea of (income) inequality decline during a late phase of the Industrial Revolution, a process that had been spearheaded by England.

The two World Wars, as well as the troubled period between them, led to a very significant decline in the share of both wealth and income owned by the richest across Western countries. This can also be seen in Fig. 2 and in Table 1. Between 1900 and 1950, the richest 10% released their grip on 12.7% of the overall wealth in the UK, 11.7% in France, 8.9% in Sweden, and 6.1% in the USA. In the same period, across the four countries, the decline in the income share of the top 10% equalled 16.1 percentage points in France, 15 in Sweden, and 13.2 in the UK. Only in the USA do we find a slight increase – by 1.3 percentage points – but here, too, some decline would be found if we took as a comparison the very eve of the USA's entry in the war. This being said, the interwar period remains the moment when the USA finally caught up with continental European levels of wealth and income inequality, turning from being one of the most egalitarian among Western countries to the most *in*egalitarian.

Table 1 Share of wealth and income of the top 10% in Europe and North America, 1850–2010 (data clustered around 50-year breakpoints, with actual year – if different from breakpoint – in parentheses)

	Top 10% (wealth)					Top 10% (income)				
	1850	1900	1950	1980	2010	1850	1900	1950	1980	2010
Denmark	*n.a.*	87.3 (1908)	71.1	67.5 (1975)	70.8	54.1 (1870)	41.9 (1903)	32.2	25.9	26.9
Finland	*n.a.*	70.6 (1909)	61.9 (1967)	63.2	54.4 (2009)	*n.a.*	*n.a.*	*n.a.*	27.7 (1990)	32.5 (1909)
France	86.7 (1847)	84.1 (1902)	72.2	51.6	55.9	*n.a.*	50	33.9	31.4	36.2
Germany	*n.a.*	*n.a.*	*n.a.*	*n.a.*	59.2	*n.a.*	39.1	34.6	31.8	39.7
Italy	*n.a.*	*n.a.*	*n.a.*	40.2 (1989)	45.7	*n.a.*	*n.a.*	*n.a.*	27.2	33.9 (2009)
Netherlands	*n.a.*	*n.a.*	*n.a.*	*n.a.*	50.7	*n.a.*	45.9 (1914)	36.7	28.5 (1981)	30.7
Norway	*n.a.*	76.3 (1912)	78.4 (1948)	58.1 (1979)	51.1	40 (1875)	42.2 (1906)	34.1	25.3	28
Sweden	78.9	86.2	77.3	54.5 (1978)	56.7 (2007)	*n.a.*	45.4 (1903)	30.37	22.8	31
Switzerland	*n.a.*	84.8 (1913)	78.8 (1949)	69.6 (1981)	82.9	*n.a.*	*n.a.*	32.3 (1949)	29.9 (1979)	33.5
UK	83.8 (1875)	92.7	79.9	50	54 (2009)	46.9 (1867)	44.5 (1911)	32.3 (1949)	28.37 (1979)	39.2
Canada	*n.a.*	82.1 (1902)	64.7	56.0 (1984)	57.0 (2012)	*n.a.*	*n.a.*	38.2	37.2	41.4
USA	65.1 (1870)	78.4 (1913)	68.3	65.1	74.5	36.5	40.6 (1910)	41.3	36.7	49.1

Sources: Wealth shares: for France, *World Wealth and Income Database* (WID). For Denmark, Finland, the Netherlands, Norway, and Switzerland, Roine and Waldenström 2015. For Germany in 2010, *Global Wealth Databook* 2017. For Italy, Brandolini et al. 2004 and *Global Wealth Databook* for 2010. For Sweden, Bengtsson et al. (2018) for 1850 and 1900, and Roine and Waldenström (2015) for other dates. For the U.K, Lindert 1986 for 1875, WID for 1900, Roine and Waldenström 2015 for 1950 and 1980, WID for 2009. For the U.S., (Sutch 2016) for 1870 and WID for other dates. For Canada, Davis and Di Matteo (2018), with some original elaboration for 1902 and 1950 (the figure for 1902 refers to the Ontario area only)

Income shares: WID for all countries and dates, exception made for the following. Lindert and Williamson (2016) for the USA in 1850 and 1910. Lindert (2000) for the UK in 1867 and 1911 (with some original elaboration)

Notes: Data for wealth refers to households (exception made for Canada 1950, which refers to adults) and is net wealth (exception made for the USA in 1870 for which only gross wealth estimates are available). Data for income refers to fiscal units and is pre-tax

The reduction in wealth inequality in this period – which also explains an important part of the reduction in overall income inequality, by trimming down disparities in capital income – seems to be due more to the damage suffered by large patrimonies than to a trickling of wealth from the top downward. Indeed, war hyperinflation and the stock market crashes of the 1920s destroyed financial capital; the wars, and especially World War II, significantly affected physical capital; many overseas properties and investments were lost (for a synthesis, see Piketty 2014, pp. 181–187). However after the end of World War II and for 30 years or so, wealth and income inequality remained relatively low, and in many cases, they fell further, at least in Europe. Looking at top income shares, this drop was particularly impressive in Sweden, where from 1950 to 1980, the top 10% lost 7.6 percentage points of the overall national income. In the case of wealth, the largest decline is found in the UK, where the top 10% saw their share drop by 29.9 percentage points. Indeed, by 1980 the UK, which had long been characterized by one of the highest levels of wealth inequality found in Western countries, had become relatively egalitarian, as can also be seen in Table 1, which includes information about the wealth and income share of the top 10% across a broader range of countries.

This overall tendency for a further reduction in income and wealth inequality came to an end from the late 1970s to early 1980s, partly as a consequence of tax reforms (see discussion in section "Institutions"). This is also the point when a difference seems to have emerged in distributive dynamics between English-speaking countries and other Western countries (if we focus on income) or between the USA and European countries (if we focus on wealth, see Fig. 2 and Table 1). In particular, the point has been made that, if we look at the income share of the top 1%, the main English-speaking countries (Australia, Canada, the UK, and the USA) followed a "U-shaped" curve during the twentieth century, with the income share of the top 1% growing quickly after 1980 and more than doubling by 2007, finally reaching or coming close to pre-World War I levels. Instead, continental European countries (plus Japan) followed an "L-shape" path, with much less tendency for income inequality to grow after 1980 (Alvaredo et al. 2013, pp. 5–6). Overall, this difference between continental Europe and English-speaking "Atlantic" countries is confirmed if we look at the income share of the top 10%, as, from 1980 to 2010, we see an increase of 12.4 percentage points in the USA and 10.8 in the UK, more than double what is found in countries like France or Finland, where the increase equalled 4.8 percentage points or Italy where it amounted to 6.7 percentage points. On the other hand, in Canada the increase in the income share of the top 10% was relatively low even by European standards (4.2 percentage points), while from ca. 1990, the OECD country that has been affected by the largest increases in inequality is Sweden. To a significant degree, this is due to (i) the growing importance of capital gains as a component of total income in Sweden, which is consistent with a picture of growing wealth concentration, as well as to (ii) the decreasing progressiveness of the fiscal system (OECD 2015; Waldenström 2009). The latter factor is probably one of the main culprits in the cross-countries increase in *both* income and wealth inequality detected for the post-1980 period. This point will be further discussed in section "How to Explain Inequality Change in the Long Run?" – but before, we need to complete our overview of very long-term trends in economic inequality by moving further back in time.

Glimpses into a More Remote Past: From Prehistory to the Classical Age

The new wave of research on inequality levels and trends in the past has also involved very early epochs. Although the evidence collected is inevitably less abundant and less precise than what we now have for the Middle Ages, it seems important to include here a brief account of recent contributions to our knowledge of economic inequality in the Classical Age and before.

Indeed, some authors have even tried to provide tentative estimates of economic inequality during prehistory. Their work relies upon both archaeological evidence and more detailed information about historical and current "primitive" societies, like the few surviving tribes of hunter-gatherers. Generally speaking, human groups that relied upon foraging were characterized by low levels of economic differentiation. However if we consider early farmers, we can already detect significant increases in economic inequality. There is, in fact, a continuum in inequality levels, from hunter-gatherers, to farmers and beyond: "substantial levels of economic inequality became characteristic of many (but far from all) populations only after the domestication of plants and animals, eventually culminating in the emergence of class societies and the hierarchical ancient states" (Bowles et al. 2010, p. 8). To give an idea of the implied changes in inequality levels, in a sample of contemporary and historical "small-scale societies," the average Gini inequality index of material wealth (including cattle, land, and other physical components of wealth) was 0.36 among hunter-gatherers, rising to 0.51–0.52 for pastoral and horticultural societies and to 0.57 for agricultural societies (Borgerhoff Mulder et al. 2009, Table S4). Wealth concentration, through the inheritance system, was key to reproducing and deepening inequality across generations – indeed, changes in the degree of inheritability of wealth seem to explain much of the variation in inequality levels detected both through prehistory (especially when comparing human societies before and after the so-called Neolithic Revolution, which began around 10,000–8000 BCE and was associated with the appearance of the first permanent settlements and the introduction of agriculture) and when analyzing the conditions experienced by different kinds of small-scale societies today (Borgerhoff Mulder et al. 2009; Bowles et al. 2010). Others underline the importance of a broad variety of "disequalizing factors" potentially leading early societies to grow unequal (as well as more socially hierarchical). Domestication of plants and animals is surely one such factor, along with resource scarcity and technological progress[6] (Scheidel 2017, pp. 33–39).

Further developments in early societies, and particularly the development of governmental institutions and the progressive formation of the first states, enrooted and deepened economic inequality: "premodern states generated unprecedented

[6]For example, there is archaeological proof that among the Chumash tribe living on the shores of California, the introduction of large oceangoing plank canoes able to venture relatively far from the coast, which occurred around 500–700 CE, led to the transition from an egalitarian society of foragers to one dominated by polygamous chiefs who controlled the canoes and organized economic, as well as military and religious activities (Scheidel 2017, pp. 34–35).

opportunities for the accumulation and concentration of material resources in the hands of the few, both by providing a measure of protection for commercial activity and by opening up new sources of personal gain for those most closely associated with the exercise of political power. In the long run, political and material inequality evolved in tandem" (Scheidel 2017, p. 43). As well during this period, the evidence of further inequality growth is mostly archaeological (although some useful written documentation survives) and involves early states like the Akkadian Kingdom from the twenty-fourth to the twenty-second century BCE, Babylon from the second millennium BCE to the sixth century BCE, or Egypt in various epochs (Diamond 1997; Scheidel 2017). It is only when we move a step further, into the Classical Age, whose starting point is conventionally placed around the eighth and the seventh century BCE, that the overall information we have allows for producing some more detailed – although still highly speculative in many respects – estimates of levels of economic inequality. For the sake of synthesis, I will focus on the case of the Roman Empire, which has been the object of a relatively large number of recent publications.

There are currently available two estimates of interpersonal inequality in the Roman Empire built from social tables. One refers to 14 CE, in the early empire period, when the Gini index was placed in the 0.364–0.394 range (Milanovic et al. 2007, p. 77). The second relates to year 150 CE, the apogee of the Roman Empire. By then, income inequality seems to have grown significantly, as the point estimate we have is a Gini of 0.413 (Scheidel and Friesen 2009; Scheidel 2017, p. 78). Both these estimates refer to the Roman Empire as a whole. More recently, an attempt has been made at integrating such measures of interpersonal inequality with new estimates of regional inequality in average income levels, to produce a tentative picture of inequality change across the Empire over a much longer period. Overall, this reconstruction suggests that income inequality grew from 14 CE to 150 CE, which was a period of growing prosperity and territorial expansion. Thereafter, from the high level of about 0.4, income inequality declined very significantly, to a bottom as low as 0.13–0.15 by the year 600 or 700, apparently going hand in hand with the empire's decline (Milanovic 2017, p. 12 and elsewhere).

We know much less about wealth inequality. The available evidence suggests that in this case, too, a first period characterized by increasing wealth concentration is to be found (from the second century BCE to the first century CE, hence covering the late Republican period and the first imperial phase), during which the size of the largest fortunes rose by a factor of 80, from 4–5 million to 300–400 million sesterces (Scheidel 2017, pp. 71–75). This tendency stopped when the Roman Empire began encountering difficulties and then entered a phase of decline. The final collapse of the state, which during the fifth century progressively lost control over its provinces in Europe and the Mediterranean area and in the end even over the core territories of Italy, led to a very substantial reduction in wealth inequality. To a large degree, this was the result of the breakdown of the "extensive networks of estates" that members of the Roman economic elite had owned across the Empire. We have evidence of this process in the abandonment of many country villas, but more importantly, archaeology offers some proof of a more general reduction in

inequality, involving as well the middle and upper-middle strata of society, through the analysis of the distribution of house sizes. In Britain, for example, the Gini index of house sizes fell from a level around 0.6 at the time of the Roman Empire to just 0.4 in the Early Middle Ages (Scheidel 2017, pp. 265–269). According to Walter Scheidel, this is proof of the leveling power of state collapse and more generally of large-scale catastrophes.

How to Explain Inequality Change in the Long Run?

From the data presented above, we get the clear impression that in the long run of history, wealth and income inequality have tended to grow almost continuously. In the last seven centuries – the period for which we have, at least for some world areas, detailed information about distributional dynamics reconstructed from archival or statistical sources or provided by reliable institutions – a significant decline in wealth and income inequality can be detected only after large-scale catastrophes: the Black Death in the fourteenth century and the two World Wars in the twentieth. If we include in the analysis the earlier period, for which we have sparser information, it seems that in the West in the last two millennia, the number of inequality-reducing catastrophes can be increased to three by adding the collapse of the Roman Empire during the fifth century. In all other periods (and allowing for some incertitude for the least-documented phases, in particular the Early Middle Ages), inequality seems to have been increasing almost monotonically. Given these historical dynamics, explaining inequality growth is somewhat trickier than accounting for inequality decline. Indeed, the factors able to determine inequality growth in the long run of history are currently the object of intense debates, which could not be covered in detail here – especially if one allows for the possibility that different forces were acting in different periods, that is, trends that seem similar across space and time might have deeply dissimilar underlying causes. The objective of this session is to provide an overview of these debates – a non-exhaustive overview, surely, but one aimed at making clear possible elements of continuity in the very long run of human history as well as at making explicit connections that research more narrowly focused on specific periods or areas has often failed to underline. As the distributive, inequality-reducing consequences of the main catastrophes have already been outlined in the first part of the chapter, the focus here will be placed on the factors leading inequality to increase during the other periods.

Economic Variables

As will be recalled from the Introduction, much of the research into long-term inequality trends has been triggered by the original contribution of Simon Kuznets (1955). Although there are strong reasons to argue that we should definitely move beyond Kuznets's approach (see below), it still seems useful to begin by briefly recalling his hypothesis, which focuses on economic development as the trigger of

inequality change during the industrial period. According to Kuznets, as well as to scholars who tweaked his analysis in the decades immediately following the publication of his seminal article (for a synthesis, see Brenner et al. 1991), income inequality followed an inverted-U path through the industrialization process (the so-called Kuznets Curve), with a rising phase at the beginning of industrialization. This path would be the consequence of economic development and particularly of the transfer of the workforce from a traditional (agrarian) sector to an advanced (industrial) one. Western countries would have experienced inequality growth during the late eighteenth and nineteenth centuries, and inequality decline from a certain point during the twentieth century. Kuznets's hypothesis referred to income inequality; however it stands to reason that it can also be applied to wealth (Lindert 1991, 2014). Indeed, some studies of Western areas reported an inverted-U path in wealth inequality during the Industrial Revolution, with a decline after the two World Wars. This is also clearly visible in Fig. 2a.

If Kuznets's argument still stands as a description of the historical path followed by inequality in Western countries, its broader implications have now been proven wrong, both if we look at the left side (the long preindustrial period) and at the right side (from ca. 1980 until today) of the Kuznets curve. Regarding preindustrial times, Kuznets seemed to imply that before ca. 1800 or 1750 at the earliest, inequality was relatively low and stable over time. However, as seen in section "How Economic Inequality Has Changed Over the Centuries," this was not the case. The first attempt at measuring inequality growth during the early modern period, involving Holland from the sixteenth to the nineteenth centuries, found evidence of continuous growth in income inequality (Van Zanden 1995; Soltow and Van Zanden 1998. Also see Fig. 1a, "Northern Low Countries"). Jan Luiten Van Zanden provided an interpretation of this phenomenon that was "Kuznetsian" in character. He argued that preindustrial inequality growth was even "over-explained" by economic growth and proposed different explanations for why economic growth could foster inequality growth: (i) urbanization (hence, transfer of workforce from a rural/backward sector to an urban/relatively advanced one), (ii) increasing skill premium, and (iii) changes in the functional distribution of income. He then argued that preindustrial inequality growth constituted only the first phase of a "super-Kuznets curve" that connected preindustrial and industrial economic growth. This interpretation was grounded in the specific context of the Dutch Republic, one of the most economically dynamic areas of early modern Europe. Recently, a similar view has been expressed by Bas Van Bavel (2016, pp. 192–193), who has connected inequality increase in this part of Europe to the development of market economies, which might have led to growth in inequality of both income and wealth through increases in the efficient scale of trade and production, growing opportunities for financial dealing and speculation, and growing investment opportunities (favoring the elites) in landed property and in shares of the public debt.

It is surely possible, even probable, that the inequality growth that has characterized the Dutch Republic during the early modern period was, at least in part, the consequence of economic growth. However, this argument could not be generalized to the rest of Europe, where income and wealth inequality growth were also found in

periods of economic stagnation, or decline.[7] This was, for example, the case in many Italian states (Alfani 2015; Alfani and Ammannati 2017; Alfani and Di Tullio 2019) as well as in the southern Low Countries (Alfani and Ryckbosch 2016), and, as will be seen, this is why much recent literature has been looking in other directions for an explanation of preindustrial inequality growth.[8]

If we now consider the right side of the Kuznets curve, it must be noted that the implication of the existence of an automatic mechanism leading to inequality decline after certain levels of development were reached involved the promise that the system itself would have provided a solution to the problems (social and other) caused by growing economic inequality in the early phases of the industrialization process. This is one of the fundamental reasons why, in Kuznets's view as well as in that of other scholars who recently associated inequality growth with economic development and welfare improvements (e.g., Deaton 2013), growing income concentration is seen as a relatively "benign" process, as it can be understood as a side effect of increasing prosperity and possibly just a temporary one. This view, however, is challenged by what we now know about long-term inequality trends. As seen above, the idea that increases in economic inequality are simply the result of economic growth is challenged quite directly by the historical experience of preindustrial Europe. The idea that inequality decline will follow development, instead, is challenged by what we have seen happening in recent times: "the recent increase [in income inequality] since around 1980… does not fit the [Kuznetsian] predicted earnings dynamics within the distribution. As an increasing number become skilled, the difference within the top should decrease, not increase as seems to be the case" (Roine and Waldenström 2015, p. 552). Moreover, the decline in income inequality in the first half of the twentieth century also does not easily fit the Kuznetsian paradigm, driven as it was, to a large degree at least, by the dynamics of capital income and not of labor income (Piketty 2006, 2014). From all the above considerations, the conclusion can be reached that "there is no mechanical relationship between inequality and industrialization or technological change. It is no more unavoidable that inequality increases in early stages of introducing new technology, than it is automatic that inequality eventually goes down" (Roine and Waldenström 2015, p. 552). For these and other reasons (but mostly simply due to the distributional trends that can be detected for the twentieth century), many scholars have argued that we should move definitely beyond the idea of the Kuznets curve – Peter Lindert, for example, has explicitly declared it to have now become obsolete (Lindert 2000, 2014).

It seems probable, however, that Kuznetsian arguments will continue to be present in the debate, at least for a while. Recently, Branko Milanovic (2016) offered an original perspective. He tried to reconcile the idea of the Kuznets curve with the

[7]This is also confirmed by comparing the available estimates of per capita GDP with the estimates of economic inequality in time: Alfani and Ryckbosch 2016; Alfani and Di Tullio 2019.

[8]Note that until now, the only European area for which we have evidence of a correlation between early modern economic stagnation and income inequality decline is Portugal (Reis 2017).

evidence now available about both the preindustrial period and the developments of the last 50 years or so. Milanovic argues that across history, we can detect a series of "Kuznets waves," that is, alternating phases of rising and then declining inequality, drawing a sequence of "inverted-U" curves. For example, an early wave would begin with pre-Black Death medieval inequality growth, followed by inequality decline only after that terrible pandemic. A second wave would have its ascending phase during the early modern period and would continue during industrialization, with a phase of declining inequality from 1914 onward. Finally, from the late 1970s, a third wave would begin, of which we are currently experiencing the ascending phase. According to Milanovic, this final phase of increasing inequality will also one day reverse, generating its own inverted-U. Leaving apart this forecast, which by definition is not the object of economic history (although history does remind us that earlier promises of inequality reduction have been disproved by real distributive dynamics),[9] there is reason to doubt whether we should label these movements, which are truly wave-like, "Kuznets waves" – as the factors leading to the swings in inequality during the preindustrial period are generally very different from those imagined by Kuznets.

An appealing characteristic of Kuznets's approach is its simplicity as well as its (at least in principle) generalizability. This characteristic is shared by Thomas Piketty's recent attempt at explaining inequality dynamics from the nineteenth century until today (Piketty 2014; Piketty and Zucman 2014). Piketty focused on wealth/income ratios as predictors of income inequality (the higher the ratios, the higher the expected income inequality). Moreover, he argued that as long as the rate of return to capital (r) is higher than the growth rate of national income (g) and as long as wealth stays highly inheritable, inequality (of *both* income and wealth) will continue to increase. Piketty argued that the relative dynamics of these two variables would reflect quite well the tendencies reconstructed by economic historians for the last two centuries, and made the forecast that without policy interventions, inequality will probably tend to grow indefinitely in the coming decades. Indeed, Piketty also provided some hints that his simple law would explain preindustrial distributive dynamics, too – as in a fairly speculative section of his now-famous book, *Capital in the Twenty-First Century*, he argued for r being constantly $>$ g across the world and throughout the period from year 0 to the eve of World War I (Piketty 2014, pp. 445–451). However, there are a number of problems when trying to apply Piketty's views to preindustrial times (note that admittedly, the preindustrial period *was not* Piketty's main focus). First, there is a problem with the supporting empirical evidence, as the available information about preindustrial growth rates of national incomes (g) is still highly hypothetical, while in the case of the return to capital (r), to my knowledge it has never been the object of systematic comparative studies for

[9]Admittedly, Milanovic's "optimism" only goes as far as hypothesizing that the current bout of inequality growth will one day peak, at a level lower than that reached in the early twentieth century due to the presence of "inequality stabilizers" like state pensions and unemployment benefits and will subsequently go down. That day, though, might be a long way off.

preindustrial economies. The most important problem, though, is that the view of a constantly growing inequality does not fit the empirical finding that the Black Death caused a century-long phase of significant inequality decline. Indeed, as this catastrophe destroyed human capital and only marginally physical and financial capital, the wealth/income ratio inevitably grew, which does not support, at least for this specific period, the view that such a ratio would be positively correlated to income and wealth inequality.

Piketty's ideas about the drivers of inequality growth have also been criticized for the nineteenth and twentieth century (and beyond) – indeed, as argued by Lindert, Piketty's arguments seem to encounter stronger support in data for the period 1810 to 1914, as for later periods "across countries, the levels and movements of the [wealth/income] ratio do not correlate well with those in income inequality" (Lindert 2014, p. 8). More generally, Piketty has been criticized for not defining in an entirely clear and satisfying way his concepts and the nature of the variables he uses, as well as for possible faults in his theory (see, e.g., Blum and Durlauf 2015).

It is beyond the aims of this chapter to discuss further the ample debate about Piketty's views – which, in the measure they are grounded in the large-scale (and very laudable) campaign of collection of new information about income and wealth inequality connected to the building of the *World Top Incomes Database* (now *World Wealth and Income Database*[10]), will surely continue to attract considerable attention by economic historians and cliometricians in the upcoming years. What, however, the apparent inability of both Kuznets's and Piketty's theories to fully explain the complexities of historical inequality dynamics seems to teach us is that probably we should look for complex and more case-specific explanations instead of trying to devise simple universal "laws." Again in the words of Lindert (and of Tony Atkinson before him), "[scholars] should invest in an eclectic approach that finds different causes for movements in different epochs" (Lindert 2000, p. 200) – a task that, it seems, economic historians are particularly well-equipped to accomplish.

Demography and Society

Among the many tentative explanations for long-term inequality growth, some have to do with demographic factors or with radical changes in social and social-economic structures. Explanations of this kind are particularly frequent in works focused on the early modern period, as many studies have hinted at a general connection between population growth and inequality growth, especially in cities (Van Zanden 1995; Alfani 2015; Ryckbosch 2016). The point here is not that cities were more unequal than villages and larger cities more unequal than smaller cities – a finding fairly well established in the literature (see for a synthesis Alfani and Ammannati 2017, pp. 1084–1085). It is instead the ability of population growth in a specific setting (a community urban or rural or a broader aggregate such as a region

[10] http://wid.world/data/

or state) to promote inequality growth within the setting that should be evaluated. However, looking at the larger possible aggregates – entire states – and on the basis of the available reconstructions of long-term inequality trends, it has recently been demonstrated that there is no automatic connection between population growth and inequality growth (Alfani and Ryckbosch 2016; Alfani and Di Tullio 2019). For example, in the Sabaudian State, the population stagnated during the seventeenth century – but wealth inequality continued its monotonic growth (Fig. 1a and Alfani 2015). Moreover, when large-scale mortality crises affected early modern populations (like at the time of the terrible plagues affecting Italy and other southern European areas, e.g., that of 1629–1631), they failed to cause significant inequality decline (Alfani and Murphy 2017; Alfani and Di Tullio 2019). As seen in section "How Economic Inequality Has Changed Over the Centuries," the fourteenth-century Black Death certainly did reduce inequality – but this was the result of its broader consequences on the land and labor market, mediated through a specific institutional setting in which a key role was played by the presence of unmitigated partible inheritance systems (instead, by when the seventeenth-century plagues took place, institutional adaptation affecting inheritance systems had occurred, aimed at protecting the largest patrimonies from undesired redistribution. See discussion in Alfani and Murphy 2017; Alfani and Di Tullio 2019).

More generally, it is not obvious why population change at the level of broader aggregates should affect inequality. One possibility is that this happens because demographic growth is positively correlated with economic growth – but this simply leads us back to the criticism of economic growth as the cause of inequality growth. Another possibility is that demographic growth is one of the causes of the waves of "proletarianization" that affected early modern Europe (see below). The trigger, in this case, would be population pressure on resources – which might explain why in the context of early modern Europe only population growth, and not population decline, is positively correlated to inequality. However, the reasons why demographic factors were able to shape inequality trends are more easily observed on a smaller scale: that of single communities, particularly cities. In fact, as early modern cities had a negative natural change (i.e., more individuals died in the city than were born there), growth in urban populations was possible exclusively through significant immigration from rural areas. Some micro-studies have shown how immigration acted as a kind of perpetual generator of inequality for cities, a process that became more intense after severe mortality crises, and – importantly – one that could occur even in the absence of economic growth, for example, simply because physical space to live in had opened within the city walls following a mortality crisis (Alfani 2010), as well as in the absence of growth in urban population/urbanization rates, as would be the case of a severe epidemic affecting cities but sparing rural communities. Note that these dynamics could fuel inequality of both wealth (as the newcomers tended to belong to the surplus rural population and usually had no property) and income (as the newcomers usually belonged to the low-skills workforce and consequently contributed to increasing the wage premium within the urban economy, as argued, e.g., by Van Zanden 1995). However, the evidence of community-level demographic growth actually explaining inequality growth at the *local* level in

preindustrial settings is currently inconclusive, as shown by econometric analyses of inequality in communities of the southern Low Countries from 1500 to 1900 (Ryckbosch 2016) and of the Florentine State from 1300 to 1800 (Alfani and Ammannati 2017).

The last, and in some regard the most important, demographic factor that has been considered a possible determinant of inequality change is urbanization. One reason for this is that, again, urbanization rates are often considered good indicators of economic growth while having some advantages over other indicators: they are easier to measure with actual direct archival data, and they are often available at a regional and subregional level. But urbanization rates are also relevant to assess the potential impact of "Kuznetsian" dynamics in preindustrial Europe. In fact, as argued by Van Zanden (1995, pp. 655–656), if we replace Kuznets's (1955) original distinction between an "industrial" and an "agrarian" sector characterized by different wage levels with that between an "urban" and a "rural" sector, also characterized by a steep wage differential (and there is little doubt that in early modern Europe, average wages were higher in the city than in the country), then "The gradual urbanization that typified . . . [the early modern] period probably contributed to a rise in income inequality through the mechanism described by Kuznets" (Van Zanden 1995, p. 656), i.e., as the simple consequence of the transfer of workforce from one sector to another (through migration from the country to the city), which is testified to by changes in urbanization rates. However, as for population at the state level, recent research could not find any clear correlation between changes in urbanization rates and inequality trends (Alfani and Ryckbosch 2016; Alfani and Di Tullio 2019).

As mentioned above, a mechanism through which population growth might have led to inequality growth is by leading to acute pressure on available resources thus triggering "proletarianization" – that is, the historical process leading a growing share of the European population to lose the ownership of the means of production, thus becoming dependent on selling their labor for wages. This view, which is clearly rooted in the Marxist tradition of economic history, in relatively recent times, has been strongly argued for by Tilly (1984). Many specific historical processes have been presented as a component of this overall tendency toward proletarianization, from the rural enclosures movement to the spread of the putting-out system. Its main aspect, though, is the crisis of small land ownership. If we focus on the areas for which we have reconstructions of long-term inequality, we find, for Italy, an ample literature detailing the crisis of small peasant property with the subsequent concentration of wealth in many areas, especially from the second half of the sixteenth century when population pressure on available resources became acute. Specifically, the crisis of peasant property has been singled out as a possible factor contributing to inequality growth in the Sabaudian State (Alfani 2015), while for the Republic of Venice, we have some evidence of proletarianization in the growing share of propertyless households throughout the early modern period (Alfani and Di Tullio 2019). Similar processes have been detected, too, in the southern Low Countries and in the Dutch Republic (Ryckbosch 2016; Alfani and Ryckbosch 2016). Differently from other possible explanatory factors like economic

growth or urbanization increase, proletarianization was a general, pan-European phenomenon (Tilly 1984, pp. 26–36; Van Zanden 1995, pp. 656–658; Alfani and Ryckbosch 2016), and consequently, it seems – at the very least – an important contributing cause to explaining inequality trends that were largely similar across the continent. However, as proletarianization was connected to population pressure and was triggered by acute phases of scarcity (especially continental-level famines), it tended to come in waves. This is why, although it was surely an important inequality-promoting factor, proletarianization seems to fail to fully account for a process that is found to be overall monotonic in practically all areas from 1500 to 1800. This is a reason to look for possible causes of inequality growth that exerted a more constant influence throughout the early modern period: as will be seen, institutional change affecting fiscal systems satisfies this requirement.

Demographic and socioeconomic-demographic explanations of inequality change are much less frequent when we move from the preindustrial period to the nineteenth and twentieth century. Surely, one reason for this is that economic variables can be observed more directly – hence there is no need to recur to such imperfect measures of economic development as population density and urbanization rates. Demography, however, is obviously part of those explanations that look at the growth rates of the workforce – or at its relative trends in different sectors – as the source of changes in income differentials between different components of the workforce. For example, in the UK and the USA during the nineteenth century, the demographic transition, which is associated with historically very high rates of population growth, promoted income inequality growth by widening the wage differential between skilled and unskilled labor, also due to the inability of the educational system of the time to allow for significant improvements in skills per worker. Later, when the demographic transition entered its second phase (characterized by shrinking fertility rates) and particularly in the first part of the twentieth century, labor income inequality tended to decline due to slower population growth as well as to a better ability of the educational system to allow for increases in the growth rate of workers' skills (Williamson and Lindert 1980; Williamson 1985). If we focus on wealth distribution instead, population dynamics are relevant as – in combination with inheritance systems – they contribute do determining whether the tendency is for wealth to become more or less concentrated as in general, the lower the rate of population growth, the higher the ability of the system to produce more wealth concentration over time through transfer of patrimonies to fewer inheritors (see, for a detailed discussion, Roine and Waldenström 2015, p. 552).

Institutions

Institutions play an important role in many possible explanations of the determinants of historical inequality trends. For example, as seen above, inheritance systems contribute to determining the distributive consequences of population growth from the early modern period until today (as they affect how wealth is transferred across generations), while educational systems contribute to explaining how demographic

dynamics can lead to changes in labor income inequality. Now the focus will be placed on a specific kind of institution: the fiscal systems. These are key to many explanations of inequality dynamics during the twentieth and twenty-first centuries and recently have been proposed as a fundamental cause or co-cause of the inequality tendencies found for the early modern period, too.

We are used to thinking of fiscal redistribution as something leading to lower inequality, of both income and wealth, as this is the common situation in most world countries today, including all the richest ones. In particular, taxation on income, which is usually progressive (hence those who earn larger incomes will pay proportionally more than those who earn lower incomes), reduces inequality in the short run, as by definition in a progressive fiscal regime, post-tax inequality is lower than pre-tax. Although these short-run effects may be small, they tend to accumulate over time, becoming potentially much more substantial in the medium and long run. Regarding wealth, for today's societies, the most important aspect to consider is the level and the structure of the inheritance tax, which directly affects wealth distribution (as personal or household wealth results from the combination of inherited and self-made wealth) and contributes in a crucial way to determining the degree of inheritability of wealth itself in a given society, with all that this entails for the expected long-term dynamics.

Generally speaking, fiscal redistribution seems to have been a major reason for the lull in inequality growth in the post-World War II decades. It is in this period that the progressivity of fiscal systems increased to levels never experienced before – or after (Atkinson 2004; Atkinson et al. 2011; Alvaredo et al. 2013). In 1975, the top rate on earned income was 83% in the UK, 70% in the USA, 72% in Italy, 60% in France, 56% in Germany, and 47% in Canada. Twenty-five years later, at the end of a long series of fiscal reforms that were initiated by President Ronald Reagan in the USA and Prime Minister Margaret Thatcher in the UK, the situation was inverted, with a top rate of 61% in France, 60% in Germany, 54% in Canada, 51% in Italy, 48% in the USA, and 40% in the UK (Messere 2003, p. 23). Top rates of inheritance taxes followed an entirely similar pattern. In 1980, the top rate was 75% in the UK, 70% in the USA, 35% in Germany, and 20% in France, but in 2013 the highest rate was found in France (45%), followed by the UK (40%), the USA (35%), and Germany (30%) (Piketty 2014, p. 644). By simply comparing these trends with the historical inequality trends presented in Fig. 2, we have strong hints that the progressive simplification of the Western fiscal systems[11] and the reduction in the top rates provided fertile ground for an increase in inequality. This was not just the consequence of a decline in the overall progressivity of the fiscal regime (hence in its ability to lead to a lower level of post-tax inequality compared to pre-tax), but probably also of how institutional change regarding taxation affected the pre-tax income distribution. It has been argued that lower top income rates favored

[11]The number of income "brackets" used by national personal income taxes has declined dramatically since 1975: in the case of the USA, for example, from 10 in 1975 to just 3 in 2000. Messere (2003), p. 23.

behaviors, for example, in wage negotiation, that led to increasing within-company wage differentials as higher potential net rewards favored more aggressive bargaining at the top, and this is independent from economic growth. Others favor a somewhat more optimistic picture, arguing that lower taxes stimulated economic activity, especially at the top of the income distribution (which, generally speaking, is the part of the distribution most advantaged by a reduction in top rates); hence the resulting inequality growth was in fact a collateral effect of quicker economic growth – although it has also been argued that there is no strong empirical evidence of a correlation between cuts in top tax rates and the pace of economic growth (for a synthesis, see Alvaredo et al. 2013, pp. 8–11). A final aspect to mention is that seemingly different fiscal developments contributed in a crucial way to determining the difference in the overall income inequality path followed, during the twentieth century, by English-speaking "Atlantic" countries (U-shaped) compared to continental European countries (L-shaped) (see discussion in section "How Economic Inequality Has Changed Over the Centuries" as well as Alvaredo et al. 2013).

If changes in fiscal systems seem to tell us much about the trends in economic inequality after World War II, there is reason to believe that they also played an important role in earlier decades, and in particular, that they contribute to explaining the tendency for inequality decline that was detected during the interwar period (at least for some countries and especially looking at wealth). Indeed, the interwar period was associated with the extension of the personal income tax. In the UK, for example, given the structure of exemptions, in 1912–1913 just 5% of all "tax units" were required to pay it. Already by 1930, the figure had risen to 40% of all tax units, and finally, by the end of World War II, the majority of the British population was required to pay it. In the same period, the top income tax rate rose from 8% to more than 40%. Across Europe, during the war and interwar period, the inheritance tax also increased significantly (Atkinson 2004; Piketty 2014). However, as seen in section "How Economic Inequality Has Changed Over the Centuries," arguably other factors played an even more important role in reducing economic inequality than the deepening of progressive taxation. For example, in the case of wealth, hyperinflation due to war (including during the years following the end of war itself) destroyed the real value of shares of the public debt, which had been owned mostly by the richest part of the population – a process that, in Piketty's words, was akin to expropriation through inflation (Piketty 2014, p. 184).

If during the twentieth century fiscal systems tended to become more progressive, prosecuting a tendency that had timidly begun during the nineteenth century, if we move further back in time, to the early modern period, we are faced with an entirely different situation as preindustrial fiscal systems were overall *regressive* – that is, the effective tax rates paid by those placed at top were lower (and considerably so) than those suffered by the bottom of society.[12] This was the consequence of a regime of

[12]Fiscal systems turned from being overall regressive and inequality enhancing to being progressive and inequality reducing at some point between the second half of the nineteenth century and the first decades of the twentieth. The exact timing is unclear because we lack specific studies of this fundamental transition.

systematic privilege, enrooted in law and institutions as well as in a culture that favored nobles over commoners, citizens over rural dwellers, and so on. With a regressive fiscal regime, post-tax inequality is higher than pre-tax inequality – and the greater the fiscal pressure, the greater the difference between pre- and post-tax distributions. Importantly, a continuous increase in per capita taxation is a distinctive feature of a fundamental historical process: the rise of the so-called fiscal-military state – that is, the emergence of the "modern" state with its deeper capacity and much greater ability to impose taxes on its subjects, chiefly to pay for the ever-increasing costs of war and defense. This process, which began in the sixteenth century, involved all European states independent of their economic conditions, as all had to play the same game if they were to protect themselves or to be able to project power outside their boundaries. For this reason, as a potential cause of the overall tendency for inequality growth reported for the early modern period, the increase in per capita taxation in the presence of a regressive fiscal system has the particularly desirable feature of being generalizable. For example, in the period ca. 1550 to 1780, per capita fiscal pressure rose by 70% in the Republic of Venice (where it was relatively high to begin with), more than trebled in the Sabaudian State, and increased sixfold in France and almost sevenfold in England and the Dutch Republic – thus increasing very significantly the ability of underlying regressive fiscal systems to foster inequality growth. Greater income inequality will produce, over time, more wealth inequality, too, by means of saving and investments. Consequently, the growth in the per capita tax burden during the early modern period can be seen as a co-cause of growth in inequality of both income and wealth (Alfani 2015; Alfani and Ryckbosch 2016; Alfani and Di Tullio 2019). This is even more clear if we consider that the main reason for collecting more and more resources – war – did not lead to inequality reduction as the consequence of state expenditure, differently from what we are used to today, when welfare and social spending represent the largest component of the public budget. Quite possibly, in preindustrial settings state expenditures further favored inequality growth, although further research on this specific aspect is much needed (for the redistributive consequences of public expenditure in the Republic of Venice, see Alfani and Di Tullio 2019).

Although it seems almost a certainty that during the early modern period the rise of the fiscal-military state played a very important role in favoring inequality growth across Europe (and beyond), this was surely not the only factor at work. Other common factors might have existed, for example, proletarianization, which, as seen above, probably played an important role in at least some specific historical phases. But more generally, it is probably wrong to focus research solely on identifying a single unifying cause of long-term inequality growth. In fact, as is arguably also the case for research on inequality trends in more recent epochs (Lindert 2000), we should openly recognize that the main causes explaining distributive dynamics can vary across space and time. As a consequence, we need to embrace complexity in explanations, pay great attention to context, and dig deep into available sources and information – that is, employ all the distinctive skills and good practices of economic historians. As an example of this way of proceeding, we can take the recent masterful study of inequality in the USA from 1700 until today by Peter Lindert and Jeff

Williamson (2016). Although they managed to provide an entirely convincing overall account of inequality dynamics over three centuries, for each specific period, they identified a range of concomitant causes able to explain the observed trends. For example, if we compare two phases of income inequality growth – 1800–1860 and 1970s–today – the proposed explanations are for 1800–1860: (i) rapid labor force growth that strained available resources (even taking into account the exceptionally high availability of natural resources in the USA compared to other advanced areas in the world in the same period) leading to rising prices of assets relative to wages, (ii) rapid technological progress favoring industries and cities and leading to widening North-South and city-countryside income gaps, and (iii) financial developments favoring the wealthy. Instead for the period 1970s–today inequality growth would be the consequence of: (i) political shifts leading to tax reforms of the kind described above, as well as to restricting of the welfare state, etc., (ii) skill-biased technological change, in particular the spread of automation, which determined a relative advantage for those with capital and skills, (iii) increasing competition in international trade (especially growing competition from Asia), which compromised wage gains at the bottom of the distribution, (iv) growing imbalances in levels of education, and (v) the rise of the financial sector favored by deregulation.

Conclusions: What Lessons from History?

This short overview of inequality trends in the long run of human history aimed to make explicit ways in which the remote past seems to shed light on recent developments, and vice versa, how the analysis of recent developments has led scholars to fruitfully explore the past with new eyes. A particularly important aspect is that the recently obtained better knowledge of the history of inequality in preindustrial times has changed how we perceive contemporary distributive dynamics by leading to the realization that the very long-run tendency has been for inequality to grow. The preindustrial period was not characterized by relatively flat distributive dynamics – on the contrary, the pace of inequality growth was probably comparable to that found in later epochs, with the possible exception of the phase of very quick income inequality growth experienced by a few countries since the 1970s. This common tendency, which is particularly clear when looking at the distribution of wealth, is interrupted only by two phases of inequality reduction, both triggered by large-scale catastrophes: the Black Death of 1347–1352 and the two World Wars of the twentieth century with the in-between troubles. If we expand the analysis to include, in a more tentative way, the first millennium, we can add to the list the progressive dissolution and ultimate "fall" of the Roman Empire during the fifth century.

The available historical evidence, then, strongly suggests that across two centuries and more, the overall tendency for inequality has been to increase, without, as it seems, any form of automatic rebalancing of the "Kuznetsian" kind. Admittedly, economic-historical research has underlined, for each period, different sets of potential causal factors leading to inequality growth – but what seems to be constant over time is that historically, it was much easier for inequality to grow than to decline,

as per inertia. Distributional developments in recent years lead us to think that this is what we should expect from the future, too. But is it really the case that only large-scale disasters were able to (temporarily) stop inequality growth? And hence should we be resigned to suffering growing disparities among human beings as the lesser of two evils? Luckily enough, history also offers some evidence that long-term inequality trends were heavily affected by human agency. The lull and even further decline in inequality after the end of World War II was also the effect of institutional innovations: the redistributive policies (in particular, strongly progressive taxation) and the development of welfare states from the 1950s to the early 1970s. It is only when those policies were weakened and that development was arrested (and at least in some countries, partially inverted) that inequality of both income and wealth found once again fertile ground to grow. In the opposite direction, the rise of the fiscal-military state, with the related increase in the per capita fiscal burden in a context of *regressive* taxation, fueled inequality growth throughout the early modern period.

The historical experience of Western countries does not necessarily teach us whether we should consider the current trend toward growth in economic inequality as an undesirable outcome or a problem per se (although there would be at least some ground to argue for that). What it does seem to teach us, though, is that altering that trend is entirely within our capacity, through our ability to change the institutions that in turn contribute to shaping our societies. At the same time, it teaches us that if we do not act, we have no reason whatsoever to expect that inequality will, one day, decline on its own. In other words, economic history allows us to identify quite clearly the boundaries of our playing field – but it is up to us to decide which kind of game we want to play.

References

Alfani G (2010) The effects of plague on the distribution of property: Ivrea, Northern Italy 1630. Popul Stud 64:61–75

Alfani G (2015) Economic inequality in northwestern Italy: a long-term view (fourteenth to eighteenth centuries). J Econ Hist 75(4):1058–1096

Alfani G (2017) The rich in historical perspective. Evidence for preindustrial Europe (ca. 1300–1800). Cliometrica 11(3):321–348

Alfani G, Ammannati F (2017) Long-term trends in economic inequality: the case of the Florentine State, ca. 1300–1800. Econ Hist Rev 70(4):1072–1102

Alfani G, Di Tullio M (2019) The Lion's share. Inequality and the rise of the fiscal state in preindustrial Europe. Cambridge University Press, Cambridge

Alfani G, Murphy T (2017) Plague and lethal epidemics in the pre-industrial world. J Econ Hist 77(1):314–343

Alfani G, Ryckbosch W (2016) Growing apart in early modern Europe? A comparison of inequality trends in Italy and the Low Countries, 1500–1800. Explor Econ Hist 62:143–153

Alvaredo F, Atkinson AB, Picketty T, Saez E (2013) The top 1 percent in international and historical perspective. J Econ Perspect 27(3):3–20

Atkinson AB (2004) Income tax and top incomes over the twentieth century. Hacienda Pública Española/Rev Econ Pública 168:123–141

Atkinson AB, Piketty T, Saez E (2011) Top incomes in the long run of history. J Econ Lit 49(1):3–71
Bengtsson E, Missiaia A, Olsson M, Svensson P (2018) Wealth inequality in Sweden, 1750–1900. Econ Hist Rev 71(3):772–779
Blum LE, Durlauf SN (2015) Capital in the twenty-first century: a review essay. J Polit Econ 123(4):749–777
Borgerhoff Mulder M, Bowles S, Hertz T et al (2009) Intergenerational wealth transmission and the dynamics of inequality in small-scale societies. Science 326:682–688
Bowles S, Smith EA, Borgerhoff Mulder M (2010) The emergence and persistence of inequality in premodern societies. Curr Anthropol 51(1):7–17
Brenner YS, Kaelble H, Thomas M (eds) (1991) Income distribution in historical perspective. Cambridge University Press, Cambridge
Canbakal J. Wealth and inequality in Ottoman Bursa, 1500–1840, paper given at the Economic History Society Annual Conference (York, 5–7 September 2013)
Credit Suisse Research Institute (2017) Global wealth databook 2017
Davis JB, Di Matteo L. Filling the gap: long run Canadian wealth inequality in international context. Research report n. 1/2018, Department of Economics, Western University (Canada)
Deaton A (2013) The great escape: health, wealth and the origins of inequality. Princeton University Press, Princeton
Diamond J (1997) Guns, germs, and steel: a short history of everybody for the last 13,000 years. Vintage, London
Kuznets S (1955) Economic growth and income inequality. Am Econ Rev 45(1):1–28
Lindert PH (1986) Unequal English wealth since 1670. J Polit Econ 94(6):1127–1162
Lindert PH (1991) Toward a comparative history of income and wealth inequality. In: Brenner YS, Kaelble H, Thomas M. (eds.) Income distribution in historical perspective. Cambridge University Press, Cambridge, pp. 212–231
Lindert PH (2000) Three centuries of inequality in Britain and America. In: Atkinson AB, Bourguignon F (eds) Handbook of income distribution. Elsevier, London, pp 167–216
Lindert PH (2014) Making the most of capital in the 21st century. NBER working paper no. 20232. National Bureau of Economic Research, Cambridge MA
Lindert PH, Williamson JG (2016) Unequal gains. American growth and inequality since 1700. Princeton University Press, Princeton
Malinowski M, Van Zanden JL (2017) Income and its distribution in preindustrial Poland. Cliometrica 11(3):375–404
Messere K (2003) Tax policy: theory and practice in OECD countries. Oxford University Press, Oxford
Milanovic B (2013) The inequality possibility frontier. Extensions and new applications. Policy research working paper. The World Bank, Washington, DC
Milanovic B (2016) Global inequality: a new approach for the age of globalization. Harvard University Press, Cambridge, MA
Milanovic B (2017) Income level and income inequality in the Euro-Mediterranean region, c. 14-700. Rev Income Wealth. Online-first version. https://doi.org/10.1111/roiw.12329
Milanovic B, Williamson JG, Lindert PH (2007) Measuring ancient inequality. World Bank policy research working paper no. 4412. National Bureau of Economic Research, Cambridge, MA
Nicolini EA, Ramos Palencia F (2016) Decomposing income inequality in a backward pre-industrial economy: old castile (Spain) in the middle of the 18th century. Econ Hist Rev 69(3):747–772
OECD (2015) OECD Income inequality data update: Sweden. http://www.oecd.org/sweden/OECD-Income-Inequality-Sweden.pdf
Piketty T (2006) The Kuznet's curve, yesterday, and tomorrow. In: Banerjee A, Benabou R, Mookerhee D (eds) Understanding poverty. Oxford University Press, Oxford, pp 63–72
Piketty T (2014) Capital in the twenty-first century. Belknap Press of Harvard University Press, Cambridge, MA

Piketty T, Zucman G (2014) Capital is Back: wealth-income ratios in rich countries, 1700–2010. Q J Econ 109(3):1255–1310
Piketty T, Postel-Vinay G, Rosenthal JL (2006) Wealth concentration in a developing economy: Paris and France, 1807–1994. Am Econ Rev 96(1):236–256
Piketty T, Postel-Vinay G, Rosenthal JL (2014) Inherited vs self-made wealth: theory and evidence from a rentier society (Paris 1872–1937). Explor Econ Hist 51:21–40
Reis J (2017) Deviant behaviour? Inequality in Portugal 1565–1770. Cliometrica 11(3):297–319
Roine J, Waldenström D (2015) Long run trends in the distribution of income and wealth. In: Atkinson A, Bourguignon F (eds) Handbook of income distribution, vol 2A. North-Holland, Amsterdam
Ryckbosch W (2016) Economic inequality and growth before the industrial revolution: the case of the Low Countries (fourteenth to nineteenth centuries). Eur Rev Econ Hist 20:1–22
Saito O (2015) Growth and inequality in the great and little divergence debate: a Japanese perspective. Econ Hist Rev 68(2):399–419
Santiago-Caballero C (2011) Income inequality in central Spain, 1690–1800. Explor Econ Hist 48(1):83–96
Scheidel W (2017) The great leveller: violence and the global history of inequality from the stone age to the present. Oxford University Press, Oxford
Scheidel W, Friesen SJ (2009) The size of the economy and the distribution of income in the Roman empire. J Roman Stud 99:61–91
Soltow L, Van Zanden J (1998) Income and wealth inequality in the Netherlands, 16th–20th centuries. Het Spinhuis, Amsterdam
Sutch R (2016) The accumulation, inheritance, and concentration of wealth during the gilded age: an exception to Thomas Piketty's analysis. Paper presented at the UCR Emeriti/ae Association, Orbach Science Library, University of California Riverside, February 4, 2016
Tilly C (1984) Demographic origins of the European proletariat. In: Levine D (ed) Proletarianization and family history. Academic, Orlando, pp 1–85
Van Bavel B (2016) The invisible hand? How market economies have emerged and declined since AD 500. Oxford University Press, Oxford
Van Zanden JL (1995) Tracing the beginning of the Kuznets curve: Western Europe during the early modern period. Econ Hist Rev 48(4):643–664
Waldenström D (2009) Lifting all boats? The evolution of income and wealth inequality over the path of development. Lund Studies in Economic History no 51, Lund University
Williamson JG (1985) Did British capitalism breed inequality? Allen & Unwin, Boston
Williamson JG, Lindert PH (1980) American inequality: a macro economic history. Academic, New York
World Wealth and Income Database (WID). https://wid.world/data/

Agricliometrics and Agricultural Change in the Nineteenth and Twentieth Centuries

Vicente Pinilla

Contents

Abstract

Before the industrial revolution, agriculture was the most important economic activity of traditional societies. The spread of industrialization processes, first throughout a large part of the western world and later across many more countries, gave rise to an abundance of literature on the role of agriculture in these processes. The initial perspectives offered by economic history, particularly for the British case, and the approaches of development economics specialists, largely based on previous studies by economic historians, became subject to

V. Pinilla (✉)
Department of Applied Economics, Faculty of Economics and Business Studies, Universidad de Zaragoza and Instituto Agroalimentario de Aragon -IA2- (Universidad de Zaragoza-CITA), Zaragoza, Spain
e-mail: vpinilla@unizar.es

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_55

reconsideration when numerous studies emerged that, from a cliometric point of view, sought to evaluate the changes experienced by agriculture and their contribution to economic growth. In this context, the objective of this study is to use these contributions to analyze the profound transformations that have occurred in agriculture around the world over the last two centuries.

Keywords
Economic history · Cliometrics · Agricliometrics · Agricultural production · Agricultural productivity · Technological change · Agricultural trade · Globalization · Agricultural policies · Agrarian institutions

Introduction

The agricultural sector has traditionally been a highly popular field of study among economic historians. Therefore, it is not surprising that since the emergence of new economic history in the 1950s, the cliometric approach has been frequently used in the analysis of the changes experienced by agriculture over the long term. Many researchers have used this methodology to explain the agricultural transformations, and in recent years a specialized forum has been created (Agricliometrics Conferences) in which cliometric agricultural historians are able to discuss their work. Three of these events have been organized to date (Zaragoza 2011 and 2015, Cambridge 2017).

This methodological approach, which is predominant in the universities of many developed countries today, has made a significant contribution to an in-depth and renewed vision of the principal changes that have taken place in the agricultural sector and the forces that have driven them. The use of explicit-theoretical approaches based on economic theory and the analysis of historical agricultural data with econometric techniques have enabled us to extend our knowledge substantially. For many regions, we have been able to establish not only by how much production has increased, but we are also able to explain whether this growth is due to a better use of inputs or improvements in productivity. It has also been possible to gain a thorough knowledge of the changes occurring in agricultural trade during the two waves of globalization and the forces that drove them. Finally, the agricultural policies, their effects, and the institutional changes have been examined from a new perspective aimed, mainly, at understanding the political economy that explained them and the reasons that enable us to understand them.

Furthermore, in publications providing an overall perspective of the agricultural change in the modern world, cliometric agricultural historians have summarized their previous contributions and renewed our understanding of the long-term evolution of the agricultural sector. Without a doubt, the most important study is the general overview by Federico (2005, 2014, and 2017) of the economic history of global agriculture in the nineteenth and twentieth centuries. Also from this perspective, views about the agricultural change in Europe (Lains and Pinilla 2009a), the global periphery (Pinilla and Willebald 2018a), or different regions of the world

(Hillbom and Svensson 2013) have been published. Moreover, economic history publications on certain relevant countries in the global economy which incorporate a study of agriculture have largely been undertaken by cliometric historians. Such is the case of the United States (Atacket al. 2000; Olmstead and Rhode 2000) and Great Britain (Allen 2004; Ó Gráda 1981; Turner 2004).

The interest of cliometric researchers in agriculture is not surprising. Until the beginning of the industrial revolution, agriculture was the principal economic activity of today's developed countries. There has been intense debate among economic historians about the role played by agriculture in the transformation of these countries from pre-industrial nations to advanced economies (Lains and Pinilla 2009b). Today, agriculture is still important in developing countries, and, although in developed nations, it has a very low weight in terms of production or the use of labor, it is still strategic as it produces the food essential for human beings. Furthermore, it is closely linked to the agri-food industry.

Within this context, the objective of this chapter is to offer a renewed vision of the changes that have taken place in global agriculture over the last two centuries in light of the research undertaken in recent decades by cliometric historians. Over the last 20 years, thanks to the surveys or joint volumes that have been published, this methodology has gone beyond the mere reporting of results in specialized academic journals. Nevertheless, it is still important to synthesize the principal results, as cliometric historians have tended to propose and verify new hypotheses while questioning some of the traditional ones.

The chapter is structured into five sections following this introduction. The first addresses the long-term evolution of agricultural production and productivity. An analysis of the technological change in agriculture is the subject of the second section. The third and fourth sections address the development of agricultural trade and the integration of agricultural product markets in the first and second globalizations, respectively. The final section analyzes some relevant aspects of the institutional change that has taken place in agriculture in the period studied in this chapter.

Long-Run Agricultural Production and Productivity

Increase in Production

The first difficulty faced by cliometric historians seeking to study the long-term changes in agricultural production and productivity was the absence of sufficiently consistent serial data for many countries. Therefore, the first task was usually to carry out a meticulous statistical reconstruction of annual agricultural production, particularly in the period leading up to the Second World War (i.e., GEHR 1991; Toutain 1992; Federico 2003a).

The availability of broad statistical series subsequently enabled the analysis of changes in production. From a global perspective, agricultural production has increased in the world as a whole over the last two centuries at a much faster pace than population growth. If we take into account that, due to the demographic

transition, population has increased at a very fast rate during this period, the capacity of agriculture to feed humankind should be recognized as an important achievement (Federico 2005).

It is obvious that the most complex estimates of agricultural production growth are those referring to the nineteenth century, given the lack of quality data or the difficulty in finding them. For example, in England there is considerable controversy surrounding the estimates carried out, and there is a lack of consensus (Allen 2005; Broadberry et al. 2015; Clark 2010; Floud et al. 2011; Muldrew 2011; Clark 2018). Kelly and Ó Gráda (2013) have examined the causes of these differences, proposing compromise estimates of agricultural output for the period of the industrial revolution, which show that the British population had broken the Malthusian ceiling as its agricultural production grew at a faster rate than population. Therefore, the food resources were clearly greater than those necessary for subsistence. Also, in Portugal and Sweden, agricultural production outperformed population growth substantially. In Scania, the granary of Sweden, agricultural production increased faster than in England between 1800 and 1850, with an annual growth rate of 1.77% (Olsson and Svensson 2010). Agricultural production also grew in Portugal in the first half of the nineteenth century at an appreciable annual rate of 0.7% (Reis 2016).

The only estimate carried out regarding the growth in agricultural production on a global scale highlights that, between 1870 and 1938, growth increased at an average annual rate of 1.3%. This growth was much greater until 1913 than during the interwar period, when it slowed down considerably (Federico 2004). The increase in production during this period varied greatly between the different regions, with faster rates in South America, the settler countries, and Europe. As production grew at a faster pace than the population, during this period agriculture contributed to improving nutritional levels, at least in those regions of the world where there was higher growth. However, this growth seems modest if we compare it with what came later. Between 1938 and 2000, world agricultural production grew at a higher rate, but particularly after 1950, it grew by more than 2% annually. Again, there was considerable regional dispersion as developing countries had growth rates of over 3% per annum after 1950, while developed countries grew more slowly, particularly during the final decades of the twentieth century (Federico 2005).

The greater availability of data for developed countries enables us to analyze the growth rate of their production in more detail (Table 1). The agricultural production of European countries grew throughout the nineteenth century until the First World War at an annual rate of between 0.5% and 1.5% with notable differences between countries. These rates remained at similar levels in the interwar period, although the dispersion across countries increased and some had rates higher or lower than the aforementioned range. The annual growth rates of settler countries were very high (usually above 6%) during the conquest of new territories and the displacement of the frontier westward. Once this process was completed, in the interwar period, the growth rate fell significantly (Federico 2004). After the war developed country growth rates accelerated considerably between 1950 and 1990, when they grew at an annual rate of over 2%. Since 1990, these rates have fallen considerably, to an annual level of less than 1%, or have stagnated.

Table 1 Rate of change in gross agricultural output, 1870–2000 (%)

	1870/1913	1913/1938	1938/1948–1952	1948–1952/1958–1960	1961/2000
Africa	n.a.	n.a.	1.72	3.10	2.25
Asia	1.11	0.58	0.31	3.64	3.54
Europe	1.34	0.76	n.a.	n.a.	0.00
Western Europe	n.a.	n.a.	0.56	2.55	0.91
North and Central America	n.a.	n.a.	2.63	1.40	1.77
South America	4.43	3.05	1.68	3.13	2.92
Oceania	n.a.	n.a.	0.81	2.85	1.68
Western Settlement	2.20	0.74	n.a.	n.a.	n.a.
World	1.06	0.72	1.34	2.69	2.27

Source: Federico (2004, 2005). Between 1870 and 1938, Asia includes only Japan, India, and Indonesia; Western Settlement includes Canada, Australia, and New Zealand; South America includes Argentina, Chile, and Uruguay. Asia between 1936–1938 and 1958–1960 does not include China. The World between 1936–1938 and 1958–1960 excludes the Socialist countries

For the second half of the twentieth century, it is possible to determine the growth rates of production and their regional differences with greater precision. In terms of the growth rates of production, a clear gap can be observed between developing countries and developed countries, as the former had higher rates while the latter had steadier rates in the expansion of their production (Alston and Pardey 2014; Pinilla and Willebald 2018b). Europe is an example of the developed countries. Its production grew very fast until the end of the 1980s, with annual rates of over 2%, but then it stagnated with minimum growth in the majority of the countries and negative growth in the old centrally planned economies (Martín-Retortillo and Pinilla 2015b). In developing countries the pattern is very different. If we take the case of Latin America, the figures show that the growth rates of agriculture remained relatively stable during the whole period (somewhat lower during the crisis years of the 1970s and 1980s), with very high levels at around 3% per year (Martín-Retortillo et al. 2019).

Growth in Productivity

Once the growth rates of production had been established, the most important contributions began to focus on analyzing the changes occurring in agricultural productivity and its determining factors. Two main strategies have been followed to approach this issue: calculating the total factor productivity (TFP) and partial productivities and principal labor productivity.

The emphasis on the study of TFP is largely due to the growing role that its improvement has had in the increase of production. According to Federico (2005), we can talk about a predominantly extensive model of agricultural growth until Second World War, based on the use of a greater amount of inputs and a small

contribution of an increased TFP, and an intensive growth model after the Second World War in which a more relevant role is attributed to the improvement in TFP. Many studies have been carried out estimating the historical growth of TFP, the majority of them for countries with currently high incomes. Although there is a clear spatial variability in the results obtained, some general trends can be drawn from them. First, the studies for Great Britain (i.e., Clark 2002; Allen 1994) or the United States (Craig and Weiss 1991) show that at the beginning of the industrialization processes, the TFP grew at an annual rate of around 0.5%. The most advanced European countries followed this trend in the second half of the nineteenth century, with higher annual rates, which in some countries exceeded 1% (Van Zanden 1991). Other European countries, such as France, Italy, Portugal, or Spain, had similar rates at around 0.7% (Grantham 1993; Federico 2003a; Lains 2003; Bringas 2000). In these countries, the TFP grew even faster during the interwar period, as the industrialization processes intensified. But without a doubt, the greatest growth acceleration of the TFP occurred during the decades following the Second World War. Agricultural economists have carried out the majority of studies for the second half of the twentieth century for short periods of time, which makes the comparison of their results difficult (Federico 2014).

Two recent cliometric studies for Europe and Latin America for the whole of the period 1950–2005 enable more robust conclusions to be drawn (Table 2). Therefore, the European case shows that the growth of agricultural production in European countries followed different paths but tended to converge. A model strongly based on increased efficiency was categorically followed by the more advanced countries from the early 1960s and by the more backward countries of the southern periphery from the early 1980s. The countries of Central and Eastern Europe had to wait for their transition to market economies in the mid-1990s before they could adopt a similar model. Latin America constitutes an interesting contrast, as the countries in this continent had medium and low incomes around 1950. In this case, efficiency gains made a rather modest contribution to the considerable increase in production,

Table 2 Annual growth rates of outputs, inputs, and TFP between 1950 and 2005/2008: the big European and Latin American countries

	Output	Labor	Land	Capital	TFP
Argentina	1.68	−0.23	1.13	3.66	−0.04
Brazil	3.97	0.28	2.23	4.57	1.90
Colombia	2.55	1.01	0.98	1.99	1.18
Mexico	3.67	0.89	0.77	3.22	1.99
France	1.48	−3.87	−0.14	2.43	2.31
Germany	1.24	−3.88	−0.22	1.00	2.48
Italy	0.89	−3.78	−0.87	3.08	1.78
Poland	0.63	−1.35	−0.44	3.01	0.49
Spain	2.34	−2.52	−0.20	3.64	2.37
United Kingdom	1.06	−1.64	−0.41	1.04	1.54

Source: Martín-Retortillo and Pinilla (2015a) and Martín-Retortillo et al. (2019). European countries between 1950 and 2005; Latin American countries between 1950 and 2008

although this contribution became increasingly larger over time and was highly significant between 1994 and 2008 (Martín-Retortillo and Pinilla 2015a; Martín-Retortillo et al. 2019).

The historical studies conducted on partial productivities highlight the wide variety of agricultural systems and their evolution. The studies conducted by O'Brien and Prados de la Escosura (1992) and Van Zanden (1991) for Europe from the end of the nineteenth century to the end of the twentieth century and the study undertaken by Hayami and Ruttan (1985) for six developed countries over a long time period from the end of the nineteenth century to the end of the twentieth century enable us to appreciate a wide variety of models. In all of them, over these 100 years, the increases in productivity of the land and of labor have been spectacular. In some countries, usually those with lower levels of land productivity, improvements in labor productivity were achieved mainly by increasing the number of hectares farmed per worker, with the United States being the most notable case. In other countries, such as Japan, it was largely the improvement in land productivity that contributed to the increase in labor productivity.

This variety in the agricultural development models can be generalized to the whole of the world, which also shows the existence of different trends in the second half of the twentieth century, particularly based on the improvement of land productivity in Asian countries or labor productivity in developed countries (Alston and Pardey 2014; Federico 2005).

Technological Change

A key study in the analysis of technological change in agriculture is, undoubtedly, the one conducted by Hayami and Ruttan (1985). It combines a theoretical proposal for its understanding, an international comparison between developed and developing countries and a historical-econometric analysis contrasting the cases of the United States and Japan. Based on Hicks' analysis of technical change, the authors proposed the theory of induced innovation to explain the technological change in agriculture, whereby the direction of this change depends on the relative prices of the factors. In the United States, as labor was a relatively scarce and expensive factor, the mechanization of agriculture was predominant in order to save on labor, while in Japan, where land was the scarce factor, biological innovations predominated. This is what Olmstead and Rhode (1993) call the "level approach" to technical change, meaning that even at constant relative factor prices, innovations try to save relative scarce factors.

Some studies seek to verify the key factors that determined the adoption of the new agricultural machinery. For the case of the United States, David (1966) concluded that the adoption of the harvester in the Midwest depended on whether certain economies of scale were obtained, which, in turn, critically depended on the variation of the relative prices of the factors. However, Olmstead (1975) pointed out that small variations in the values of the variables that intervened in the model had significant repercussions on the profitability threshold farm size for adopting

machinery and that the hire or joint purchase of machinery enabled this size to be reduced. In Europe, Reis (1982) concluded that in order to adopt the steam thresher in Portugal, the variation in the price of labor, the cereal yields, and the hiring or joint purchase of these machines were key factors. For Italy, the diffusion of the steam thresher from the 1870s was determined by the cost of the capital and was facilitated because these machines were purchased by specialized entrepreneurs, who in turn rented them to farmers and landowners (Federico 2003b).

An additional problem in the analysis of technological change in agriculture is the classification of the innovations as land or labor saving. Some of them can produce effects simultaneously in both directions (Federico 2005; Olmstead and Rhode 1995). For example, in the case of the United States, "tractorization" substantially reduced the land required to feed draught animals (Olmstead and Rhode 2001).

However, the analysis of technological change in agriculture must be more complex than studying the mechanical application of the theory of induced innovation, as shown in the in-depth review of the North American case carried out by Olmstead and Rhode (2008, 2015). They highlighted the importance that biological innovations (not mechanical) had throughout the nineteenth century and the first third of the twentieth century. Biological innovations were important for the country's major crops, particularly through the import of genetic material from other places and the growing capacity to combat plagues and insects that threatened harvests. Furthermore, one of the most important, though often overlooked, consequences for land productivity was the increasing capacity to control and eradicate epidemics suffered by animals for human consumption. Finally, they showed that the evolution of the relative prices of land and labor in the United States moved in the opposite direction as that forecast by Hayami and Ruttan. Throughout the nineteenth century and until 1910, the ratio between the value of land and agricultural wages increased progressively, i.e., there was higher growth in land prices than in labor prices (Olmstead and Rhode 1993).

Innovations designed to increase the productivity of land have also been important in European agriculture. Until the Second World War, the use of chemical fertilizers and pesticides substantially contributed to increasing the productivity of land (Van Zanden 1991). During the first decades after the war, European agriculture intensified its yields. In the case of arid or semiarid areas (not only in Europe), the extension and intensification of irrigation has also been fundamental. In Europe, the irrigated area increased from 8.7 million hectares to 19.1 million hectares, 10 million of which are located in Mediterranean Europe (Martín-Retortillo and Pinilla 2015a). The contribution of irrigation to the increase in agricultural production is remarkable. For example, Cazcarro et al. (2015) estimate that between 1935 and 2006, 45% of the increase in Spanish vegetable production was due to the increased area of irrigation, and 41% was due to increased productivity of irrigated land.

Several studies have sought to demonstrate the importance of both foreign and domestic demand in encouraging the adoption of technological innovations, which could increase supply without increasing prices. This is an important argument to explain the virtuous cycle that triggered the English industrial revolution, as the

demand for food from the cities played a fundamental role in the growth of agricultural productivity (Allen 2003b). In the case of European countries, Van Zanden (1991) highlights the importance of domestic demand and overall economic growth to explain the improvement in agricultural productivity. Also, in the case of Germany, agriculture reacted to urban and industrial development rather than shaping it (Kopsidis and Hockmann 2010; Kopsidis and Wolf 2012; Pfister and Kopsidis 2015). In the case of France in the first half of the nineteenth century, demand was the main driver of the improvement in productivity (Grantham 1989). In Sweden, institutional change and growing markets for agricultural products boosted agricultural production and productivity between 1700 and 1860 (Olsson and Svensson 2010). The same argument has been used to explain the growth in North American agriculture throughout the twentieth century, emphasizing the importance of labor market adjustments and the relocation of the agricultural labor force to the nonagricultural sector as the principal cause of increased productivity in agriculture (Gardner 2002). In general, both internal and external demand have been identified as being important factors for the development of European agriculture (Lains and Pinilla 2009a).

Supply factors also played an important role in the technological change and the improvement in productivity. Returning to Hayami and Ruttan (1985), it seems obvious that, with all the precautions that we have already indicated, relative prices were important for understanding the direction in which innovation was moving. Here we can add some other key factors. First, the role of the public research centres has been an important factor driving innovation, more so in some sectors than others (Ruttan 2002). Due to the difficulties in defining property rights of some innovations, the public sector organized an agricultural research system that has produced remarkable results, which are clearly reflected in the case of the green revolution. Therefore, organized public (and also private) investment in agricultural R&D was a primary driver of rapid growth in productivity in the second half of the twentieth century (Alston et al. 2009; Federico 2014). Furthermore, as we shall see later, the growth of agricultural cooperatives facilitated the diffusion of innovations. Finally, access to credit was a determining factor for adopting innovations, as many required large capital investments.

The First Wave of Globalization and the Growth of Agricultural and Food Trade

Market Integration and Agricultural Trade

The upsurge in trade, together with mass transoceanic migrations of workers and capital movements, were the essential components of the first wave of globalization from the early decades of the nineteenth century (O'Rourke and Williamson 1999). A strict definition of globalization such as that commonly used by economic historians requires an increasing integration of factor and product markets

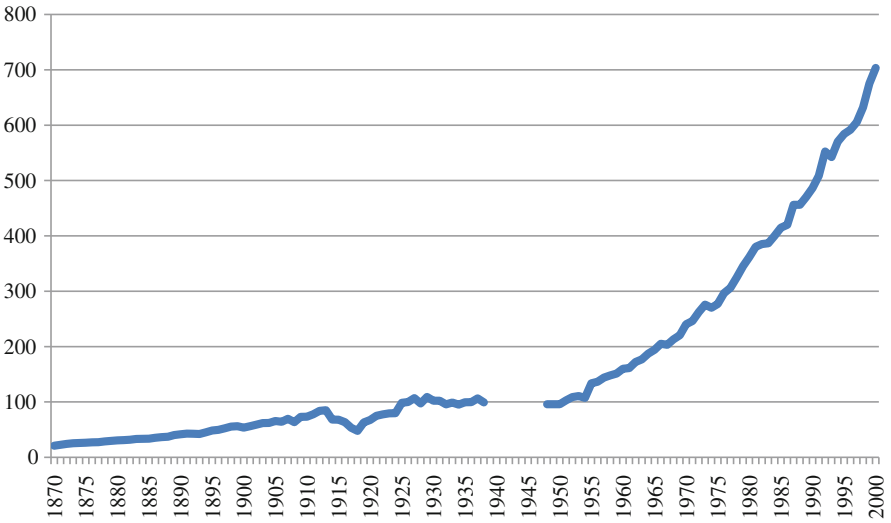


Fig. 1 International trade in agricultural and food products, 1870–2000 (constant prices, index numbers, 1934=100). (Source: Own elaboration based on Aparicio et al. (2009) and González et al. (2016))

(O'Rourke and Williamson 2002). In the case of goods markets, those with a greater weight in economic activity, such as textiles or grain, played a decisive role.

In the first wave of globalization, international trade in agricultural goods expanded rapidly (Fig. 1). The estimations performed by Lewis (1952, 1981) for primary products as a whole indicate an annual growth rate of 3.7% between 1850 and 1900. In the first third of the twentieth century, agricultural trade grew at an annual rate of 1.4%. However, this period was characterized by enormous contrasts. Until 1914, the growth rate of agricultural trade was similar to that of the nineteenth century. It subsequently fell during the First World War, recovered in the 1920s, and suffered a considerable decrease in the 1930s due to the Great Depression (Aparicio et al. 2009).

During the first wave of globalization, agricultural products were a key component of international trade, representing approximately half of the total. During this period, trade was essentially interindustrial, that is, the predominant tendency was the exchange of raw materials and food products for manufactured goods. Therefore, the Heckscher-Ohlin models based on comparative advantage usually provide a convincing explanation of the specialization trends and trade patterns.

During this first wave of globalization, the integration and formation of the agricultural product markets was essential (Federico 2018). The wheat market was fundamental, as wheat was a staple product and its trade accounted for a high volume of total agricultural trade.

There is an abundance of evidence suggesting that a massive international integration process of the wheat market took place between the end of the Napoleonic wars and approximately 1870, as measured through the convergence

of its prices and the reduction of its dispersion (Jacks 2006). The northwestern European countries led this process, reaching high levels of integration before any other region (Chilosi et al. 2013). The protectionist measures implemented after the "grain invasion" in the final decades of the nineteenth century partially reversed this integration process, which resumed at the beginning of the twentieth century. The domestic markets also partially experienced this integration process (Federico 2012a; Uebele 2011). In the case of integration between Asia and the international market, the process was more continuous until the First World War with no regression at the end of the nineteenth century (Chilosi and Federico 2015). After the First World War and during the 1920s, the dispersion in prices was once again very low, but the 1930s depression disintegrated the wheat market once more (Federico and Pearsson 2007; Hynes et al. 2012). In this process, at least in the European market, the contribution of the domestic and international markets to the long-run process of integration was similar. Its main drivers were political events, trade liberalization, and the fall in transportation costs (Federico 2011).

The European demand for wheat, mostly from Britain, was the key factor in the growth of its trade in the nineteenth century. The abolition of the Corn Laws in Britain in 1846 was one decisive factor in the liberalization process that took place (Sharp and Weisdorf 2013). However, at the end of the century, the situation for the European producers became complicated by the increasing imports of wheat from the East (the Russian Empire) and from the West (the United States), at prices which were lower than those usually seen in Europe. The largest countries, such as Germany, France, Italy, and Spain, reacted by imposing protectionist barriers, which gave a certain margin to their farmers to compete with foreign production, with the condition that they modernize their farms by introducing innovations such as chemical fertilizers or machinery. Less common was the practice implemented by countries such as the United Kingdom, the Netherlands, and Denmark, which decided to maintain their free trade policies. These different responses have been explained by the fact that the grain invasion implied different shocks in different European countries and also had different effects on income distribution depending on the importance of grain production and agriculture (O'Rourke 1997). The capacity of influence of the different sectors affected by the crisis in each country was, therefore, decisive in explaining the type of trade policy adopted (Lehmann and Volckart 2011).

Until the First World War, trade in wheat was characterized by the stability of demand from the European free trade countries, while that proceeding from the protectionist countries was highly irregular. The war and the decline of European production, together with the end of Russian exports, led to a growth in production in overseas countries, while in the rest of the world production decreased. Exports from those countries also increased significantly. In the 1920s, the market began to display unequivocal signs of saturation. During the 1930s worldwide exports of wheat decreased. This fall was approximately 65 million quintals between 1928–1932 and 1934–1938 (Aparicio and Pinilla 2019). The principal explanation can be found in the increasing self-sufficiency of the principal wheat consumers due to the increase in protectionist measures. From 1929 onward, the traditionally

protectionist countries increased their tariffs on wheat imports and, in 1932, adopted additional measures to reinforce their protectionist policies. In this latter year, even the United Kingdom abandoned its free trade policy and adopted some type of protective measures.

Market integration processes also took place in other agricultural and processed agricultural products. In the case of Mediterranean fruit and vegetables, which were hardly consumed outside of their area of production before the first wave of globalization, there was a strong increase in trade and intense international competition. As exports from Southern Europe increased, new producers entered the market, with California being the most noteworthy, seeking to win market share from the traditional producers (Morilla et al. 1999). Spain obtained a prominent leadership position when, over the first third of the twentieth century, it became the world's number one exporter, with its sales increasing more quickly from the end of the nineteenth century. Rising incomes in the more developed countries and technological change in agriculture specializing in these products were key to the growth of this trade. Spanish exporters also benefited from the increasing integration of international markets, especially through declining transport costs and, to a lesser extent, trade liberalization (Pinilla and Ayuda 2010). The results of the transatlantic competition between the Mediterranean countries and the new producers were varied. Thanks to its technological leadership, efficient marketing, and protectionist policies, California was able to seize the North American market from the Mediterranean producers. However, the latter retained the European markets, thanks to highly competitive prices despite a certain technological delay. Even so, the cost of losing the North American market, which had the highest growth rate in the world, was high. For Spain, which was the leader in the trade of these products, the cost was considerable. Considering only the orange market, in the counterfactual case of a nonexistent Californian product, Spain's GDP would have been 0.8% higher in 1910 (Pinilla and Ayuda 2009). In the case of raisins, the only example where the Californians eliminated the Mediterranean competitors, the cost was extremely high and their production in the old continent stagnated or decreased substantially (Morilla et al. 1999).

In other cases, the competition was mainly intra-European, such as the case of olive oil, where Italy specialized in the highest quality and value segment, while Spain dominated the lower-quality segment (Ramon 2000).

The trade of wine increased from the mid-nineteenth century (Anderson and Pinilla 2018). It grew strongly during the second half of the nineteenth century due to the increase in demand both in countries in the north of Europe that traditionally did not consume wine and in the destination countries of European Mediterranean emigrants. The arrival of the phylloxera plague in France led to imports of wine from Spain and other countries in order to supply its domestic market and maintain its exports. France ultimately took a leadership position as both an exporter and later as an importer of wine (Ayuda et al. 2018). The plague implied a progressive orientation in France toward the high-quality segment, while other Mediterranean countries drifted toward the low-quality segment of the market. French trade policy, which opened its duty-free doors to wine from its Algerian

colony, penalized the rest of the producers with high tariffs once the Algerian production was able to replace their exports beginning in the 1890s. The wines that entered France, the leading importing market, were highly sensitive to tariff policy. Each one percent increase in the tariff reduced the market share of imports by 1.8% in the long term (Pinilla and Ayuda 2002). From the end of the nineteenth century, new competitors entered the wine market from the new world, namely, Argentina, Chile, the United States, Australia, and South Africa. Before the last quarter of the twentieth century, the new competing countries did not threaten the European markets, but helped by their fierce protectionist policies, they were able to reduce the imports of European wines slightly. The use of gravity models to explain the determinants of wine exports has revealed their sensitivity to tariff increases (Pinilla and Serrano 2008). The wine market also shows how the end of the first wave of globalization was the result of deliberate measures taken by different countries to hamper trade flows. So the French, despite specializing in high-quality wines, suffered enormously from the tariff increases caused by the new Soviet State during the Depression and Prohibition in North America (Ayuda et al. 2018).

The textile fiber markets also operated on an international scale. In the case of silk, there were three main competing countries: Japan, China, and Italy. International demand grew, and the producers responded by expanding their supply, with Japan gaining the greatest market share. China fared the worst, due to its technological delay and poor competitiveness. The silk market illustrates how two of the three principal producers (Japan and Italy) took advantage of the opportunities of an expanding demand to improve their productivity and expand their exports (Federico 1997).

In other cases, such as the Anglo-Danish trade of butter, the integration between these markets took place extremely early (Lampe and Sharp 2015a).

Export-Led Growth

One of the most relevant consequences of the advance in globalization was a new division of labor on an international scale. In this respect, those countries of the world with a clear technological and industrial leadership position, such as Britain and other Western European nations, specialized in the production and export of manufactured goods. On the contrary, other regions of the world, such as the settler countries, Africa, and Asia, took advantage of the strong demand for agricultural raw materials and food to specialize their production and supply these goods to countries with higher levels of industrialization.

In order to be able to produce the products demanded by the industrialized world in the settler countries, the development of this specialization implied the displacement, confinement, or extermination of native peoples when they were colonized. In these countries, the agri-export sector acted as a driver of their economic development, and throughout the first wave of globalization not only did they spectacularly expand their frontiers and increase the size of their economies, but they also reached high levels of income per inhabitant. Before the First World War, Australia,

Argentina, Canada, and New Zealand were among those countries with the highest per capita income. However, the jump from this agri-export specialization to industrialization was not always straightforward, and the results were varied. The progress of Australia, Canada, and New Zealand after the First World War was clearly superior to that of Argentina and Uruguay. According to Willebald (2007), there was a positive relationship between higher degrees of inequality and production-trade specialization in goods of a low aggregate value and a lower capacity to absorb more advanced technology, which was a problem when industry became the leading sector. Therefore, in the second half of the twentieth century, the production structures, which had produced an undeniable success in the exporting era in some of these countries, limited the structural change and technological processes that were essential for the culmination of the industrialization and diversification of exports (Willebald and Bértola 2013).

Not all of the countries seeking to base their economic development on agricultural exports had the same success. Even expanding their exports, countries located in tropical areas of America, Africa, and Asia were not able to develop their economies sufficiently enough to close the gap with the industrial center of the world. The growth of exports was much slower, and the linkages with the rest of the economy were very weak. As a consequence, income levels remained low, and these economies did not experience far-reaching transformations (Aparicio et al. 2018).

Other countries combined the development of an export sector of agricultural raw materials and food with the beginning of their industrialization process. Therefore, some peripheral countries of Europe, such as Italy and Spain, combined an early specialization in the export of agricultural products to their neighbors in Northern Europe with the development of a slow industrialization process. Their position in the international markets of agricultural products was strong, and this sector made an appreciable contribution to their development (Federico 1994; Clar and Pinilla 2009)

Finally, the case of the United States is noteworthy. It experienced a rapid and far-reaching industrialization while simultaneously pushing its frontier ever westward to colonize an immense territory (Atack et al. 2000). This resulted in large increases in both grain and cotton production. The exports of only two products, cotton and wheat, constituted a substantial part of total US goods exported well into the nineteenth century (North 1966). This concentration of exports in only one or two products was also common in the settler countries (Anderson 2018).

Agricultural Trade in the Second Wave of Globalization

The Second World War dealt a harsh blow to trade, which had already been significantly affected by the Great Depression and the policies implemented during it (Hynes et al. 2012). South American agricultural exports contracted by about 40% during the war (Pinilla and Aparicio 2015). For the world as a whole, in 1948–1950, agricultural trade was still 4% lower than that of 1934–1938. At the beginning of the 1950s, trade once again exceeded the prewar level, although in 1952–1954, it had only grown by 9% with respect to the prewar level (González et al. 2016).

More important than the fall in the level of trade was that of the terms of trade, which experienced a considerable deterioration in the interwar period, during which there was a structural rupture in the series, and the improvement during the Second World War did not compensate for this decline (Ocampo and Parra-Lancourt 2010).

After the postwar recovery period, world agricultural trade displayed remarkable contrasts during the second globalization that started around 1950. On the one hand, it grew at an even faster rate than during the first wave of globalization, but on the other hand, it suffered a spectacular drop in its relative share of total international trade.

Agricultural trade grew faster in this period than during the first third of the twentieth century (Fig. 1). This growth was also higher than in the expansion phase during the second half of the nineteenth century, when it grew at an average annual rate of 3.7%, whereas in the second half of the twentieth century, it grew at a rate of 4.0% (Aparicio et al. 2009). A series of econometric studies have sought to explain the causes of this rapid growth with two different methodologies: time series and gravity models.

In the first case, the rapid growth of agricultural trade is explained mostly by the increase in income and, to a lesser degree, by the fall in real agricultural prices and exchange rate stability. Transport costs and tariff protection, measured through the nominal protection coefficient, remained stable in the long term without fostering agricultural trade (Serrano and Pinilla 2010).

The cliometric research based on gravity models provides us with richer information about this evolution. First, it offers a powerful explanation for the change in the composition of agricultural trade. The trend in trade shows a substantial loss in the relative importance of commodities, while processed and high-value products gained considerable weight. This explains why we can observe the home market effect in the processed and high-value products (not in commodities), which were closely related to the progressive concentration of the agri-food industry in the developed world. Therefore, the gravity equation has enabled us to verify that the increasing returns/product differentiation theory appropriately explains the growth in trade in high-value and processed products and that the homogeneous goods/relative factor abundance theory can explain trade growth in agricultural commodities. Finally, the group of high value-added products and processed foods benefited from greater trade liberalization, as we have seen. Thus, trade flows between high-income nations, where differentiated and reference-priced products were concentrated, benefited earlier from a progressive liberalization of regional markets. Beginning in the 1990s, these products enjoyed a more liberalized trade, with a new boost from regional trade agreements (RTAs) among economies of the South and liberalization, which the Uruguay Round produced for the food and drink industry. In contrast, the markets for traditional export groups and homogeneous goods, which characterize South-North trade flows, remained tightly controlled, and the GATT had no effect on the liberalization of their trade. This suggests that closed markets were a key factor in the decline in homogeneous goods trade flows, compared to the rise in the trade in differentiated products and reference-price goods (Serrano and Pinilla 2014).

In addition, the gravity model can be used to explain the important changes that have taken place in regional shares of agricultural and food product trade. A highly interesting contrast can be observed between Europe, particularly after the Treaty of Rome, and Latin America.

The European case is surprising from the outset. In the first wave of globalization, before the Second World War, Europe, led by England and the other industrialized countries, was the world's principal importer of agricultural products and food, accounting for approximately 60% of world imports. The share decreased steadily throughout the second half of the twentieth century, falling to less than 50%. By dividing these imports into inter-European and intra-European exports, we can observe two contrary trends. Trade between the member countries of the European Union (EU) began to grow quickly after the customs union was formed in the early 1960s. At the same time, agricultural imports from non-EU countries lost importance very quickly in relative terms, and at the end of the twentieth century, their percentage share of world imports was just half of what it had been in the 1950s. In absolute terms, these imports were much lower than those from EU member countries.

Before Second World War, Europe represented only 17% of world exports of agricultural and food products but by the end of the century accounted for almost 50%. Again, the explanation lies in the strong increase in trade between EU member countries (Pinilla and Serrano 2009). Gravity models once again help us to understand the causes of these trajectories. First, exports from EU countries from 1963 to 2000 were stimulated by their high concentration in products that had a home market effect, characteristic of a pattern of intra-industrial trade associated with the growing concentration of the international agri-food industry within the EU. The expansion of these countries' markets had a major impact. This proved to be even more important as a driver of exports than growth in their export markets as intense trade specialization in industrially processed agricultural and food products created economies of scale and generated a home market effect that increased export levels. Moreover, the development of the EU enormously affected the growth of imports from the countries joining it, and the slow growth of imports from outside countries was affected by the growing agricultural self-sufficiency of the EU. Models for intra-EU trade show that membership created trade between the partners (Serrano and Pinilla 2011a).

The evolution of the agricultural trade of Latin America constitutes a striking contrast to Europe. In the second half of the twentieth century, Latin American countries lost a substantial part of their share of worldwide exports of agricultural and food products. These countries specialized intensively in the export of commodities during the "first globalization," to the point where the excellent economic results achieved by some republics can be attributed entirely to export-led growth models (Martín-Retortillo et al. 2018). Disincentives for farm exports produced by import substitution industrialization (ISI) strategies followed by these countries partly explain the decline in Latin America's relevance as a partner in world trade.

Additional factors important in helping us understand what happened in Latin America have been identified by studies using gravity models. They highlight the importance of the profound regionalization of farm trade in this period. In addition, the failure of Latin American attempts to liberalize trade in the region delayed the

creation of trade until much later. Latin American countries failed to make any major changes in the composition of their farm and food exports until very late in the period, and as a result, they remained anchored to low-demand products. Moreover, the region's agro-export specialization fundamentally consisted of basic goods and did not benefit from any long-run home market effect. The high levels of protection facing Latin American exporters throughout practically the whole of the period made it difficult for them to perform better in their target markets. The exclusion of agricultural products from the GATT agreements until almost the end of the century imposed a high cost on countries specializing in farm exports. At the same time, the European customs union created a trade distortion that affected Latin American exports by removing trade barriers between EU partners and raising a tariff wall against outside imports (Serrano and Pinilla 2016).

Another important characteristic of the evolution of agricultural trade in the second wave of globalization is its loss of relative importance with respect to total world trade. While agricultural and food products accounted for 43.0% of total world trade in 1951, this share had shrunk to just 6.7% at current values by 2000. One of the most important reasons for this significant loss of importance is the relative fall in prices. This is evident when we consider the difference between the drop in agricultural trade in terms of value and the more moderate (albeit important) decline in terms of volume, which demonstrates an extremely serious fall in relative prices. The aggregate index of the real prices of agricultural and food products presents a structural break in the level (AO2) in 1976 and in the trend (IO2) in 1977, which suggests that they suffered the impact of the 1970s oil crisis with a lag. Thus, the deterioration in terms of the trade of agricultural and food products was strong and clear in the second half of the last century and especially impacted countries that specialized in the export of the most basic products (Serrano and Pinilla 2011b).

With regard to the causes of the loss of share in terms of volume, one reason was the generalized protectionism in the international markets for agricultural products (Anderson 2009, 2016). While other types of trade, such as manufacturing, enjoyed a greater multilateral liberalization of their markets, strong market intervention caused agricultural trade growth to be based on the proliferation and success of regional trade agreements, in addition to important changes in consumption patterns related to rising income levels. Therefore, the slower growth in farm trade had a lot to do with the significant fall in agriculture's share of world GDP. The smaller share of intra-industrial trade for the majority of agricultural products was also crucial. The home market effect for agricultural exchanges was weak, which explains why these markets grew less dynamically than those of manufactured goods and total trade (Serrano and Pinilla 2012).

Public Intervention in the Agricultural Sector

In the laissez-faire period, public intervention in the agricultural sector was scarce and highly focused on trade policy. Throughout the nineteenth century, the states adopted a fairly passive role with respect not only to agriculture but also in terms of

economic activity as a whole. One area where the states did take a hands-on approach was the creation of agricultural experiment stations, which appeared in several countries. The objective was for these centers to collaborate with agricultural producers, facilitating the adoption of the latest innovations. In some cases, their role was highly relevant.

In contrast, despite the liberal paradigm advocating free trade, there were varying degrees of tariff protection in many countries, particularly for wheat. In Europe, the first half of the nineteenth century was dominated by highly protectionist policies to safeguard national wheat production. The abolition of the Corn Laws and its effects on British agriculture aroused the interest of cliometricians.

Some studies indicated that the impact of these protectionist measures on British agriculture was only moderate until the 1840s. Williamson (1990) used a counterfactual model to argue that the protection of British grain markets had only a moderate impact on the domestic market. The liberalization arising from the abolition of these laws, within a context of robust demand, implied the rapid increase of exports to Great Britain from different countries. However, during the first post-abolition decades, prices remained more or less stable in the British market, although they tended to increase in the markets of the exporting countries (Gallego 2004). This demand was also favored by the fall in transport prices and trade liberalization (O'Rourke and Williamson 1999). However, some studies have given more importance to the trade liberalization processes or the reduction in other trade costs than the reduction in transport costs, at least until the last quarter of the nineteenth century (Jacks 2006). The trade liberalization processes, brought about through the signing of many trade agreements, were particularly important during that period (Lampe and Sharp 2015b). The technological innovations also contributed to a decrease in the production costs of the exporting countries and, from the end of the 1870s, grain from the Russian Empire, and the Americas began to arrive in Europe at prices that were lower than the production costs of the European farmers.

The so-called European grain invasion generated numerous responses across the continent, although the large countries, such as Germany, France, Italy, and Spain reacted by substantially increasing their tariffs on grain. However, other countries, such as the United Kingdom and Denmark, kept their markets open. The wide variety of responses has been explained, at least partially, by the fact that the shocks suffered due to the invasion were also very diverse, on both prices and income distribution (O'Rourke 1997). However, we should note that this protectionist reaction did not imply a general increase of tariffs on other agricultural products, nor was it essential for the survival of European agriculture (Federico 2005). The estimates of the nominal rate of assistance for several Western European countries show that the support received by farmers was very low until the First World War and only grew substantially in the 1930s (Swinnen 2009).

Until the 1930s, direct public intervention in agriculture was very low, although at the end of the nineteenth century, there were some cases that went against the laissez-faire policy, such as the federal intervention to control animal health in the United States, mainly through the creation of the Bureau of Animal Industry in 1884. This organization exerted substantial influence, and its actions to eradicate the principal

diseases that decimated the livestock population had considerable success (Olmstead and Rhode 2015). Interventions that sought to prevent price reductions were also carried out in the 1930s. These policies had far-reaching effects as a consequence of the Great Depression. The passing of the Agricultural Adjustment Act of 1933 marked a divide in the scale of public intervention in agriculture and had an enormous influence on future policies both in the United States and in other developed countries (Libecap 1998). On the other hand, the fixing of minimum prices for the purchase of produce from farmers required the restriction of this production and, therefore, implied a profound intervention, which completely distorted the functioning of the agricultural product market in the United States. We cannot find policies in any other country with such a high capacity to limit the role of the market in agriculture. However, regulatory policies also began to be extended, normally on a sectoral basis, as in the case of the regulation of wine production in France (Chevet et al. 2018).

The Second World War and the needs that it created for the warring parties increased government intervention in agriculture, and in many countries the regulatory measures of production, consumption, or prices were extended. These precedents are fundamental for understanding the policies that were designed after 1945, as the tendency to intervene persisted (Federico 2012b, 2017).

The North American policy implemented in 1933 was to have profound consequences for the functioning of the international market in agricultural products. Its objective in 1933 was to raise the purchasing power of most agricultural products to their 1909–1914 parity ratios (Olmstead and Rhode 2000). Gains in agricultural efficiency and output were not matched by a comparable advance in farm incomes. This farm income problem was soon perceived as structural rather than temporary, and intervention measures were implemented in order to raise and stabilize agricultural income (González et al. 2016).

Policy developments in rich countries over the period following Second World War tended to ignore the supply side and were mainly focused on raising farm returns through price supports. One paradoxical outcome of those policies was the stimulation of production, thus aggravating the problem of surpluses (Johnson 1987). The farm problem and the notion of an equitable income for those engaged in agriculture were at the heart of most policy statements in postwar industrialized countries. However, agricultural policy was also driven by other considerations. As Johnson has pointed out, these included national self-sufficiency or autarky for food, reducing balance of payment difficulties, and benefits to consumers in the form of an assured source of supply and stable prices.

As food shortages disappeared, price supports were maintained and even reinforced, thus revealing that the objective of sustaining higher incomes in agriculture was the major driving force behind policymaking. Of course, national farm policies strongly reflected the private and public interests underlying their national political economy systems. In the case of the United States, there is no doubt that postwar agricultural policy was deeply influenced by the farm lobby. These policies, which were traditionally justified as welfare policies for farm households, have been questioned in some cliometric studies. Spoerer (2015) has calculated that these

policies were much more beneficial to European agriculture than any welfare policy could have been. Rent-seeking behavior by agricultural lobbies emerged as a fundamental determinant of these policies.

International agricultural trade after Second World War was greatly distorted by domestic farm policies. Intervention was widespread among industrialized countries and was actually permitted by the set of international rules regulating trade. On the one hand, agricultural trade was severely restricted by import control measures, but on the other hand, it was actually expanded by the use of export subsidies and restitutions. World agriculture markets were in disarray because of the distortions in prices and trade, the substantial costs imposed on taxpayers and consumers, the inefficient expansion of farm output in the industrial countries, and its associated effects on developing countries (Tyres and Anderson 1992). In the GATT, which was the principal mechanism for reducing tariffs in the second half of the twentieth century, special rules were applied to agricultural trade, and agricultural protectionism was largely untouched. The GATT rules were adapted to the agricultural policies of the developed countries, particularly the United States. This meant that, in the European integration process, it was not necessary to adapt agricultural policy to international trade rules.

Countries tended to go from taxing to subsidizing agriculture depending on other sectors and the course of their economic development. Lindert (1991) identified the existence of two clear patterns. Developed countries tended to protect their agricultural sector with protectionist policies and direct intervention in the markets, regulating prices or establishing transfers to farmers (development pattern). Developing countries did just the opposite, taxing exportable-good agriculture and protecting import-competing agriculture (anti-trade pattern). This pattern has been quantitatively confirmed with the estimates of the nominal rate of assistance (NRA). The NRA is "the percentage by which government policies have raised gross returns to farmers above what they would be without government intervention, or lowered them, if the NRA is below zero" (Anderson 2009: 11). The figures are telling that the NRA was positive in weighted average terms in the developed world at least after 1955, the first year for which data are available. Thus, the public policies of developed countries increased farm incomes by 44% in Western Europe, 39% in Japan, and 13% in the United States in the years 1955–1959. In later years, support to farmers grew considerably, especially in Western Europe and Japan. The NRA estimates for European countries show a continuous growth from the beginning of the 1950s until they reached very high values after the implementation of the Common Agricultural Policy (Swinnen 2009). The pattern identified by Lindert has been verified econometrically as a quadratic relationship between the NRA and per capita GDP. In developing countries with low levels of income, an increase in income reduces the NRA, while for higher-income levels, the NRA increases (Anderson et al. 2010). During these years in developed countries, electoral regimes and coalition governments also influenced agricultural policies. The adoption of proportional representation in Europe is related to heavier support for farmers (Fernández 2016). In developing countries, however, the pattern after Second World War was the direct or indirect taxation of farmers due to the implementation

of import-substituting industrialization strategies, which led to the overvaluation of their currencies. The consequence was the depression of price incentives so that their farmers turned toward export markets (Anderson 2009).

Agrarian Institutional Change

The analysis of institutional change has been at the heart of the research of economic historians in recent decades, undoubtedly driven by the theoretical and empirical development proposed by Douglas North (1991, 1999). The agricultural institutions have experienced extensive transformations over the last two centuries. Even when the discussion addresses how much the institutional change has contributed to the increase in production, there is broad consensus regarding its significance (Wallis 2018). The configuration of a market economy in which the economic agents respond to the price signals is decisive in not only explaining the growth in agricultural output in Europe and other countries colonized by the Europeans since the beginning of the nineteenth century but also for understanding the processes of technological change, which enabled a substantial increase in productivity.

Property Rights, Agrarian Contracts, and Labor

In the case of agriculture, the redefinition of property rights over the land and transformations in agricultural contracts were highly important (Federico 2014; Libecap 2018). From the eighteenth century and throughout the nineteenth century, the majority of European societies considerably modified an institutional framework in which the property rights over land were not clearly defined or were shared and frequently limited the capacity of landowners to act. The long shadow cast by seigniorial rights still extended across many European countries. Therefore, one of the primary objectives of the liberal revolutions, beginning with France, was the establishment of an institutional framework characteristic of a market economy. In the case of land, this implied ending the partitioning of rights where this system still existed and ensuring the capacity of owners to freely dispose of their land and act according to their interests, with clarity regarding their ownership rights. Church property rights were undermined in Catholic European countries, resulting in large areas of land coming on to the market.

The agricultural contracts have also changed over the last two centuries, displaying an enormous spatial variability. The types of contracts that exist are endogenous to each society, and cliometric research has sought to explain them accordingly. It is well known that the establishment of market-economy-style institutions did not imply that agriculture would thereafter be organized principally through large companies using salaried workers. On the contrary, family farms not only survived, but they became the dominant form of organizing agricultural production. Even so, organizational forms were varied and benefitted from coexistence advantages. Large farms that used salaried workers coexisted with family-run farms

where the land was either owned or leased. Elsewhere, particularly in Africa and Asia, the plantations, usually owned by European colonists who employed local labor, gained considerable weight. However, after decolonization they began to lose significance and were replaced primarily by small family farms due to the combination of their inherent efficiency, a more level playing field in terms of policy support, and institutional innovations to coordinate their production (Byerlee and Viswanathanm 2018).

The study of lease contracts has attracted much attention. Decades ago this was seen as feudal survival and a type of inefficient contract. However, research has since revealed them to be perfectly rational, adapting to the circumstances and characteristics of the countries where they prevailed, such as the crop mix, the social environment, or the characteristics of the agents involved (Carmona and Simpson 2012; Garrido 2017). In short, agriculture has special characteristics derived from the difficulties of obtaining perfect information, the risks assumed, the supervision costs, and, in general, the transaction costs (Federico 2006). In this way, the choice of contract has been explored in different cliometric studies. In their study on the variety of contracts – wage payment, crop sharing, and land rental – in the south of the United States, Alston and Higgs (1982) concluded that their variations over time and space depended on the relative resource endowments of the contracting parties, the prevailing risk conditions, and the transaction costs of alternative contractual arrangements. Federico (2006) shows that the total share of tenant land, under both sharecropping and fixed-rent contracts, followed a significantly declining trend due to the combined outcome of economic motivations, institutional changes, and specific policies, such as the agricultural reforms implemented in the twentieth century. In Western Europe, this decline was particularly important, most of all during the second half of the twentieth century (Swinnen 2002).

The case of Spain also shows how the number of landless workers decreased as a result of the falling ratio between land prices and rural wages. This made plots of land cheaper for landless workers to rent and buy and led to a structural change that drained the rural population from the countryside (Carmona and Rosés 2012; Carmona et al. 2018). This process did not work that well in southern Spain where landless laborers continued to constitute a significant part of the agricultural population. In short, the modernization of agriculture, with changes in technology, the prices of factors of production, or the products themselves could also have contributed to the disappearance of sharecropping contracts that had prevailed for a long time.

Considerable social conflict surrounded the agricultural contracts and the existing pattern of land ownership. In some countries, criticism of the existence of large properties or tenants asserting their rights to be owners generated serious social conflict. During the interwar period and after Second World War, a variety of agricultural reforms were implemented in different countries, resulting in the transfer of land ownership from large landowners to small farmers, landless workers, or tenants (Federico 2005). Some had a significant impact due to the volume of land transferred, such as Mexico during the 1930s and Japan between 1947 and 1950, when almost 40% of the land changed hands (Hayami and Yamada 1991). However,

the most important land ownership transfer processes were the collectivizations that took place in the Soviet Union, the Eastern European countries after 1945, and China after the communist party came to power. The collectivization of Soviet agriculture was the first that was carried out, and it served as a model for the other cases to follow. Soviet leaders justified this as the inevitable outcome of the recurring procurement crises during the years of the New Economic Policy (NEP). Gregory and Mokhtari (1993), however, argue that they were really caused by state pricing policy. The accelerated industrialization as a major economic objective was another justification of this collectivization process, although in reality its contribution to economic growth was very small and could have been achieved under the NEP framework (Allen 2003a).

Another fundamental aspect in the institutional change in agriculture was the elimination of coercive ways of providing work. In Eastern Europe, serfdom still existed in the nineteenth century and was eliminated over the course of that century (Federico 2014). In Imperial Russia, this abolition produced substantial increases in agricultural productivity, industrial growth, and peasant living standards (Markevich and Zhuravskaya 2018). Outside of Europe, slavery was commonplace in many regions, though it was eliminated throughout the nineteenth century. The United States is not only the best case study but was a favorite subject for early cliometricians (Olmstead and Rhode 2018; Sutch 2018). One of the foundational studies of cliometrics (Conrad and Meyer 1958) broke with firmly established beliefs by arguing that slave labor in the Southern United States was a profitable activity. The investment in slaves provided returns that were comparable or higher than the alternative investments (Fogel and Engerman 1974; Yasuba 1961). Other studies reveal that there was no economic reason that would have led to the disappearance of slavery in the South (Sutch 1965), although the growth prospects of the southern economy were not promising due to their incapacity to foster innovation and economic diversification (Wright 2006).

Agricultural Cooperatives

The emergence and development of agricultural cooperatives from the end of the nineteenth century in western countries led to the practice of collective purchases of new innovations, such as fertilizers or machinery. Cooperatives also provided credit for their members and helped with the processing, marketing, and distribution of farm products. These new forms of cooperation sought to improve the competitive possibilities of the farmers in response to the technological change or the problems derived from the agricultural depression at the turn of the century. In this way, family-run farms sought to better adapt to the new economic conditions arising at the end of the nineteenth century.

Coperatives enjoyed highly unequal degrees of success and explaining that a variety has occupied the attention of many economic historians. Some have identified pre-existing social capital as the key to explaining the unequal performance of agricultural cooperatives (Fernández 2014; Guinnane 2001; Henriksen 1999). The

trust between farmers emerged as a decisive factor for understanding these differences, with a whole variety of factors that could determine it. The existence of networks that allow for social interaction may also have been favored by the existing common lands and irrigation communities, such as in Spain, providing the social networks that facilitated the building of mutual trust (Beltrán Tapia 2012). The existence of this previous social capital, however, is not a guarantee that cooperatives would emerge, as other factors could prevent this (Garrido 2014; Henriksen et al. 2015). Other factors that affected the performance of cooperatives include the existing institutional framework, the predominance of small- and medium-sized family farms, the existence of well-trained human capital, the density of production, and product specialization (Henriksen et al. 2012).

The characteristics of the products were also important for the spread of cooperatives. They were much more important in the production and marketing of butter in the north of Europe than in that of wine in the south. Fernández and Simpson (2017) explain this contrast by pointing out that while in some products cooperative production could improve product quality and competitiveness in high-value markets, wine cooperatives produced mostly low-quality wine due to environmental conditions and measurement problems. However, specializing in a certain type of production did not guarantee the success of cooperatives. The successful Danish case in butter production contrasts with the poor results in Ireland (Ó Gráda 1977). The existence of serious social and political conflict on the island generated distrust between potential partners, in contrast to the religious and cultural homogeneity of Denmark (O'Rourke 2007).

The Privatization of the Common Lands

Finally, to conclude this section, we will analyze the advances made in the understanding of the role played by the common lands in economic development. In early cliometric studies, the privatization of the communal regime was seen as a precondition to foster economic growth (McCloskey 1975; North and Thomas 1978). Private property rights were believed to be essential for stimulating innovation and investment and therefore generating growth. They also indicate that an inevitable consequence of the existence of common property regimes was resource overexploitation. However, since the 1980s, the essential points of these arguments have been thoroughly revised. Common property regimes can be efficient and sustainable, which, in turn, has led to the reassessment of the possible role that they have played in economic development (Allen 1982, 1992; Clark 1998; Van Zanden 1999; De Moor et al. 2002, De Moor 2009; Ostrom 1990; 2005).

As a result, the link established between the enclosures and the British agricultural revolution has also been questioned. Allen (1992, 1999, 2003b) reveals that before the parliamentary enclosures began, the agricultural revolution was already underway, and agricultural productivity growth had also taken place on open fields. He also notes that the privatization of the common lands deprived the lower rural classes of their rights of access, which, in many cases, gave rise to the worsening of

their standard of living, forcing them to migrate to the towns (Humphries 1990). It is also important to highlight how this series of more recent studies has emphasized the fact that the commons were not open-access resources. They were regulated by local communities by way of a series of formal and informal rules that guaranteed their sustainability (De Moor 2009; Beltrán Tapia 2016).

The case of privatization of the common lands in Spain has been used to highlight that this process did not promote the growth of agricultural productivity. In contrast, it is associated with deteriorating living standards of the poorest segment of the rural population (similar to the case of Great Britain), lower educational attainments due to the limited capacity of town councils to finance primary education, and the deterioration of social capital (Beltrán Tapia 2016).

Conclusion

The emergence of cliometrics in economic history revolutionized this discipline (Haupert 2015). From the beginning of the new economic history, research in agriculture has been important to this new methodological perspective. Although this rupture with the traditional methods used by agricultural historians emerged in the United States, it has since spread to other continents, particularly Europe, and since the 1990s has grown in importance. Today a large number of economic historians base their research on economic theoretical foundations and use econometric methodology. The organization of three international conferences over the last decade to discuss the progress made in this field has established agricliometrics as a growing field of study.

Cliometricians have addressed an enormous variety of topics in the study of agricultural history, but they have concentrated a good part of their efforts on a few main subjects, primarily the study of the evolution of agricultural production and the role played by the increased productivity in its growth. The studies carried out highlight the different speeds in the growth rates of production over the last two centuries and their regional differences. Furthermore, the increasing role of the gains in productivity increase is evident, particularly after the Second World War.

The relevance of the two waves of globalization and their effects on different economies help explain the interest generated by the analysis of agricultural trade and the articulation and development processes of the agricultural product and food markets. Agricultural products played a key role in the first wave of globalization, and the analysis focuses on the integration processes of the different product markets and the drivers of these processes. In the second globalization, although agricultural trade grew even more quickly, its relative weight with respect to total trade fell sharply due to the low elasticity of demand for agricultural products, the diminished importance of its intra-industrial trade, the lower degree of differentiation of its products, and the high level of protectionism, which contrasts with the strong liberalization experienced by manufactured products.

The growing role of public intervention from the 1930s oriented many studies toward the role of government both in relation to trade policies adopted and

agricultural policies followed. Trade policies were relaxed until the end of the nineteenth century and the grain invasion experienced by Europe, and consequently protectionism was increased. However, it was the Great Depression of the 1930s that led to the collapse of the first globalization and had a profound effect on international agricultural trade. The high levels of protectionism were maintained during the decades following the Second World War, and only at the end of the twentieth century were tentative attempts made to liberalize trade.

Finally, the analysis of different agricultural institutions, such as ownership, agricultural contracts, privatization of common lands, and cooperatives, has also constituted a fertile field for cliometric research. This work highlights the importance of institutional change and the reasons explaining the continuity or modification of certain agricultural contracts as well as the unequal growth of agricultural cooperatives.

Acknowledgments This study has received financial support from Spain's Ministry of Science and Innovation, project ECO2015-65582-P, the Government of Aragon, through the Research Group 'S55_17R, and from the European Regional Development Fund.

References

Allen RC (1982) The efficiency and distributional consequences of eighteenth century enclosures. Econ J 92(368):937–953

Allen RC (1992) Enclosures and the yeomen. Oxford University Press, Oxford

Allen RC (1994) Agriculture during the industrial revolution. In: Floud R, McCloskey D (eds) The economic history of Britain since 1700, vol 1: 1700–1860. Cambridge University Press, Cambridge, MA, pp 96–122

Allen RC (1999) Tracking the agricultural revolution. Econ Hist Rev 52:209–235

Allen RC (2003a) Farm to factory. A reinterpretation of the Soviet industrial revolution. Princeton University Press, Princeton

Allen RC (2003b) Progress and poverty in early modern Europe. Econ Hist Rev 56(3):403–443

Allen RC (2004) Agriculture during the industrial revolution, 1700–1850, The Cambridge economic history of modern Britain, vol 1. Cambridge University Press, Cambridge, pp 96–116

Allen RC (2005) English and Welsh agriculture, 1300–1850: output, inputs, and income. Oxford University working paper. https://www.nuffield.ox.ac.uk/media/2161/allen-eandw.pdf

Alston LJ, Higgs R (1982) Contractual mix in southern agriculture since the civil war: facts, hypotheses, and tests. J Econ Hist 42(2):327–353

Alston J, Pardey PG (2014) Agriculture in the global economy. J Econ Perspect 28(1):121–146

Alston JM, Beddow JM, Pardey PG (2009) Agricultural research, productivity, and food prices in the long run. Science 325(4):1209–1210

Anderson K (2009) Distortions to agricultural incentives. A global perspective, 1955–2007. Palgrave Macmillan and the World Bank, Washington/New York

Anderson K (2016) Agricultural trade, policy reforms, and global food security. Palgrave McMillan, London/New York

Anderson K (2018) Agricultural development in Australia in the face of occasional mining booms: 1845 to 2015. In: Pinilla V, Willebald H (eds) Agricultural development in the world periphery: a global economic history approach. Palgrave Macmillan, London, pp 365–388

Anderson K, Pinilla V (2018) Global overview. In: Anderson K, Pinilla V (eds) Wine globalization: a new comparative history. Cambridge University Press, New York, pp 24–54

Anderson K, Croser J, Sandri D, Valenzuela E (2010) Agricultural distortion patterns since the 1950s: what needs explaining? In: Anderson K (ed) The political economy of agricultural price distortions. Cambridge University Press, New York, pp 25–80

Aparicio G, Pinilla V (2019) International trade in cereals and the collapse of the first wave of globalization, 1900–1938. J Glob Hist 14(1):44–67

Aparicio G, Pinilla V, Serrano R (2009) Europe and the international agricultural and food trade, 1870-2000. In: Lains P, Pinilla V (eds) Agriculture and economic development in Europe since 1870. Routledge, London, pp 52–75

Aparicio G, González-Esteban A, Pinilla V, Serrano R (2018) The world periphery in global agricultural and food trade, 1900–2000. In: Pinilla V, Willebald H (eds) Agricultural development in the world periphery: a global economic history approach. Palgrave Macmillan, London, pp 63–88

Atack J, Bateman F, Parker WN (2000) Northern agriculture and westward movement. In: Engerman SL, Gallman R (eds) The Cambridge economic history of the United States, vol II. Cambridge University Press, Cambridge, MA, pp 285–328

Ayuda, MI, Ferrer-Pérez, H, Pinilla, V (2018) How to become a leader in an emerging new global market: the determinants of French wine exports, 1848–1938. EHES working papers in economic history, 124

Beltrán Tapia F (2012) Commons, social capital, and the emergence of agricultural cooperatives in early twentieth century Spain. Eur Rev Econ Hist 16(4):511–528

Beltrán Tapia F (2016) Common lands and economic development in Spain. Revista de Historia Economica/J Iber Lat Am Econ Hist 34(1):111–133

Bringas MA (2000) La productividad de los factores en la agricultura española (1752–1935). Banco de España, Madrid

Broadberry S, Campbell B, Klein A, Overton M, van Leeuwen B (2015) British economic growth, 1270–1870. Cambridge University Press, Cambridge, MA

Byerlee D, Viswanathanm PK (2018) Plantations and economic development in the twentieth century: the end of an era? In: Pinilla V, Willebald H (eds) Agricultural development in the world periphery: a global economic history approach. Palgrave Macmillan, London, pp 89–118

Carmona J, Rosés J (2012) Land markets and agrarian backwardness (Spain, 1904–1934). Eur Rev Econ Hist 16(1):74–76

Carmona J, Simpson J (2012) Explaining contract choice: vertical integration, sharecropping, and wine in Europe, 1859–1950. Econ Hist Rev 65(3):887–909

Carmona J, Rosés J, Simpson J (2018) The question of land access and the Spanish land reform of 1932. Econ Hist Rev. https://doi.org/10.1111/ehr.12654

Cazcarro I, Duarte R, Martín-Retortillo M, Pinilla V, Serrano A (2015) Water scarcity and agricultural growth: from curse to blessing? A case study of Spain. In: Badía-Miró M, Pinilla V, Willebald H (eds) Natural resources and economic growth: learning from history. Routledge, London, pp 339–361

Chevet JM, Fernandez E, Giraud-Héraud E, Pinilla V (2018) France. In: Anderson K, Pinilla V (eds) Wine globalization: a new comparative history. Cambridge University Press, New York, pp 55–91

Chilosi D, Federico G (2015) Early globalizations: the integration of Asia in the world economy, 1800–1938. Explor Econ Hist 57:1–18

Chilosi D, Murphy TE Studer R, Coşkun Tunçer A (2013) Europe's many integrations: geography and grain markets, 1620–1913. Explor Econ Hist 50:46–68

Clar E, Pinilla V (2009) Agriculture and economic development in Spain, 1870–1973. In: Lains P, Pinilla V (eds) Agriculture and economic development in Europe since 1870. Routledge, London, pp 311–332

Clark G (1998) Commons sense: common property rights, efficiency, and institutional change. J Econ Hist 58(1):73–102

Clark G (2002) Farmland rental values and agrarian history: England and Wales, 1500–1912. Eur Rev Econ Hist 6:281–309

Clark G (2010) The macroeconomic aggregates for England, 1209–2008. Res Econ Hist 27:51–140
Clark G (2018) Growth or stagnation? Farming in England, 1200–1800. Econ Hist Rev 71(1):55–81
Conrad AH, Meyer JR (1958) The economics of slavery in the antebellum South. J Polit Econ 66:95–122
Craig LA, Weiss T (1991) Hours at work and total factor productivity growth in nineteenth-century U.S. agriculture. Adv Agric Econ Hist 1:1–30
David P (1966) The mechanization of reaping in the ante-bellum Midwest. In: Rosovsky H (ed) Industrialization in two systems: essays in honor of Alexander Gerschenkron. Wiley, New York, pp 3–39
de Moor T (2009) Avoiding tragedies: a flemish common and its commoners under the pressure of social and economic change during the eighteenth century. Econ Hist Rev 62(1):1–22
de Moor T, Shaw-Taylor L, Warde P (eds) (2002) The management of common land in North West Europe, ca. 1500–1850. Brepols, Turnhout
Federico G (1994) Agricoltura e sviluppo (1820–1950): verso una reinterpretazione. In: Ciocca PL (ed) Il progresso economico dell'Italia. Permenenze, discontinuitá, límiti. Il Mulino, Milan, pp 81–207
Federico G (1997) An economic history of silk industry 1830–1930. Cambridge University Press, Cambridge, MA
Federico G (2003a) Le nuove stime della produzione agrícola italiana, 1860–1910: primi resultati ed implicazioni. Rivista di Storia Economica 19:359–381
Federico G (2003b) A capital intensive innovation in a capital-scarce world: steam-threshing in nineteenth century Italy. Adv Agric Econ Hist 2:75–114
Federico G (2004) The growth of world agricultural production, 1800–1938. Res Econ Hist 22:125–181
Federico G (2005) Feeding the world: an economic history of agriculture, 1800–2000. Princeton University Press, Princeton
Federico G (2006) The 'real' puzzle of sharecropping: why is it disappearing? Contin Chang 21(2):261–285
Federico G (2011) When did European markets integrate? Eur Rev Econ Hist 15:93–126
Federico G (2012a) How much do we know about market integration in Europe? Econ Hist Rev 65(2):470–497
Federico G (2012b) Natura non Fecit saltus: the 1930s as the discontinuity in the history of European agriculture. In: Brassley P, Segers Y, Van Mollen L (eds) War, agriculture, and food. Rural Europe from the 1930s to the 1950s. Routledge, London, pp 15–32
Federico G (2014) Growth, specialization, and organization of world agriculture. In: Neal L, Williamson JG (eds) The Cambridge history of capitalism, The spread of capitalism: from 1848 to the present, vol 2. Cambridge University Press, Cambridge, MA, pp 47–81
Federico G (2017) The economic history of agriculture since 1800. In: McNeill JR, Pomeranz K (eds) The Cambridge world history, Production, destruction and connection, 1750–Present, vol 7. Cambridge University Press, Cambridge, MA, pp 83–105
Federico G (2018) Market integration. In: Diebolt C, Haupert M (eds) Handbook of cliometrics, 2nd edn. Springer, Berlin/Heildelberg
Federico G, Pearsson KG (2007) Market integration and convergence in the world wheat market, 1800–2000. In: Hatton TJ, O'Rourke KH, Taylor AM (eds) The new comparative economic history: essays in honour of Jeffrey G. Williamson. Cambridge University Press, Cambridge, MA, pp 87–114
Fernández E (2014) Trust, religion, and cooperation in western agriculture, 1880–1930. Econ Hist Rev 67(3):678–698
Fernández E (2016) Politics, coalitions, and support of farmers, 1920–1975. Eur Rev Econ Hist 20(1):102–122
Fernández E, Simpson J (2017) Product quality or market regulation? Explaining the slow growth of Europe's wine cooperatives, 1880–1980. Econ Hist Rev 70(1):122–142

Floud R, Fogel RW, Harris B, Hong SC (2011) The changing body: health, nutrition, and human development in the Western world since 1700. Cambridge University Press, Cambridge, MA
Fogel RW, Engerman SL (1974) Time on the cross: the economics of American negro slavery. Little, Brown, Boston
Gallego D (2004) La formación de los precios del trigo en España (1820–1869): el contexto internacional. Hist Agrar 34:61–102
Gardner BL (2002) American agriculture in the twentieth century: how it flourished and what it cost. Harvard University Press, Cambridge, MA
Garrido S (2014) Plenty of trust, no much cooperation: social capital and collective action in early twentieth eastern Spain. Eur Rev Econ Hist 18(4):413–432
Garrido S (2017) Sharecropping was sometimes efficient: sharecropping with compensation for improvements in European viticulture. Econ Hist Rev 3(70):977–1003
González AL, Pinilla V, Serrano R (2016) International agricultural markets after the war, 1945–1960. In: Martiin C, Pan-Montojo J, Brassley P (eds) Agriculture in capitalist Europe, 1945–1960. From food shortages to food surpluses. Routledge, London, pp 64–84
Grantham G (1989) Agricultural supply during the industrial revolution: French evidence and European implications. J Econ Hist XLIX(1):43–72
Grantham G (1993) Divisions of labour: agricultural productivity and occupational specialization in pre-industrial France. Econ Hist Rev XLVI(3):478–502
Gregory P, Mokhtari M (1993) State grain purchases, relative prices and the Soviet grain procurement crisis. Explor Econ Hist 30:182–194
Grupo de Estudios de Historia Rural (1991) Estadísticas Históricas de la producción agraria española, 1850–1935. Ministerio de Agricultura, Madrid
Guinnane T (2001) Cooperatives as information machines: German rural credit cooperatives, 1883–1914. J Econ Hist 61(2):366–389
Haupert M (2015) History of cliometrics. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer, Berlin/Heildelberg, pp 3–32
Hayami Y, Ruttan VW (1985) Agricultural development; an international perspective. Johns Hopkins University Press, Baltimore
Hayami Y, Yamada S (1991) The agricultural development of Japan: a century's perspective. University of Tokyo Press, Tokyo
Henriksen I (1999) Avoiding lock-in: cooperative creameries in Denmark, 1882–1903. Eur Rev Econ Hist 3:57–78
Henriksen I, Hviid M, Sharp P (2012) Law and peace: contracts and the success of the Danish dairy cooperatives. J Econ Hist 72:197–224
Henriksen I, Mclaughin E, Sharp P (2015) Contracts and cooperation: the relative failure of the Irish dairy industry in the late nineteenth century reconsidered. Eur Rev Econ Hist 19(4):412–431
Hillbom E, Svensson P (eds) (2013) Agricultural transformation in a global history perspective. Routledge, London
Humphries J (1990) Enclosures, common rights, and women: the proletarisation of families in the late eighteenth and early nineteenth centuries. J Econ Hist 50(1):17–42
Hynes W, Jacks DS, O'Rourke KH (2012) Commodity market disintegration in the interwar period. Eur Rev Econ Hist 16(2):119–143
Jacks D (2006) What drove 19th century commodity market integration? Explor Econ Hist 43(3):383–412
Johnson DG (1987) World agriculture in disarray. MacMillan, London
Kelly M, Ó Gráda C (2013) Numerare est errare: agricultural output and food supply in England before and during the industrial revolution. J Econ Hist 73(4):1132–1163
Kopsidis M, Hockmann H (2010) Technical change in Westphalian peasant agriculture and the rise of the Ruhr, circa 1830–1880. Eur Rev Econ Hist 14(2):209–237
Kopsidis M, Wolf N (2012) Agricultural productivity across Prussia during the industrial revolution: a Thünen perspective. J Econ Hist 72(3):634–670

Lains P (2003) New wine in old bottles: output and productivity trends in Portuguese agriculture, 1850–1950. Eur Rev Econ Hist 7(1):43–72
Lains P, Pinilla V (eds) (2009a) Agriculture and economic development in Europe since 1870. Routledge, London
Lains P, Pinilla V (2009b) Introduction. In: Lains P, Pinilla V (eds) Agriculture and economic development in Europe since 1870. Routledge, London, pp 1–24
Lampe M, Sharp P (2015a) How the Danes discovered Britain: the international integration of the Danish dairy industry before 1880. Eur Rev Econ Hist 19(4):432–453
Lampe M, Sharp P (2015b) Cliometric approaches to international trade. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer, Berlin/Heildelberg, pp 295–330
Lehmann S, Volckart O (2011) The political economy of agricultural protection: Sweden 1887. Eur Rev Econ Hist 15(1):29–59
Lewis AW (1952) World production, prices and trade, 1870–1960. Manch Sch Econ Soc XX(2):105–138
Lewis AW (1981) The rate of growth of world trade, 1830–1973. In: Grassman S, a Lundberg E (eds) The world economic order. Palgrave Macmillan, London, pp 11–74
Libecap G (1998) The great depression and the regulating state: federal government regulation of agriculture. In: Bordo M, Goldin C, White EN (eds) The defining moment. The great depression and the American economy in the twentieth century. University of Chicago Press, Chicago, pp 181–226
Libecap G (2018) Property rights. In: Diebolt C, Haupert M (eds) Handbook of cliometrics, 2nd edn. Springer, Berlin/Heildelberg
Lindert P (1991) Historical patterns of agricultural policy. In: Timmer PC (ed) Agriculture and the State. Growth, employment, and poverty in developing countries. Cornell University Press, Ithaca, pp 1–29
Markevich A, Zhuravskaya E (2018) The economic effects of the abolition of serfdom: evidence from the Russian Empire. Am Econ Rev 108(4-5):1074–1117
Martín-Retortillo M, Pinilla V (2015a) Patterns and causes of growth of European agricultural production, 1950–2005. Agric Hist Rev 63(I):132–159
Martín-Retortillo M, Pinilla V (2015b) On the causes of economic growth in Europe: why did agricultural labour productivity not converge between 1950 and 2005? Cliometrica 9:359–396
Martín-Retortillo M, Pinilla V, Velazco J, Willebald H (2018) The goose that laid the golden eggs? Agricultural development in latin America in the twentieth century. In: Pinilla V, Willebald H (eds) Agricultural development in the world periphery: a global economic history approach. Palgrave Macmillan, London, pp 337–364
Martín-Retortillo M, Pinilla V, Velazco J, Willebald H (2019) The dynamics of latin American agricultural production growth since 1950. J Lat Am Stud. https://doi.org/10.1017/S0022216X18001141
McCloskey DN (1975) The persistence of English common fields. In: Parker WN, Jones EL (eds) European peasants and their markets: essays in agrarian economic history. Princeton University Press, Princeton, pp 73–119
Morilla J, Olmstead A, Rhode PW (1999) "Horn of plenty": the globalization of Mediterranean horticulture and the economic development of southern Europe, 1880–1930. J Econ Hist 59(2):316–352
Muldrew C (2011) Food, energy, and the creation of industriousness. Oxford University Press, Oxford
North D (1966) The economic growth of the United States: 1790–1860. W. W Norton & Company, Englewood Cliffs
North D (1991) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge, MA
North D (1999) Understanding the process of economic change. Princeton University Press, Princeton
North D, Thomas R (1978) The first economic revolution. Econ Hist Rev 30:229–241

Ó Gráda C (1977) The beginnings of the Irish creamery system, 1880–1914. Econ Hist Rev XXX:284–305
Ó Gráda C (1981) Agricultural decline, 1860–1914. In: Floud RC, McClosley DN (eds) An economic history of Britain since 1700, vol II. Cambridge University Press, Cambridge, MA, pp 157–197
O'Brien PK, Prados de la Escosura L (1992) Agricultural productivity and European industrialization, 1890–1980. Econ Hist Rev 45(3):514–536
O'Rourke KH (1997) The European grain invasion, 1800–1913. J Econ Hist 57:775–801
O'Rourke KH (2007) Culture, conflict and cooperation: Irish dairying before the Great War. Econ J 117:1357–1379
O'Rourke KH, Williamson JG (1999) Globalization and history. The evolution of the nineteenth Atlantic economy. The MIT Press, Cambridge, MA
O'Rourke KH, Williamson JG (2002) When did globalization begin? Eur Rev Econ Hist 6:23–50
Ocampo JA, Parra-Lancourt MA (2010) The terms of trade for commodities since the mid-19th century. Rev Hist Econ/J Iber Lat Am Econ Hist 28(1):11–44
Olmstead AL (1975) The mechanization of reaping and mowing in American agriculture, 1833–1870. J Econ Hist XXXV:327–352
Olmstead AL, Rhode PW (1993) Induced innovation in American agriculture: a reconsideration. J Polit Econ 101(1):100–118
Olmstead AL, Rhode PW (1995) Beyond the threshold: an analysis of the characteristics and behavior of early reaper adopters. J Econ Hist 55(1):27–57
Olmstead AL, Rhode PW (2000) The transformation of Northern agriculture. In: Engerman SL, Gallman R (eds) The Cambridge economic history of the United States, vol III. Cambridge University Press, Cambridge, MA, pp 693–742
Olmstead AL, Rhode PW (2001) Reshaping the landscape: the impact and diffusion of the tractor in American agriculture, 1910–1960. J Econ Hist 61(3):663–698
Olmstead AL, Rhode PW (2008) Creating abundance. Biological innovation and American agricultural development. Cambridge University Press, New York
Olmstead AL, Rhode PW (2015) Arresting contagion: science, policy, and conflicts over animal disease control. Harvard University Press, Cambridge, MA
Olmstead AL, Rhode PW (2018) Cotton, slavery, and the new history of capitalism. Explor Econ Hist 67:1–17
Olsson M, Svensson P (2010) Agricultural growth and institutions: Sweden, 1700–1860. Eur Rev Econ Hist 14(2):275–304
Ostrom E (1990) Governing the commons. The evolution of institutions for collective action. Cambridge University Press, New York
Ostrom E (2005) Understanding institutional diversity. Princeton University Press, Princeton
Pfister U, Kopsidis M (2015) Institutions versus demand: determinants of agricultural development in Saxony, 1660–1850. Eur Rev Econ Hist 19(3):275–293
Pinilla V, Aparicio G (2015) Navigating in troubled waters: South American exports of food and agricultural products, 1900–1950. Rev Hist Econ/J Iber Lat Am Econ Hist 33(2):223–255
Pinilla V, Ayuda MI (2002) The political economy of the wine trade: Spanish exports and the international market, 1890–1935. Eur Rev Econ Hist 6:51–85
Pinilla V, Ayuda MI (2009) Foreign markets, globalisation and agricultural change in Spain. In: Pinilla V (ed) Markets and agricultural change in Europe from the 13th to the 20th century. Brepols Publishers, Turnhout, pp 173–208
Pinilla V, Ayuda MI (2010) Taking advantage of globalization? Spain and the building of the international market in Mediterranean horticultural products, 1850–1935. Eur Rev Econ Hist 14(2):239–274
Pinilla V, Serrano R (2008) The agricultural and food trade in the first globalization: Spanish table wine exports 1871 to 1935 – a case study. J Wine Econ 3(2):132–148
Pinilla V, Serrano R (2009) Agricultural and food trade in the European Community since 1961. In: Patel K (ed) Fertile ground for Europe? The history of European integration and the common agricultural policy since 1945. Nomos, Baden-Baden, pp 270–273

Pinilla V, Willebald H (eds) (2018a) Agricultural development in the world periphery: a global economic history approach. Palgrave-Macmillan, London
Pinilla V, Willebald H (2018b) Agricultural development in the world periphery: a general overview. In: Pinilla V, Willebald H (eds) Agricultural development in the world periphery: a global economic history approach. Palgrave-Macmillan, London, pp 3–28
Ramon R (2000) Specialization in the international market for olive oil before world war II. In: Pamuk S, Williamson JG (eds) The mediterranean response to globalization before 1950. Routledge, London, pp 159–198
Reis J (1982) Latifúndio e progresso técnico:a difusão da debulha mecânica no Alentejo, 1860–1930. Análise Soc XVIII(71):371–433
Reis J (2016) Gross agricultural output: a quantitative, unified perspective, 1500–1850. In: Freire D, Lains P (eds) An agrarian history of Portugal, 1000–2000. Economic development on the European frontier. Brill, Leiden, pp 166–196
Ruttan VW (2002) Productivity growth in world agriculture: sources and constraints. J Econ Perspect 16(4):161–184
Serrano R, Pinilla V (2010) Causes of world trade growth in agricultural and food products, 1951–2000: a demand function approach. Appl Econ 42(27):3503–3518
Serrano R, Pinilla V (2011a) Agricultural and food trade in European Union countries, 1963–2000: a gravity equation approach. Économies et SociÉtÉs, SÉrie Histoire Économique quantitative, AF 43(1):191–219
Serrano R, Pinilla V (2011b) Terms of trade for agricultural and food products, 1951–2000. Rev Hist Econ/J Iber Lat Am Econ Hist 29(2):213–243
Serrano R, Pinilla V (2012) The long-run decline in the share of agricultural and food products in international trade: a gravity equation approach of its causes. Appl Econ 44(32):2199–2210
Serrano R, Pinilla V (2014) Changes in the structure of world trade in the agrifood industry: the impact of the home market effect and regional liberalization from a long term perspective, 1963–2010. Agribus Int J 30(2):165–183
Serrano R, Pinilla V (2016) The declining role of latin America in global agricultural trade, 1963–2000. J Lat Am Stud 48(1):115–146
Sharp P, Weisdorf J (2013) Globalization revisited: market integration and the wheat trade between North America and Britain from the eighteenth century. Explor Econ Hist 50(1):88–98
Spoerer M (2015) Agricultural protection and support in the European Economic Community, 1962–92: rent-seeking or welfare policy? Eur Rev Econ Hist 19(2):195–214
Sutch R (1965) The profitability of ante bellum slavery−revisited. South Econ J 31:365–377
Sutch R (2018) Slavery. In: Diebolt C, Haupert M (eds) Handbook of cliometrics, 2nd edn. Springer, Berlin/Heildelberg
Swinnen J (2002) Political reforms, rural crises, and land tenure in Western Europe. Food Policy 27:371–394
Swinnen J (2009) The growth of agricultural protection in Europe in the 19th and 20th centuries. World Econ 32(11):1499–1537
Toutain JF (1992) La production agricole de la France de 1810 à 1990: départements et régions: croissance, productivité, structures. Presses Universitaires de Grenoble, Grenoble
Turner M (2004) Agriculture, 1860–1914. In: Flour R, Johnson P (eds) The Cambridge economic history of modern Britain, Economic maturity, 1860–1938, vol II. Cambridge University Press, Cambridge, MA, pp 161–189
Tyres R, Anderson K (1992) Disarray in world food markets: a quantitative assessment. Cambridge University Press, Hong Kong
Uebele M (2011) National and international market integration in the 19th century: evidence from comovement. Explor Econ Hist 48(2):226–242
Van Zanden JL (1991) The first green revolution: the growth of production and productivity in European agriculture. Econ Hist Rev XLIV(2):215–239
Van Zanden JL (1999) The development of agricultural productivity in Europe, 1500–1800. In: Van Bavel BJP, Thoen E (eds) Land productivity an agro-systems in the North Sea area. Middle Ages- 20th century. Elements for comparison. Brepols, Turnhout, pp 357–375

Wallis J (2018) Institutions and institutional change. In: Diebolt C, Haupert M (eds) Handbook of cliometrics, 2nd edn. Springer, Berlin/Heildelberg
Willebald H (2007) Desigualdad y especialización en el crecimiento de las economías templadas de nuevo asentamiento, 1870–1940. Rev Hist Econ/J Iber Lat Am Econ Hist XXV(2):291–345
Willebald H, Bértola L (2013) Uneven development paths among settler societies. In: Lloyd C, Metzer J, Sutch R (eds) Settler economies in world history. Brill, Leiden, pp 105–140
Williamson JG (1990) The impact of the Corn Laws just prior to repeal. Explor Econ Hist 27(2):123–156
Wright G (2006) Slavery and American economic development. LSU Press, Baton Rouge
Yasuba Y (1961) The profitability and viability of plantation slavery in the United States. Econ Stud Q 12:60–67

Nutrition, the Biological Standard of Living, and Cliometrics

Lee A. Craig

Contents

Abstract

In much of the world today, populations are richer, taller, and enjoy longer healthier lives than their counterparts in the past. Cliometricians debate the extent to which this "health transition" was the result of nutritional improvements or other factors, such as the increase in public health infrastructure that followed mastery of the germ theory of disease. Although the long-run trend in health was positive, in the nineteenth century, many Western countries experienced cyclical downturns in the biological standard of living, the so-called antebellum puzzle. While the long-run trends in the growth of real GDP, income, and wages were positive, as the presence of the antebellum puzzle suggests, the onset of industrialization was accompanied by an increase in inequality, the stagnation of the expectation of life at birth, increases in morbidity, declines in mean adult stature, and an erosion in the consumption of net nutrients. Taken together, this experience has been labeled the "Malthusian squeeze."

L. A. Craig (✉)
Department of Economics, North Carolina State University, Raleigh, NC, USA
e-mail: lacraig@ncsu.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_22

Keywords
Nutrition · Standard of living · Stature · Mortality · Morbidity · Obesity

Introduction: Nutrition and the Standard of Living

In the opening passage of his *Structure and Change in Economic History*, Nobel Laureate Douglass North states: "I take it as the task of economic history to explain the structure and performance of economies through time" (North 1981, p. 3). As an indicator of an economy's performance over time, economists typically use real gross domestic product (GDP), which is the value of final goods and services produced in an economy, usually on an annualized basis. The word "economy" in this context refers to the geographical boundaries of a nation state. Mathematically, real GDP is the sum of the product of prices (adjusted for inflation, which makes the measure "real") and the final quantities of goods and services produced; thus, GDP measures production rather than consumption.[1] Writing nearly a century and a half before the creation of the national income and product accounts, of which GDP is a key component, Adam Smith admonished his readers that, when it came to comparing the performance of one country's economy relative to that of another, "Consumption is the sole end and purpose of all production" (Smith 1976, vol. 2, p. 179). Here, Smith was grasping for a concept economists would subsequently refer to as the "living standard."

Since real GDP reflects aggregate economic activity, a geographically large, but poor, country might have a GDP that exceeded that of a small, but rich, country, thus masking the very performance North would have us explain. In response to this "size problem," economists use real GDP per capita as the typical indicator of the living standard within a modern nation state, and the growth of this measure is assumed to reflect improvements in the standard of living. Another Nobel Laureate, Simon Kuznets (1966) – referred to as the "patron saint" of national income accounting[2] – popularized the expression "modern economic growth," by which he meant real GDP growth that exceeded population growth by enough, and for long enough, that periodic downturns would not disrupt the long-run improvement in the standard of living, as measured by real GDP per capita. In Western civilization, before the onset of modern economic growth in the early nineteenth century, the long-run average annual compounded rate of growth of real GDP per capita was essentially zero (Clark 2007, p. 2). Only with the Industrial Revolution, so the argument goes, did the West escape from the Malthusian world in which short-run economic growth was eventually matched by population growth. In the United States, to offer one example of an early-developing country, real GDP per capita expanded at 1.3% annually from

[1]The difference between the production of goods and their consumption is captured, at least partly, by the "change in business inventories" component of GDP.

[2]Feldstein 1990, p. 10

1800 to 1860, 1.6% from 1860 to 1910, and it has grown at 2.0% annually over the past century.[3]

Despite the 30-fold increase in real GDP per capita over the past two centuries or so, Kuznets himself recognized that the use of GDP to measure the standard of living was not without its faults. Specifically, real GDP per capita failed to reflect the distribution of income, and it ignored many "quality-of-life issues."[4] In his Nobel lecture, Kuznets noted that industrialization, and the high rate of economic growth that came with it, could have a negative impact on the standard of living more broadly defined (1973, p. 257).

In the late 1970s, one of Kuznets's students, Robert Fogel, yet another Nobel Laureate, and some of his colleagues and students, began to explore the use of biological indicators of the standard of living, many of which were tied to nutrition (Fogel et al. 1979). Among the most prominent of these indicators were mortality and human stature. Survival, as captured by infant mortality rates and life expectancy, reflected the ultimate standard by which the quality of life could be judged. Stature represents a more subtle measure. The consumption of net nutrients, i.e., those beyond that exhausted during work or fighting disease, determines whether an individual will achieve a genetically programmed height potential and, more indirectly, life expectancy. Although like real GDP per capita, stature allows researchers to characterize economic performance, it differs from GDP in that stature more directly reflects consumption, specifically the consumption of nutrients, and the physical costs (i.e., work and disease) associated with the productive activities captured by GDP. As such, it better reflects the distribution of consumption and living standards.

Of course, the consumption of net nutrients is also affected by the time and intensity of labor, as well as by working and living conditions. Physically demanding occupations and those that require more intense effort increase these demands and thus leave fewer nutrients for growth. Disease also imposes demands on the body, and it spreads more easily in urban and industrial environments relative to rural and agricultural ones. It follows that mean adult stature offers a valuable indicator of a population's biological standard of living. In other words, it is a "cumulative indicator of net nutritional status over the growth years" (Cuff 2005, p. 10).

Today, as measured by real GDP per capita, populations in developing countries are richer, on average, than their counterparts in the past; they are taller; and they enjoy longer healthier lives. Fogel and Costa (1997) and Floud et al. (2011) characterize the increase in stature and the reduction in mortality that accompanied it as the product of the "techno-physio evolution," and Costa (2013) refers to the extension of the life span and improvements in health as the "health transition." This chapter reviews the research on nutrition's role in these important changes.

[3] Williamson 2013

[4] It also omits many types of economic activity, though this was not a point of emphasis for Kuznets.

Nutrition, the Health Transition, and the Techno-Physio Evolution

The expression "health transition" refers to improvements in a set of outcomes that exemplify the health of a population. One body of research on the transition has focused on the reduction in age-specific mortality, which resulted in an increase in life expectancy. From the late nineteenth century through the middle of the twentieth century, among the countries that experienced an increase in life expectancy, the largest decreases in mortality occurred at the youngest ages. As Haines and Steckel observe, "The reduction in mortality reflected largely a sharp decrease in mortality due to infectious diseases. . .the young, especially infants and children, who had suffered the highest fatality rates from many of these diseases, were the principal beneficiaries of the sharp reduction in their incidence" (2000, p. 647). Stolnitz (1955) was among the first to argue that, because the mortality declines were geographically widespread, at the macro-level, the mastery of the germ theory of disease, rather than nutritional improvements, must be among the causes of the decline. This led scholars who followed Stolnitz to largely attribute the increase in life expectancy to improvements in health-care technologies including both science-based medical care, such as vaccinations, and public health infrastructure (Preston 1975).[5] Subsequent micro-level research on clean water and sewer systems, to offer examples of the type of infrastructure likely to reduce mortality, supported this position (Troesken 2004). Deaton (2003, 2006) supports it further by noting that "it is clear that public health measures, particularly the provision of clean water and better sanitation. . .were the fundamental forces for mortality reduction during the century from 1850 to 1950" (2006, p. 110).

The mortality declines resulting from mastery of the germ theory of disease leave in question the net impact of these changes on the overall morbidity of a society. The early improvements in public health technologies largely eliminated infectious disease as a killer among the young, which, holding other factors constant, puts downward pressure on overall morbidity while increasing life expectancy. At the same time, the increase in life expectancy increased the incidence of chronic morbidity among the (now more numerous) elderly, leaving the net impact on morbidity ambiguous, a debate about which we will say more below.

In contrast to the view that the health transition was largely one of technological change in medical care and public health, which was supported primarily by public spending, McKeown (1976) points out that much of the decrease in mortality came before the major medical breakthroughs, and so he places a greater weight on the role of nutritional improvements, which occurred earlier than those associated with medical care, a point conceded, if grudgingly, by Deaton (2006, p. 110).[6]

[5]It was this article that generated the famous "Preston curve," which shows life expectancy as a function of income (or real GDP) per capita. The first derivative of the resulting function is positive, but the second is negative, suggesting that, beyond some point, increases in income do not lead to further increases in life expectancy.

[6]To be fair, Preston also mentioned nutrition as a causal element; to wit, "Income, food, and literacy were unquestionably placing limits on levels of life expectancy. . ." (1975, p. 240).

McKeown's position is supported by Fogel (2004) who emphasizes the "synergism" between nutrition and income: More net nutrition yields more productive workers, who in turn earn higher incomes, which are used to purchase more nutrition and so forth. Furthermore, as Floud et al. (2011) note, there is an intergenerational component to the synergism between nutrition and income, arguing that the nutritional status of one generation determines the life span and productivity (and hence income) of the members of that generation, which in turn determines, at least partly, the nutrition, life span and productivity of the next generation, and so on.

This "supremacy of nutrition" argument forms the foundation of the explanation in Floud et al. of the "techno-physio evolution." It is based on what Komlos refers to as *Homo sapiens*' "evolutionary advantage," which is humankind's capacity to "adapt to the availability of nourishment" (2012, p. 4). In short, Floud et al. and Komlos reject the Malthusian rhetoric that framed much of the earlier debate on these issues. The standard juxtaposition of a Malthusian world, with high death rates, no understanding of the germ theory of disease, and no public health spending or technologies to speak of, with a post-Malthusian world, with low death rates, a mastery of the germ theory, and extensive public works projects aimed at public health, is a false one. Nutrition is the key to the techno-physio evolution and health transition because death was not the only, or even the main, biological response to nutritional want. Rather than having a mass die-off in the face of want, humans could simply be smaller and less healthy.

Indeed, by detailed estimation of the supply and demand of nutrients in the past,[7] and the use of Waaler curves – which plot height, weight, and mortality risk – Floud et al. show that populations in the past adjusted their size to the availability of nutrients. But it is important to note that being smaller also meant being weaker, sicker, and more likely to die earlier. As one reviewer summarized this position, "In other words, smaller was not equal. Malthus was right to focus on the misery of past populations. His model just did not well reflect all of the dimensions of that misery" (Craig 2013, p. 114).

The late nineteenth- and early twentieth-century period was also one marked by Kuznets's modern economic growth, as well as the reasonably steady growth of agricultural production in much of the early-developing world. The synergy between these trends – which resulted in the reduction in morbidity and mortality, as well as the growth of labor productivity and incomes from the Industrial Revolution (and some scholars would argue an agricultural revolution that either preceded or accompanied it) – explains the techno-physio revolution and much of the health transition. However, a challenging "food puzzle" remains to be explained. According to Clark et al. (1995), the puzzle is that, between 1750 and 1850, a period during which the techno-physio evolution was underway, the British population and incomes increased much more rapidly than food production. In evaluating various explanations of the puzzle, Clark et al. focus on changing patterns in consumption. Specifically, they argue that as populations became more urban, they increased their

[7]Which involves allocating nutrition across various physical activities.

consumption of certain processed foods, including alcohol, tea, and sugar, and their solution to the puzzle rests on a combination of variations in the income elasticity of food and changes in the relative prices of various foods and beverages. In contrast, Floud et al. argue that the income elasticity employed by scholars looking at the puzzle is too high, because, during the early phase of industrialization, incomes increased more rapidly than the physically smaller populations at that time could increase their consumption of nutrients. That is, they were too small to consume enough additional nutrition to justify spending their – now higher – incomes on food. It follows that the high-earning urban workers and their families increased their consumption of other goods, including alcohol, caffeine, and sugar.

These issues have public policy implications. For those who place the weight on public-sector spending for scientific medical research and public health infrastructure, the government, rather than the market, is the key to increasing the biological standard of living. Easterlin notes that, in societies that eventually broke out of the Malthusian trap, historically, "free market institutions have functioned poorly to control major infectious diseases" and hence mortality (2004, p. 125). Floud et al. do not disagree that government played an important, and positive, role in improving the consumption of nutrients. They argue that from the dawn of the early modern era, famines in the West were "man-made rather than natural disasters." Nutritional crises occurred "not because there was not enough grain to go around, but because the demand for inventories pushed prices so high that laborers lacked the cash to purchase the grain."[8] Governments lessened the severity of the crises by intervening in the markets for food, primarily through price controls and the pooling of output in public granaries. Thus, the authors conclude that "By the start of the nineteenth century, famines had been conquered in England, not because the weather had shifted, or because of improvements in technology, but because [of] government policy. . . ."[9]

The Biological Standard of Living and the Antebellum Puzzle

The debates concerning the source of the techno-physio evolution and the health transition tend to focus primarily on the long-run trends in the underlying indicators. However, a number of Western countries experienced cyclical downturns in the biological standard of living during the nineteenth century. In the United States, for example, the mean adult stature of native-born white males declined by roughly 2.5 cm between 1800 and 1860 (Haines et al. 2003), a period during which real per capita GDP grew at a robust rate by historical standards (1.3% per annum). Mean stature further eroded after 1870, bottoming out in 1880 (Treme and Craig 2013). Margo and Steckel (1983) and Komlos (1987) were among the first to explore this

[8]Floud et al. 2011, p. 117

[9]Floud et al. 2011, pp. 116–118

deviation from the long-run trend in the biological standard of living, an event which Komlos and Coclanis (1997) labeled the "antebellum puzzle."

The discovery of the erosion of the biological standard of living in the United States during the early decades of the nineteenth century led scholars to explore the trends and cycles in other countries that were experiencing the onset of modern economic growth during the same period. These investigations found that, as was the case in the United States, the behavior of height was cyclical in many countries. A common pattern was that mean adult stature increased early in the nineteenth century, decreased mid-century, and began to increase again near the end of the century. The magnitude of the downturns varied from country to country. In the United Kingdom, the difference in adult stature between birth cohorts of 1810 and 1850 was almost 5.8 cm; the Netherlands experienced a downturn of 2.5 cm between 1830 and 1860; in Denmark the decline was 1.7 cm between 1820 and 1850; in Sweden the decline was 0.3 cm between 1830 and 1840 (Treme and Craig 2013, Fig. 1).[10]

Explanations of the antebellum puzzle focus on four factors: (1) a decline in the mean consumption of net nutrients, (2) a growing inequality in the distribution of income, (3) Kuznets's "negative results" associated with industrialization (including an increase in the intensity of work) and urbanization, and (4) a deterioration in the epidemiological environment resulting from improvements in transportation and urbanization.[11] The last two of these we consider below, in the section on technological change; however, items (1) and (2) we consider here.

With respect to the trend in net nutrition, in his path-breaking study of the nutritional status of West Point cadets, Komlos argued that "nutritional intake was declining [in the United States] in the late antebellum period. The availability of nutrients declined because food output did not keep pace with the demands placed upon it. . ." (1987, pp. 909–910). He estimated that the per capita consumption of nutrients declined by roughly 10% during the 1840s (1987, Table 8). That position was subsequently challenged by Gallman, who argued that "diet surely improved – and importantly – between the 1820s and the1850s" (1996, p. 199). Following subsequent debate in the literature, Komlos and Coclanis (1997) showed that an important feature of the period was a change in the relative prices of nutrients, which was caused by the commercialization of agriculture and which in turn caused households to substitute out of the consumption of meat, a key source of protein, and into carbohydrates. Evidence provided by Craig and Weiss (1997) on the commercialization of agriculture, and Craig and Hammond (2013) on the allocation of nutrients between free and slave populations, is broadly supportive of the Komlos and Coclanis argument.

As for the argument that the distribution of income became more unequal during the period, there are two related issues: First, did income in fact become more

[10]Interestingly, US slaves and the very rich did not experience a downturn in the biological standard of living (Craig and Hammond 2013; Sunder and Woitek 2005).

[11]This list is broadly consistent with, though not identical to, that found in Komlos (1987, p. 905).

unequally distributed, and second, if it did, then how exactly did that fact manifest itself in the erosion of the net nutrition of the US population? The evidence collected and analyzed by cliometricians would seem to support the proposition that wealth (and probably income) became less equally distributed over the course of the nineteenth century (Lindert and Williamson 1980, pp. 33–95). Of course, that observation still leaves one with the question of how the growing inequality led to the erosion of net nutrition for much of the population. Although the income elasticity of food tends to be relatively low, it is positive, and mean incomes and wealth were rising during the period; thus, food intake should have increased, as long as the food supply kept pace, which Komlos argues it did not. While that issue remains much debated, less controversial is the Komlos-Coclanis argument that changes in the relative price of nutrients, with that of meat and dairy products outpacing grains, caused farm families to substitute out of protein-laden products, which were more likely to promote adult stature, and into grains. In short, agricultural output was increasing; the prices of farm products were falling; and mean wealth and income, as measured by real GDP per capita, were rising, but this increase was disproportionately enjoyed by those in the richer tails of the distributions of wealth and income. For those further down the socioeconomic ladder, changes in the relative price of nutrients (increasing for protein-heavy foods, decreasing for carbohydrate-rich foods) caused the substitution effect to overwhelm the income effect and, for much of the population, led to a reduction of the net consumption of nutrients. Thus, the changing distribution of income and the changing relative price of nutrients together countered the positive impact of modern economic growth and led to the erosion of the biological standard of living.

Nutrition, Stature, and Income

Komlos's path-breaking microlevel study of the heights and weights of West Point cadets highlighted the relationship between nutrition, stature, and real output. Two important questions emerged from that study. One concerned the empirical relationship between net nutritional intake during the growth years and adult stature: Specifically, at the margin, how much net nutrition would be necessary to generate an additional centimeter in adult stature? The second was, again at the margin, how much additional real output would that marginal centimeter yield?

Unfortunately, the scarcity of data that would allow the matching of access to nutrients with individual consumers prohibits a detailed study of the type that would be ideal for answering those two questions. However, employing the sample of Union Army recruits compiled by Fogel et al. (1979) and microlevel data from the US Censuses of Population and Agriculture, Haines et al. (2003) were able to match the availability of nutrients in the county in which a recruit was born, and other variables, with the recruit's adult stature. Their argument was that recruits who grew up in counties that generated a "surplus" of food were more likely to have access to nutrients during their growth years and therefore should be taller as adults. The model they specified was:

$$\text{Height}_i = f(\text{Nutrition}_i, X_i) + \varepsilon_i \quad (1)$$

where Height_i is the adult stature of the ith Union Army recruit, Nutrition_i is the daily nutritional surplus generated in the county in which the recruit was born,[12] X_i is a vector of other location or individual-specific variables that would be expected to influence adult stature, and ε_i is an error term distributed $(0, \sigma^2)$.[13]

When it comes to generating stature, some nutrients are better than others. Baten and Murray (2000) show that protein is a particularly important component of the nutrition-stature relationship. Thus, Haines et al. focused on surplus protein production as their measure of access to nutrition. Their results suggest that being born in a county that generated a daily agricultural surplus equal to one-half of one standard deviation above the mean for US counties in 1840 (roughly 35 g of protein per capita, the equivalent of 120 g of beef or ten slices of whole wheat bread) would have yielded an additional 0.125 cm in adult stature (Haines et al. 2003, p. 405).

As for the relationship between income and stature, using national-level data, Steckel (1995, p. 1914) estimated the following equation:

$$\text{Height}_j = g(\text{Income}_j, Y_j) + \mu_j \quad (2)$$

where Height_j is the mean adult stature of the jth country, Income_j is per capita income in the jth country, Y_j is a vector of other country-specific variables, and ε_i is an error term distributed $(0, \sigma^2)$.[14]

The results indicate that, at the mean, the height elasticity of income is between 0.20 and 0.25. Craig et al. (2004) inverted this relationship to find the impact of stature on per capita income. Combining the coefficients from Haines et al. and Steckel, Craig et al. estimate that an additional 3.5 g of protein per capita per day, roughly a slice of whole wheat bread, would have generated an additional 0.0125 cm in mean adult stature, which would have in turn generated an additional 0.05% in national income (more than $2,000 per person in the United States today).[15] It is important to note that this estimate represents a permanent increase in income, not simply a one-time shock. The long-run impacts on the standard of living from these improvements in nutrition are arguably quite large by any reasonable comparison.

[12]As defined by Atack and Bateman (1987), this measure represents the surplus of nutrients beyond that consumed by the county's human and livestock populations.

[13]The other variables reflect the impacts of urbanization, the individual's occupation, the wealth of the region in which he grew up, and a dummy variable capturing whether or not the region had access to water or rail transportation.

[14]The other variables include or reflect the impacts of wealth, region, urbanization, and the age distribution of the population.

[15]Steckel estimated stature as a function of income; thus, this coefficient represents the inverse of his estimate.

Nutrition, Mortality, and Morbidity

A considerable body of research links the health transition with the mortality transition. Specifically, the expectation of life has increased dramatically in both early- and later-developing countries. (In the United States, for whites, it has gone from 39.5 in 1850 to 78.9 today, whereas in Mexico, to offer an example of a later-developing country, the figure has risen from 25.3 to over 70 today (Haines and Steckel 2000, pp. 696–698)). As noted, McKeown (1976) emphasized the role of nutrition in this improvement, and the findings of subsequent microlevel research indicate that historically there was a strong relationship between nutrition and longevity (Cuff 2005; Haines 1996). For example, US counties that generated agricultural surpluses, as defined above, had mortality rates that were 30% lower than those in deficit counties (Haines et al. 2003, p. 394).

Despite the unambiguous long-run positive trend in life expectancy, just as was the case with stature, life expectancy declined across many countries in the mid-nineteenth century. In the United States, life expectancy at age 20 declined by roughly 6 years during the first half of the century (Pope 1992), and, again, as was the case with stature, at least some of this decline was associated with the erosion in the consumption of net nutrients (Floud et al. 2011; Komlos 1987, 1996).

The scholarly work on trends and cycles in mortality should be reviewed in light of related work on morbidity. Research on morbidity tends to follow one of two paths. One path traces the mortality-disease nexus. Into the nineteenth century, outbreaks of infectious diseases were often serious enough and/or widespread enough to bring about substantial increase in death rates. According to Flinn, bubonic plague was the most effective and persistent killer, with outbreaks that "moved around Europe throughout most of the [early-modern] period. . .there were few years when the disease was not attacking somewhere" (1981, p. 51). Epidemics that resulted in an increase in the death rate tended to disrupt the economy directly through the labor market, in which the lost production of the sick and the dead reduced output. Thus, famine could follow plague, even without a harvest failure. Severe epidemics also disrupted trade. Disease often entered a region through port cities, and the subsequent damage to trade – primarily through lost labor, but also through the isolation that came from quarantine and fear – could be substantial (Craig and Garcia-Iglesias 2010).

The negative shocks of severe epidemics are relatively easy to identify, and on occasion, they could cause dramatic increases in mortality rates. However, by the mid-nineteenth century, other than the occasional local outbreak of, for example, cholera or yellow fever, food supplies were secure enough (and the appreciation of public health measures were widespread enough) that, at least in the West, disease- and famine-caused mortality crises were a thing of the past. As such, for most of the population, morbidity became a chronic, rather than an immediately life-threatening, concern. Using microlevel data from British friendly societies, Riley (1990, 1997) found that the prevalence of morbidity increased with the decline in mortality that accompanied modern economic growth. One interpretation of Riley's finding is that the decline in mortality was accompanied by longer – albeit, on an individual basis,

fewer – episodes of illness, as the health-care industry evolved to help workers better manage their illnesses. Thus, the implication of Riley's research is that overall, the health of Western societies has deteriorated over time as improvement in public and private health technologies have allowed the sick to survive illness that in the past would have killed them.

This conclusion is not without controversy. Not only does it conflict, at least qualitatively, with McKeown's argument that nutrition, rather than public health spending or improvements in the provision of private health care drove the health transition, it also conflicts with much recent empirical work on morbidity. For example, Dora Costa surveys a wide range of studies across several countries that, collectively, show an unambiguous improvement in the health of Western populations since the late nineteenth century. Comparing cohorts from the US Civil War with those from the late twentieth century, Costa (2000) and Costa et al. (2007) show that the incidence among the elderly of a wide set of conditions declined over time, and comparing early twentieth-century populations with those from later in the century, she notes that "mothers of the children born in the 1910s and 1930s were shorter, showed signs of malnutrition, had high blood pressure during pregnancy, and were more likely to be syphilitic than mothers who gave birth in the 1960s and in 1988" (2013, Table 3).

There is a fundamental mathematical element that lies at the heart of these positions, which, it should be noted, are not all mutually exclusive. If we think of disease as a state that lasts through time, then the prevalence of morbidity for an individual can be expressed as:

$$M_i = \int_0^T h(t)\, dt \tag{3}$$

where M_i is the amount of time over the course of the ith individual's life that she spends sick, h(t) characterizes her path of illness through time, and T is the length of her life. Assuming h can be expressed as a function of, among other things, nutrition and public and private health-care technologies, then McKeown's position can be summarized as follows: A decrease in h, largely caused by nutritional improvements, in turn led to an increase in T, leaving the net impact on lifetime morbidity, M, ambiguous. In short, over time, people ate better and lived longer, but that increased their exposure to disease. Riley's position can be interpreted as the long-run increase in T, a result of modern economic growth and improvements in health-care technologies, unambiguously increased lifetime exposure to morbidity. In short, as people lived longer, they spent more time sick, though the doctors got better at treating their illnesses. Finally, Costa's position can be interpreted as follows: Despite the increase in T, improvements in nutrition and health-care technologies unambiguously decreased the prevalence of morbidity. In short, people now enjoy longer *and* healthier lives than they did in the past.

In related research connecting morbidity directly to mortality, Alter and Riley (1989) offer an "insult accumulation model," in which exposure to disease – i.e., an

"insult" – weakens survivors in ways that cannot be easily erased through subsequent nutritional gains; thus, survivors are susceptible to future insults and higher mortality rates than would have been the case in the absence of the insult. Using microlevel data from the Union Army sample, Lee (2003) challenges this view. He finds that adults raised in locales with high childhood mortality rates, such as urban areas, had lower mortality rates while they were in the army than those from rural areas. This finding indicates that surviving prior insults gave one an advantage – rather than, as Alter and Riley suggest, a disadvantage – when faced with future insults.

Nutrition and Technological Change

Technological change has played an important, though often confounding, role in the techno-physio revolution and the health transition. At its most basic level, the nutritional access of a population is tied to the food supply from either domestic production or trade. Most of the world's population today no longer faces the food crises that kept population growth in check for millennia. The average annual compounded growth rate of the world's population over the past five centuries is an order of magnitude larger than it was over the prior 2000 years, 0.71% versus 0.03%.[16] This is largely the result of a series of technological changes in agriculture. Often referred to as "revolutions" in the literature, by some accounts, the first of these occurred before the Industrial Revolution and involved countless small-scale improvements by small-holding, open-field farmers. As evidence in support of this interpretation, Allen (1999) estimates that farm output in England doubled between 1520 and 1740. A second revolution, this one revolving around the mechanization of agriculture, occurred in the mid- to late nineteenth century. Craig and Weiss (2000) estimate that in the United States, average annual total factor productivity growth, a standard economic indicator of the pace of technological change, expanded from near zero in the two decades before 1860 to roughly 0.70% in the four decades after that date. Finally, in the twentieth century, the so-called Green Revolution, which was marked by the development and expanding use of chemical pesticides, herbicides, and fertilizers, increased agricultural output in much of the world. Whereas in 1500, the vast majority of the world's labor force was directly employed in the production of food and fiber; in modern developed economies, only 1% or 2% is so employed today, and in many countries, the most prominent nutritional problem is a surplus of food leading to an epidemic of obesity (Komlos et al. 2008).

Holding other factors constant, in the long run, the resulting increase in food production unambiguously reduced mortality and morbidity, though as the antebellum puzzle suggests, the positive trend was not without a cyclical component. Transportation improvements, also frequently referred to as a "revolution," proved to be an ambiguous influence on the relationship between the increase in food

[16]United Nations 1999, Table 1

production and the biological standard of living. Technological change in transportation reduces costs and results in an increase in the amount of goods shipped, which would tend to increase total consumption, including the consumption of nutrients. However, transportation also had two negative impacts on net nutrition.

As Komlos and Coclanis note, in the nineteenth century, transportation improvements were accompanied by urbanization, which increased the demand for food supplied to the market, which in turn impacted the nutritional status of both urban and farm populations. In urban areas, there were now more people purchasing food in the market, and although the per unit cost of transporting food declined, "a larger share of the population had to pay transportation costs in order to obtain its nutrients . . ." (1997, p. 448). On the farm, the commercialization of agriculture led to an increase in the production and export of cash crops, most notably cotton in the South, at the expense of a healthier more diverse diet that had previously been heavy in animal-based protein.

> The price of meat and dairy products rose with the distance from the point of production, thereby impinging on the amounts demanded. Hence, a decline in the number of livestock per capita implies that the consumption of an important determinant of nutritional status, animal protein, was declining . . . Cash, in other words, did not automatically translate into higher nutritional status in the early industrial era . . . . (1997, pp. 447–448)

So both urban and rural populations suffered nutritionally as a result of these changes.

The other problem transportation improvements caused for the biological standard of living was an expansion of the disease nexus. Urbanization, the expansion of which was facilitated by the industrial and transportation revolutions, was strongly correlated with increases in mortality rates and the erosion in other indicators of the biological standard of living. In their study of the antebellum puzzle, Haines et al. (2003) estimate that in the United States, 10 percentage point increase in a county's urbanization rate resulted in a 7.5% increase in the crude death rate, and US Army recruits from urban areas were, on average, an inch shorter than those from rural areas. Supporting that finding more broadly, Fogel (1986) estimates that urbanization explains approximately 20% of the US stature decline during the same time period.

In addition, these changes also led to an increase in the intensity of farm labor. Average hours at work in agriculture increased during the period in question (Craig and Weiss 2000). Recall that it is net nutrition that allows the body to prosper, and the body consumes nutrients while fighting disease and exerting the energy needed for work. The expansion of the disease nexus and the increase in hours at work associated with the commercialization of agriculture, all of which were at least partly dependent on improvements in transportation, led to an increase in the body's demand for nutrients at a time when, for much of the population, their supply was increasingly challenged.

There was one technological innovation that unambiguously led to an improvement in the standard of living, as traditionally measured by income, as well as the biological standard of living: mechanical refrigeration. The physics of refrigeration

had been well understood thousands of years before the widespread adoption of mechanical refrigeration in the late nineteenth century and early twentieth century. The first US patent for a mechanical refrigerator was issued in 1853. The early machines were too costly to build and maintain, and too unreliable, to be widely adopted. Only after a set of "rather mundane improvements in the machine-tool industry, related metallurgical improvements, the development of high-pressure seals, and the addition of the electric motor" were perfected and adopted was a low-cost refrigerator feasible (Goodwin et al. 2002). All of these changes came together near the end of the nineteenth century.

Refrigeration smoothed the seasonal price and supply swings that had plagued agricultural markets since time immemorial, but it also allowed farmers to maintain herds beyond the slaughter season, and without reducing slaughter rates, overall herd sizes could be increased: "In short, refrigeration did more than simply allow a farmer to hold a hog off the market today for future slaughter. It allowed the farmer to slaughter a hog today *and* hold another hog off the market for later slaughter" (Craig and Holt 2008, p. 111). This had an impact on the overall production of meat and dairy products, which together represented roughly 30% of total agricultural output, and agriculture was 25% of GDP in 1900. Craig and Holt estimate that the market value of agricultural output increased by 2.28% as a result of the adoption of mechanical refrigeration, which would represent a 0.17% increase in GDP. With respect to nutrition, Craig et al. estimate that mechanical refrigeration led to a 0.75% increase in the consumption of calories, a 1.25% increase in the consumption of protein, and a 1.26% increase in household incomes (2004, p. 333). Note that these were permanent additions to the long-run growth paths of these variables, not one-time increases. As Gordon (2000) observed, refrigeration was truly one of the "great inventions."

The debates surrounding the behavior of the biological standard of living during industrialization are part of a larger debate among cliometricians concerning the impact of the Industrial Revolution. Broadly speaking, scholars who focus on the long-run positive trends in the growth of real GDP, income, and wages have been labeled "optimists," whereas those who focus on the increase in inequality, stagnation of the expectation of life at birth, increases in morbidity, declines in mean adult stature, and erosion in the consumption of net nutrients have been labeled "pessimists" (Craig 2006). After several decades of cliometric debate, it is clear that a long-run view of these changes, where long-run here means a century or so after the onset of industrialization, leads to an optimistic conclusion. However, viewed from the perspective of many, perhaps most, of those who lived through the transition from the world of Malthus to the Gilded Age, there was probably a deterioration in the biological standard of living, an experience Haines et al. labeled the "Malthusian squeeze" (2003, p. 408).

Conclusion

Today, populations in the early-developing countries are richer, taller, and enjoy longer healthier lives than their counterparts in the past. Cliometricians have debated the extent to which the techno-physio evolution and the health transition resulted

from nutritional improvements or other factors, such as the increase in public health infrastructure that followed mastery of the germ theory of disease. Although the long-run trends in the growth of real GDP, income, wages, and the biological standard of living were positive, as the presence of the antebellum puzzle suggests, the onset of industrialization was accompanied by an increase in inequality, the stagnation of the expectation of life at birth, increases in morbidity, declines in mean adult stature, and an erosion in the consumption of net nutrients. Taken together, this experience has been labeled the Malthusian squeeze.

Whereas much of the cliometric research on nutrition focuses on its role in the techno-physio revolution and the health transition, modern developed societies face an entirely different nutritional issue: obesity. As Treme and Craig note, "There is a biological maximum to the mean stature of a population, and for those populations enjoying a surplus of nutrients, further consumption would merely lead to obesity" (2013, p. s131). Although the phenomenon is widespread among rich and not-so-rich societies, the United States offers perhaps the most striking example of the trend. Over the past 30 years, the incidence of obesity has doubled, and roughly one in three US adults is obese, as measured by BMI (Flegal et al. 2012). The social cost of this epidemic is enormous. Cawley and Meyerhofer (2012) estimate that more than 20% of US health-care expenditures are directly attributable to obesity.

Despite much recent study, the causes of mass obesity are not well understood. Cawley et al. (2010), employing cross-sectional econometric work with instrumental variables, show that income differences explain only a very small component of weight differences. This finding suggests that one possible culprit of the rise of obesity, the modern economic growth that helped developing countries generate the techno-physio revolution and escape from the Malthusian world, can be ruled out as the cause of obesity. Komlos et al. (2008) suggest that modern consumer technologies, such as the television and automobile, as well as labor-saving production technologies, explain the rising prevalence of obesity. In any case, the topic represents the next frontier in empirical research on nutrition and its economic and social impacts.

References

Allen R (1999) Tracking the agricultural revolution in England. Econ Hist Rev 52(2):209–235

Alter G, Riley J (1989) Frailty, sickness and death: models of morbidity and mortality in historical populations. Popul Stud 43(1):25–45

Atack J, Bateman F (1987) To their own soil: agriculture in the antebellum north. Iowa State University Press, Ames

Baten J, Murray JE (2000) Heights of men and women in 19th-century Bavaria: economic, nutritional, and disease influences. Explor Econ Hist 37(4):351–369

Cawley J, Meyerhofer C (2012) The medical care costs of obesity: an instrumental variables approach. J Health Econ 31(1):219–230

Cawley J, Moran JR, Simon KI (2010) The impact of income on the weigh of elderly Americans. Health Econ 19(8):979–993

Clark G (2007) A farewell to alms: a brief economic history of the world. Princeton University Press, Princeton

Clark G, Huberman M, Lindert P (1995) A British food puzzle, 1770–1850. Econ Hist Rev 48(2):215–237
Costa D (2000) Understanding the twentieth century decline in chronic conditions among older men. Demography 37(1):53–72
Costa D (2013) Health and the economy in the United States, from 1750 to the present. NBER working paper no 19685, © National Bureau of Economic Research
Costa D, Helmchen L, Wilson S (2007) Race, infectious disease, and the arteriosclerosis. Proc Natl Acad Sci 104:13291–13224
Craig LA (2006) A review of Timothy cuff's the hidden cost of economic development: the biological standard of living in antebellum Pennsylvania. Reviewed for Eh.net
Craig LA (2013) The changing body: health, nutrition, and human development in the western world since 1700: a review essay. Econ Hum Biol 11(1):113–116
Craig LA, Garcia-Iglesias C (2010) Business cycles. In: Broadberry S, O'Rourke K (eds) An economic history of modern Europe: vol 1: 1700–1870. Cambridge University Press, Cambridge, pp 122–146
Craig LA, Hammond R (2013) Nutrition and signaling in slave markets: a new look at a puzzle within the antebellum puzzle. Cliometrica 7(2):189–206
Craig LA, Holt M (2008) Did refrigeration kill the Hog-Corn cycle? In: Rosenbloom J (ed) Quantitative economic history: the good of counting: essays in honor of Thomas Weiss. Routledge, London, pp 100–118
Craig LA, Weiss T (1997) Long-term changes in the business of farming: hours at work and the rise of the marketable surplus. Paper presented at the international business history conference, Glasgow, July
Craig LA, Weiss T (2000) Hours at work and total factor productivity growth in 19th-century U.S. agriculture. Adv Agric Econ Hist 1(1):1–30
Craig LA, Goodwin B, Grennes T (2004) The effect of mechanical refrigeration on nutrition in the United States. Soc Sci Hist 28(3):325–336
Cuff T (2005) The hidden cost of economic development: the biological standard of living in antebellum Pennsylvania. Ashgate, Aldershot
Deaton A (2003) Health, inequality, and economic development. J Econ Lit 41(1):113–158
Deaton A (2006) The great escape: a review of Robert Fogel's the escape from hunger and premature death, 1700–2010. J Econ Lit 44(1):106–114
Easterlin R (2004) The reluctant economist: perspectives on economics, economic history and demography. Cambridge University Press, Cambridge, UK
Feldstein M (1990) Luncheon in honor of individuals and institutions participating in the first income and wealth conference. In: Berndt ER, Triplett JE (eds) Fifty years of economic measurement: the jubilee of the conference on research in income and wealth. University of Chicago Press, Chicago, pp 9–18
Flegal KM, Carroll MD, Kit BK, Ogden CL (2012) Prevalence of obesity and trends in the distribution of BMI among U.S. adults, 1999–2010. JAMA 307(5):E1–E7
Flinn MW (1981) The European demographic system, 1500–1820. Johns Hopkins, Baltimore
Floud R, Fogel RW, Harris B, Hong SC (2011) The changing body: health, nutrition, and human development in the western world since 1700. Cambridge University Press, Cambridge
Fogel R (1986) Nutrition and the decline in mortality since 1700. In: Engerman S, Gallman R (eds) Long-term factors in American economic growth. University of Chicago Press, Chicago, pp 439–556
Fogel R (2004) The escape from hunger and premature death, 1700–2100: Europe, America and the third world. Cambridge University Press, Cambridge, UK
Fogel R, Costa D (1997) A theory of the technophysio evolution, with some implications for forecasting population, health care costs, and pension costs. Demography 34(1):49–66
Fogel RW, Engerman SL, Floud R, Steckel RH, Trussell J, Wachter KW, Margo R, Sokoloff K, Villaflor G (1979) The economic and demographic significance of secular changes in human stature: the U.S. 1750–1960. NBER working paper, © National Bureau of Economic Research
Gallman R (1996) Dietary change in antebellum America. J Econ Hist 56(1):193–201

Goodwin B, Craig LA, Grennes T (2002) Mechanical refrigeration and the integration of perishable commodity markets. Explor Econ Hist 39(2):154–182
Gordon R (2000) Does the "new economy" measure up to the great inventions of the past? NBER working paper no 7833, © National Bureau of Economic Research
Haines MR (1996) Estimated life tables of the United States, 1850–1900. Hist Methods 32(4):149–169
Haines MR, Steckel R (2000) A population history of North America. Cambridge University Press, Cambridge
Haines MR, Craig LA, Weiss T (2003) The short and the dead: nutrition, mortality, and the 'antebellum puzzle' in the United States. J Econ Hist 63(2):385–416
Komlos J (1987) The height and weight of west point cadets: dietary change in antebellum America. J Econ Hist 47(4):897–927
Komlos J (1996) Anomalies in economic history: toward a resolution of the "antebellum puzzle." J Econ Hist 56(1):202–214
Komlos J (2012) A three-decade "Kuhnian" history of the antebellum puzzle: explaining the shrinking of the US population at the onset of modern economic growth. University of Munich discussion papers in economics 2012–10. http://epub.ub.uni-muenchen.de/12758/
Komlos J, Coclanis P (1997) On the 'puzzling' antebellum cycle of the biological standard of living: the case of Georgia. Explor Econ Hist 34(4):433–459
Komlos J, Breitfelder A, Sunder M (2008) The transition to post-industrial BMI values among US children. NBER working paper no 13898, © National Bureau of Economic Research
Kuznets S (1966) Modern economic growth: rate, structure and spread. Yale University Press, New Haven
Kuznets S (1973) Modern economic growth: findings and reflections. Am Econ Rev 63(3):247–258
Lee C (2003) Prior exposure to disease and later health and mortality: evidence from civil war medical records. In: Costa DL (ed) Health and labor force participation over the life cycle. University of Chicago Press, Chicago, pp 51–88
Lindert PH, Williamson JG (1980) American inequality: a macroeconomic history. Academic, New York
Margo R, Steckel R (1983) The heights of native-born whites during the antebellum period. J Econ Hist 43(1):167–174
McKeown T (1976) The modern rise of population. Arnold, London
North D (1981) Structure and change in economic history. W.W. Norton, New York
Pope C (1992) Adult mortality in America before 1900: a view from family histories. In: Goldin C, Rockoff H (eds) Strategic factors in nineteenth-century American economic history. University of Chicago Press, Chicago, pp 267–296
Preston S (1975) The changing relation between mortality and level of economic development. Popul Stud 29(2):231–248
Riley J (1990) The risk of being sick: morbidity trends in four countries. Popul Dev Rev 16(3):403–432
Riley J (1997) Sick, not dead: the health of British workingmen during the mortality decline. Johns Hopkins, Baltimore
Smith A (1976) An inquiry into the nature and causes of the wealth of nations. University of Chicago Press, Chicago
Steckel RH (1995) Stature and the standard of living. J Econ Lit 33(4):1903–1941
Stolnitz G (1955) A century of international mortality trends: I. Popul Stud 9(1):24–55
Sunder M, Woitek U (2005) Boom, bust, and the human body: further evidence on the relationship between height and business cycles. Econ Hum Biol 3(3):450–466
Treme J, Craig LA (2013) Urbanization, health, and human stature. Bull Econ Res 65(S1): s130–s141
Troesken W (2004) Water race and disease. MIT Press, Cambridge
United Nations (1999) The world at six billion. http://www.un.org/esa/population/publications/sixbillion/sixbilpart1.pdf. Accessed 17 Jan 2014
Williamson SH (2013) The annual real nominal GDP for the United States, 1790–2012, MeasuringWorth.com, August. http://www.measuringworth.com/usgdp/. Accessed 20 Jan 2014

Improvements in Health and the Organization and Development of Health Care and Health Insurance Markets

Gregory T. Niemesh and Melissa A. Thomasson

Contents

Abstract

This chapter describes the gains in health in the twentieth century and the development of the markets of health care and health insurance. It first provides an overview of the literature documenting the gains in public health that led to mortality transition in the United States in the late nineteenth and early twentieth centuries. Clean water, sanitation and electrification helped reduce mortality, as did food and milk inspection, the elimination of parasites such as malaria and

G. T. Niemesh (✉) · M. A. Thomasson (✉)
Department of Economics, Miami University, Oxford, OH, USA

NBER, Cambridge, MA, USA
e-mail: niemesgt@miamioh.edu; mthomasson@miamioh.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_77

hookworm, and food fortification. As the century progressed, advances in science and technology, combined with reforms in physician education and licensing led to improvements in medical care and health. These changes increased the cost of medical care and led to the development of health insurance markets.

Keywords
Health care and insurance · Public health · Hospitals · Medical education

Introduction

In the late nineteenth and early twentieth centuries, infant mortality and life expectancy improved as mortality transitioned from infectious disease mortality to chronic disease mortality beginning in 1915 (Meeker 1972; Haines 2001). Early in this period, gains in longevity can be largely attributed to gains in public health. For the most part, medical care did not contribute to improved mortality because it was largely ineffective. A large body of research in economic history shows that gains in public health during this period led to significant improvements in health outcomes. Clean water, sanitation, and electrification helped reduce mortality, as did food and milk inspection. The eradication of parasites such as hookworm and malaria, and the fortification of foods to eliminate nutritional deficiencies also played roles in decreasing mortality and improving other outcomes.

In the first few decades of the twentieth century, advances in science and technology, coupled with improvements in physician education, led to improved health and quality of life. Discoveries such as diphtheria antitoxin, salvarsan for syphilis, and antibiotics led to more awareness of the effectiveness of medical care. As these discoveries increased the demand for medical care, they led to increases in medical costs and the development of the U.S. health insurance system. Occupational licensing of physicians and other medical professionals evolved both as medical schools sought to improve the quality of new physicians, and to protect consumers from unqualified practitioners. In this chapter, we focus on summarizing the literature in these different strands of research. We first discuss the literature on the improvements in health generated by public health initiatives, and then summarize studies that examine how changes in physician education and licensure affected the demand and supply of medical services. Finally, we provide an overview of studies that look at the emergence and impact of health insurance in the United States.

Improvements in Public Health

The definition of "public health" evolved with the changing health problems afflicting the population. The nineteenth- and early twentieth-century definition focused on the elimination of contagious infectious diseases (cholera, typhoid, dysentery) through "community action to avoid disease and other threats to the health and welfare of individuals and the community at large" (Duffy 1992, p. 1). Public

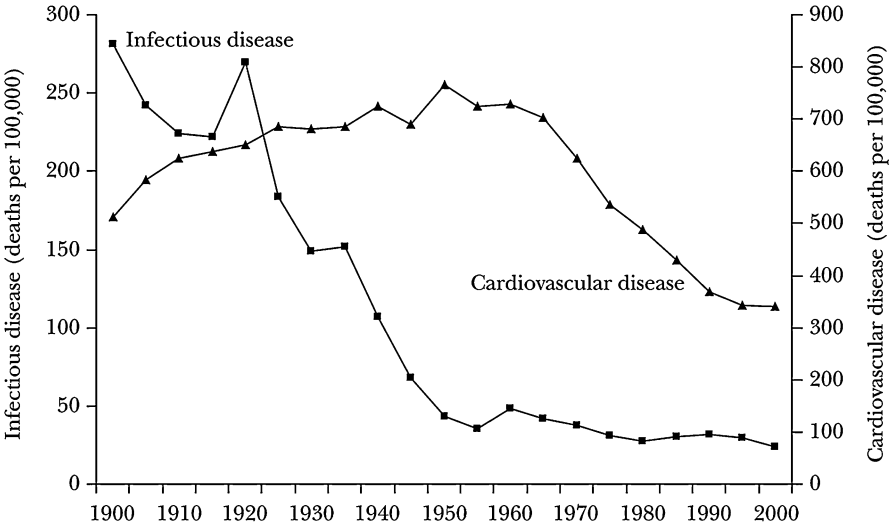


Fig. 1 Mortality from infectious disease and cardiovascular disease, United States, 1900–2000. (Source: Reprinted from Cutler et al. (2006). Data are from the Centers for Disease Control and Prevention, National Center for Health Statistics, and are age adjusted.)

health initiatives are often undertaken on the part of governments – local, state, and federal – and not through voluntary exchange in the health care market between physician and patient. The success of the early public health initiatives can be seen in Fig. 1: an initially high but rapidly falling mortality rate from infectious disease during the first half of the twentieth century in the United States. By at least the 1930s, the chronic diseases of old age had become the biggest killers. In response, the definition of public health expanded to include "actively promoting health rather than simply maintaining it" (Duffy 1992, p. 1), with a focus on the social determinants of health as a way to address mortality from chronic diseases. For example, poverty is viewed as an underlying cause of high mortality rates. Therefore, guaranteeing everyone a standard of living sufficient for a healthy existence became an aim for public health organizations.

It is important to emphasize the contributions of economic history to understanding the disease elimination aspects of the early public health movement. One of the main findings of the literature is that much of the improvement in health early in the twentieth century came from public health interventions. Early initiatives that nearly universally improved health included water purification and sanitation. Later in the century, nutritional fortification also improved health for Americans. Other public health efforts focused on eradicating parasites, such as hookworm and malaria.

Water Purification and Sewage Systems

Public health interventions such as water filtration, sanitation, refuse collection, and food inspection played a significant role in the decline of mortality in the early

twentieth century. Using city-level data, David Cutler and Grant Miller (2005) find that water filtration and chlorination account for almost half of the mortality reduction in major cities, 75% of the decrease in infant mortality, and two-thirds of the decrease in child mortality. Similarly, Werner Troesken (2004) finds that deaths from typhoid and other waterborne diseases were nearly eliminated by the expansion of public water and sewer services. Looking at Massachusetts, Alsan and Goldin (forthcoming) show that the combination of safe water and sewerage lowered child mortality by 26.6 log points and was more effective than either intervention alone. Joseph P. Ferrie and Werner Troesken (2008) find that reducing typhoid also led to a reduction in nonwaterborne diseases, including death rates from gastroenteritis, tuberculosis, pneumonia, influenza, bronchitis, heart disease, and kidney disease (known as the Mills-Reincke phenomenon).

While water purification systems reduced deaths from typhoid, they may have led to increased morbidity and mortality in cities that used lead pipes to deliver the water, depending on the age of the pipe and the acidity of the local water supply. In two studies, Troesken (2006, 2008) reports that lead resulted in increases of 25–50% for infant mortality and stillbirths in 1900 for the average Massachusetts town. As cities abandoned lead water pipes in the early 1900s, the deleterious effects of lead exposure on infant mortality were reduced, although a lack of data makes it difficult to estimate the size of the reduction. Karen Clay et al. (2014) follow up on this research and look at the effect of waterborne lead exposure on infant mortality in U.S. cities over the first two decades of the twentieth century. They use city-level variation in water acidity and the type of pipe used by cities (lead, iron or concrete), and confirm that lead pipes led to significant increases in infant mortality. Specifically, they find that increasing pH from the 25th percentile to the 50th percentile (since water becomes less acidic, it leaches lead more slowly) would reduce infant mortality by 7–33%.

Early public health improvements occurred nearly exclusively at the state and local levels (Preston and Haines 1991). State and local public health departments operated to reduce mortality in a variety of ways. Louis P. Cain and Elyce Rotella (2001) use data from 48 U.S. cities to examine the impact of sanitation expenditures on death rates. They find that spending on sewer systems and refuse collection reduced death rates from typhoid, dysentery, and diarrhea. Municipalities also engaged in street cleaning and the distribution of diphtheria antitoxin (Condron and Crimmins-Gardner 1978; Meckel 1990). State and local legislation was aimed at protecting food and milk and preventing the spread of disease. Mokyr and Stein (1996) state that by 1905, 32 states had laws preventing the adulteration of milk.

Several different types of laws were also enacted to stem the spread of tuberculosis. Between 1900 and 1917, state and local governments passed laws that required TB reporting. They also enacted disinfection laws, spitting bans, and common drinking cup bans (Anderson et al. forthcoming). Anderson et al. examine these laws and find that reporting laws led to a 6% reduction in the TB mortality rate, and that the establishment of a state-run sanatorium reduced TB by 4%. This finding is similar, but larger in magnitude to that found by Alex Hollingsworth (2013) in his study of sanitaria in North Carolina. However, a paper by Karen Clay et al. (2018)

suggests that community-based health interventions targeted at reducing tuberculosis did not reduce tuberculosis relative to control cities over time, although they did reduce infant mortality.

Scholars interested in examining state and local spending and their impact on outcomes can rely on several sources of data. At the state level, *Financial Statistics of States* provides similar information to that of *Financial Statistics of Cities* beginning in 1915. Richard E. Sylla et al. (1993) have published relevant data in State and Local Government Sources and Uses of Funds: Twentieth Century Statistics (ICPSR Study 6304). For municipalities, the *Bureau of Labor Statistics Bulletins* #24, 30, 36, and 42 (1899–1902) provide data on municipal finances for 1899–1902, and *Census Bulletin* #20 (United States Bureau of the Census 1904) provides similar information for 1902–1903. *The United States Census Bureau published Statistics of Cities* (1907, various years) to report annual data on municipal-level health related expenditures for cities over 30,000 between 1905 and 1908. *The United States Census Bureau also published Financial Statistics of Cities* (1909, various years), which provides similar information for the years 1909–1913, 1915–1919, and 1921–1930. The specific categories related to health include information on health conservation and sanitation cost payments; health conservation and sanitation outlays; charities, corrections, and hospital cost payments; and charities, corrections, and hospital outlays.

Public Health Education and Information

In addition to direct public health spending on refuse collection, infrastructure, and the passage of laws designed to facilitate the spread of disease, some public health initiatives revolved around educating families about hygiene and infant and child care. Grant Miller (2008) suggests that the enfranchisement of women led to large shifts in public spending on hygiene campaigns that led to increases in child survival. While Miller focuses on data from *Financial Statistics of Cities* (so that his mechanism about how public expenditures led to reductions in child mortality is a black box), other studies have used data on public health activities to identify the channels that were most effective in improving health outcomes. For example, Carolyn Moehling and Melissa A. Thomasson (2014) examine activities undertaken by public health authorities under the Sheppard-Towner program, in which the federal government gave matching funds to states to engage in infant and child care and hygiene. They find that the most effective interventions were those that provided one-on-one care, such as nurse visits and health centers, compared to activities such as classes, conferences, and demonstrations.

The Eradication of Parasites

In the early 1900s, up to 30% of the population was infected with malaria (Kitchens 2013a). The effects of malaria vary from stunted physical stature and impaired

cognitive development to death (Hong 2007; Bleakley 2010). Several New Deal agencies and programs contributed to the decline of malaria. By paying farmers to take land out of cultivation, the Agricultural Adjustment Act (AAA) led farm laborers to leave mosquito-breeding grounds (Humphreys 2001). Alan Barreca et al. (2012) estimate that this out-migration accounts for about 10% of the decline in malaria between 1930 and 1940. Carl Kitchens (2013a, b) examines the impact of other New Deal programs on malaria. Using county-level data from Georgia between 1932 and 1941, Kitchens (2013a) demonstrates that drainage projects constructed under the auspices of the Works Progress Administration (WPA) explain more than 40% of the observed reduction in malaria over the period.

Kitchens (2013b) finds different results when he looks at the construction of dams under the Tennessee Valley Authority (TVA). He uses county-level panel data from Alabama and Tennessee, and finds that the dams created a vast increase in coastline suited for mosquito breeding. Despite subsequent efforts of the TVA to control mosquitos, Kitchens calculates that the TVA led to a significant increase in loss of life due to malaria that reduced the fiscal benefit of dam construction by 24%.

Other public health efforts focused on eradicating another Southern parasite: hookworm. Transmitted through the soil, hookworm eventually lodges in the intestines of its victim. It causes lethargy and anemia, but is rarely fatal. Nevertheless, its effects can lead to decreased productivity and make it difficult for children to cognitively focus. For example, Garland Brinkley (1997) shows that the sharp decline in Southern agricultural output after the Civil War can be attributed to increased rates of hookworm infection.

In 1910, the Rockefeller Sanitary Commission (RSC) estimated that 40% of Southern schoolchildren were affected with hookworm. It engaged in an eradication campaign, and sent health care workers to dispense de-worming medication. Hoyt Bleakley (2007) finds that children living in areas with greater rates of hookworm infection prior to the RSC's campaign showed greater gains in school enrollment, attendance, and literacy than those living in areas with lower rates of infection. Thus, eradicating hookworm could account for closing up to half the literacy gap between the North and the South, and reducing the income gap by up to 20%.

Improvements in Diet

Nutritional insufficiency – whether stemming from parasitic infection or poor diet – is correlated with reduced economic outcomes. Numerous economists and historians have noted that nutritional improvements (measured by caloric and/or protein intake) are correlated with gains in both income and health (Higgs 1971; Fogel 1994; Steckel 1995; Floud et al. 2011). Most recently, Gregory Niemesh (2015) adds to this literature by measuring the impact of the first federal requirement in 1943 to fortify bread with iron. Iron deficiency in infants and children causes developmental delays and behavioral problems, and reduces productive capacity in adults. Niemesh uses pretreatment variation in iron consumption, and shows that the law led to increases in income and educational attainment in areas with lower levels of iron

consumption prior to the mandate. James Feyrer et al. (2017) similarly examine the impact of salt iodization in the U.S. in the 1920s on later cognitive outcomes. Their findings suggest that iodized salt raised IQ for those who were most deficient, but also increased thyroid-related deaths, particularly among older individuals. Karen Clay et al. (2018) examine the rise and fall of pellagra, a disease caused by niacin deficiency in the American South that is characterized by dermatitis, diarrhea, and dementia. They argue that cotton production in the South displaced local food production, leading to increases in pellagra rates. Using a difference-in-differences framework, they leverage the arrival and spread of the boll weevil to identify the relationship between cotton production and pellagra. They find that pellagra death rates fell between 23% and 40% more in high cotton counties than in low cotton counties after the boll weevil arrived. They also find that after 1937, state fortification laws helped to eliminate pellagra.

The Growth of the Market for Medical Care

Reforms in Medical Education and the Changing Public Perception of Hospitals

At the turn of the twentieth century, the formal health care market was much simpler than today and made up a smaller portion of economic activity. Consumer expenditures on medical care in 1900 (about $384 million) accounted for around 2% of Gross Domestic Product (Craig 2006; Sutch 2006), compared to around 18% in 2017 (Centers for Medicare & Medicaid Services 2018). Medical care was inexpensive because it was ineffective, so people did not need health insurance to pay for medical expenses (Thomasson 2002). Hospitals were generally shunned by people of means, and functioned as almshouse and places for those without families to care for them. Respectable people of means had physicians visit them in their homes.

At the turn of the twentieth century, physician training in the U.S. was largely substandard compared to medical education in Europe, but several medical schools had already made significant strides in reforming the quality of medical education. Johns Hopkins, Harvard, Michigan, and Pennsylvania had all increased admissions requirements, lengthened their terms, and moved from apprenticeships to a focus on clinical instruction. By the time Abraham Flexner (Flexner et al. 1910) published his well-known report on the quality of medical education in the United States, the reform movement was well underway (Moehling et al. 2018). A key component of modern medical education was the alliance between medical schools and clinical training in hospitals. As physician training moved to hospitals, so too did patients. At the turn of the twentieth century, few people even considered going to hospitals and preferred to have physicians visit their homes. This changed significantly as the public gradually became aware of scientific progress and visited physicians trained under the new regime. Although hospitals were considered second-rate and germ-ridden in 1900, by the second decade of the century, some products (such as Lysol) featured advertising boasting how they were used in hospitals. As news from Europe

during World War I showed doctors saving lives on the battlefield, the public perception of hospitals grew more favorable.

Occupational Licensing of Health Care Providers

While the first physician licensure laws in the United States were passed in the 1870s, most were very lax until the 1910s. For example, even as late as 1906, 13 states still allowed people who had not graduated from medical school to become physicians (Ludmerer 1985). Supporters of medical education reform had long pushed for stricter licensing requirements, but Flexner's report spurred legislators to implement higher standards to ensure that would-be physicians could meet new standards. States continued to ratchet up the entry and graduation requirements throughout the 1910s. Term lengths were lengthened to 4 years, and states began to require premedical education of at least 2 and sometimes 4 years. Many schools did not have the resources to invest in reforms, and subsequently closed or merged with other schools. In response, the number of medical schools fell from 160 in 1900 to 81 in 1922 (American Medical Association, Council on Medical Education 1922, p. 633).

Figure 2 displays the widespread adoption of stricter education qualifications for entry into medical school. No states required premedical education in 1907. By 1915, only 8 years later, 35 states had mandatory baccalaureate education as a requirement to admission to medical school. Figure 3 shows the simultaneous

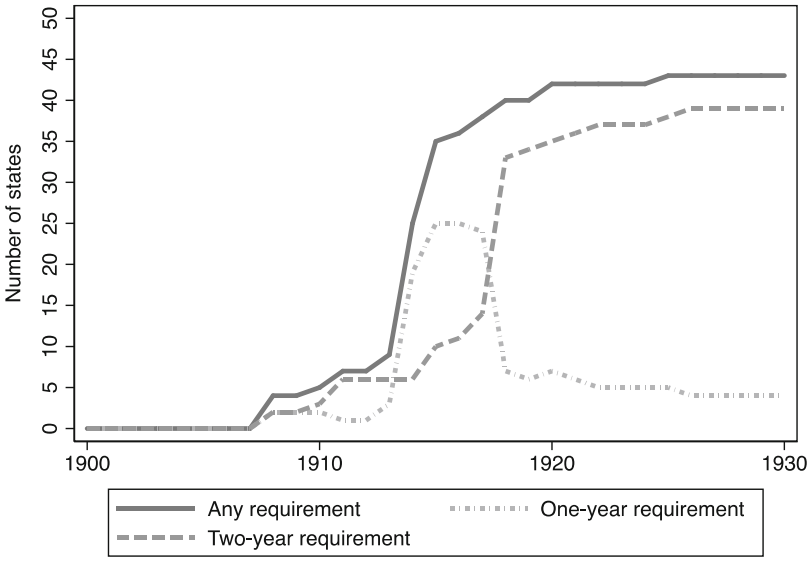


Fig. 2 States with premedical college requirements for physician licensure. (Source: "Medical Education in the United States," *Journal of American Medical Association*, Aug. 27, 1932: p. 746)

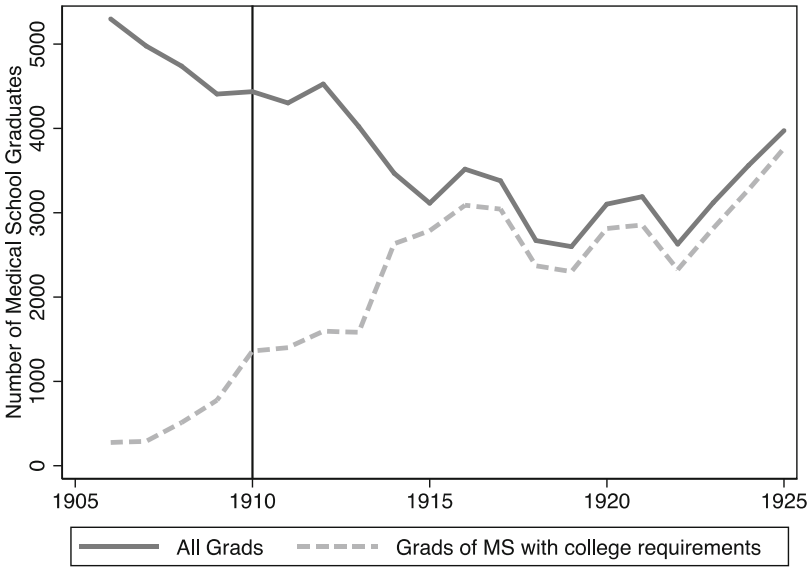


Fig. 3 Number of medical school graduates by college requirements. (Sources: Premedical college requirements are from American Medical Association Council on Medical Education (1919, 1923). Medical schools and graduates from American Medical Association Council on Medical Education (1905–1910, 1911–1914, 1915–1920).)

increase in the skill of new medical school graduates and the decrease in the supply. The number of new physicians produced by American medical schools had been gradually falling prior to the publication of the Flexner report in 1910, but the increased premedical education requirements accelerated the trend. As attending a medical school became increasingly more expensive – indirect costs from lost wages and direct costs from tuition of attending a baccalaureate institution – the number of medical school graduates declined. On the other hand, the skill of the potential entrants increased, measured as fulfilling the more strict premedical education requirements.

Economic historians have tried to explain the early adoption of occupational licensure laws by health professions, as well as the effect of widespread coverage of occupations across the various states. Their work focuses on the impact of licensing on supply, price, and quality, and they leverage their results to distinguish between two economic theories that explain the existence of occupational regulations. One theory is that professions are motivated by their own self-interest to lobby governments to enact occupational licensure requirements. As requirements for entry into the profession become stricter, competition for incumbent practitioners becomes more limited. Supply falls, so that prices and wages increase above what would have occurred in the absence of a license system. A related idea is that of regulatory capture, which occurs when practitioners take control of the licensing apparatus meant to regulate entry into an occupation. For example, physicians make up the

majority of members of state medical boards. With the licensing apparatus captured, boards could enact regulations in the interest of their profession (e.g., restrict supply and increase wages) at the expense of consumers (e.g., higher prices and less access). Underlying this view is that licensure provides no benefit to the consumer by increasing physician quality, or at least any quality gains that do occur do not fully compensate for the loss in welfare from higher prices and lower quantity of health services provided.

Alternatively, licensure might have emerged to solve an asymmetric information problem between the sellers of health services and patients. Physician quality may be unobservable. Since patients do not have the ability to discern "quacks" from good doctors, occupational licensure can help solve the asymmetric information problem by eliminating the low-quality providers from the market. Education requirements or an entrance exam to the profession proves a provider is of high quality. With the low-quality competition gone, the remaining high-quality providers may see an increase in wages. However, in contrast to the pure self-interest of regulatory capture, the consumers benefit on net from the quality increases in this case.

Physicians

Law and Kim (2005) document the effects of physician licensure laws on the supply and quality of doctors during the Progressive Era using licensure data spanning 1870–1930. They use cross-state variation in the timing of when specific parts of a licensure law were passed in a state-level decadal panel. In general, passage of a licensure law or an increase in entry requirements reduced entry to the profession and reduced the supply of doctors per capita. The requirements most related to declining entry were the implementation of 2-year and 4-year premedical education requirements; they find little evidence that the initial licensure statutes of the 1870s affected entry.

In addition, they find evidence that physician licensure may have reduced asymmetric information and improved patient outcomes. Using the same empirical framework, they find suggestive evidence that physician quality improved in response to the premedical education requirements during the 1910s. States with premedical requirements experienced relatively larger declines in maternal mortality and appendicitis mortality, negative outcomes that physician behavior might be able to mitigate during this period. However, the author's find no effect of greater licensure requirements on either overall mortality or the infant mortality rate.

Law and Kim (2005) provide additional evidence that that licensure reduced information asymmetry and improved quality. Professions that a priori one would expect to more acutely suffer from informational asymmetries were more successful in restricting entry (physicians, dentists, and veterinarians as opposed to plumbers, electricians, and barbers). Their results suggest this increased quality came at the expense of a reduction in entry into the profession.

Using panel data on medical school enrollments from 1906 to 1932, Moehling, Niemesh and Thomasson (2018) measure how medical school and state licensing

requirements may have differentially affected women physicians. They look at the effect of premedical college requirements at the school level on the percentage of female enrollment, and also include a third measure of reform: whether the school required graduates to complete an internship. In addition, they examine how state-level requirements affected enrollments. Their results indicate that as schools increased their standards for admission and required hospital internships, the shares of women enrolled fell (Fig. 4). Similarly, when a state imposed a hospital internship requirement for obtaining a physician license, the share of women in medical schools in that state decreased. These results run counter to the findings of Law and Marks (2009), who find no adverse effects of state licensing laws on the likelihood a woman reported herself as a physician in the census from 1870 to 1930. The differences may arise because Law and Marks do not examine the effect of the requirement to complete a hospital internship. In addition, they use individual level data from the IPUMS samples of the population censuses. While large from the perspective of historical data sets, the small number of women physicians during this period makes the variation used to identify the effects of changes in the licensing laws very small.

While these studies look at the effect of licensing on physician supply, focusing solely on the state-level supply impacts of licensure masks the geographic redistribution of physicians that may occur as the market for physician services adjusts to a new spatial equilibrium. Moehling et al. (2018) examine the rural/urban location choices of physicians in the first decades of the twentieth century to test whether they were related to the changes in medical education during the period. Physicians were becoming a more urban profession over the course of the early twentieth century, as was the population of the country as a whole. However, physicians trained in medical schools that required premedical education were significantly more likely to set up practice in an urban area than physicians that graduated from schools without premed education requirements. Physicians trained in "modern" medical schools were more strongly attracted to professional amenities located in urban areas such as hospital beds and larger physician communities than were physicians trained in lower quality schools. As states enacted increasingly strict admission requirements to medical school, not only did the total supply of physicians decrease, but the composition of new graduates was more likely to locate in an urban area. These two factors combined led a reduction in access to physician services in rural areas, increasing an already existing urban-rural divide.

Midwives

The studies above describe the effect of medical licensure on physicians. Yet during the Progressive Era, midwives also became more highly regulated. At the turn of the twentieth century, half of all births were attended by midwives, with the other half attended by physicians. Only 5% of births (mostly those by indigent women or unwed mothers) took place in hospitals (Wertz and Wertz 1989). Midwives during this period were (almost entirely) women who learned the craft through helping more

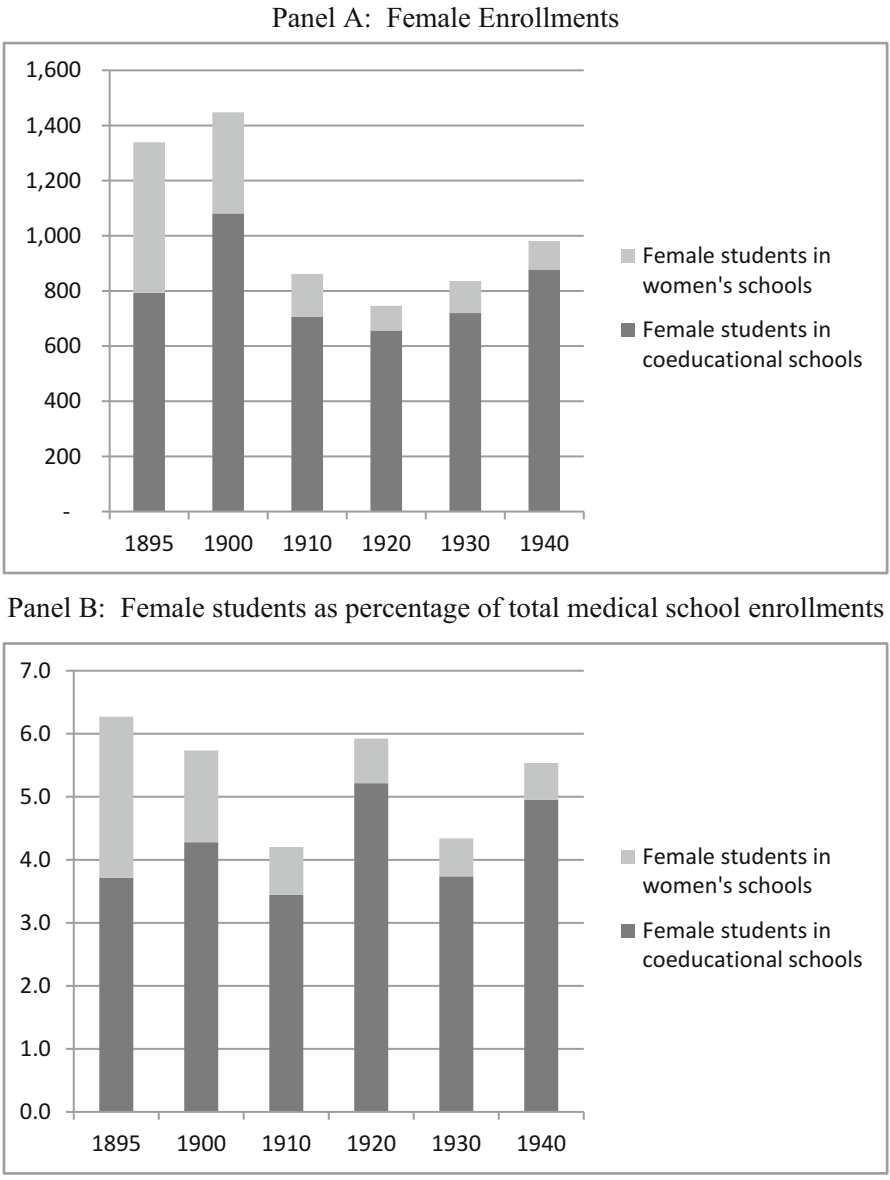


Fig. 4 Female medical school enrollment, 1895–1940. (Sources: United States Commissioner of Education (United States Commissioner of Education 1895 and 1900); American Medical Association Council on Medical Education (1910–1940))

experienced midwives. Formal training was not provided. Moreover, the market for midwives was wholly unregulated prior to a string of licensure laws enacted in the early twentieth century in an effort to increase the quality of midwifery services.

Anderson et al. (2016) analyze the supply and mortality outcomes associated with 22 states and at least a dozen municipalities enacting licensing requirements over the 1900–1940 period. The stringency of entry requirements varied tremendously across the states. "Applicants for licenses in Mississippi were judged based on their character, cleanliness and intelligence, but were not required to take an exam or graduate from a school of midwifery. By contrast, midwives in California, Washington and Wisconsin were required to graduate from a recognized school of midwifery and to pass an examination administered by their State Board of Medical Examiners" (Anderson et al. 2016, p. 3). The introduction of licensing requirements for midwives was associated with a 6–7% reduction in maternal mortality.

The Impact of Medical Care on Health

As medical education improved and medical care advanced, it became more effective. Thomasson and Treber (2008) show that the while medicalization of childbirth did not initially reduce maternal mortality, maternal mortality declined once sulfa drugs became available in 1937. Sulfa provided physicians with their first effective treatment against a range of bacterial infections. Jayachandran et al. (2010) show that in addition to reducing maternal mortality, sulfa use also reduced deaths from pneumonia and scarlet fever. They estimate that sulfa reduced overall mortality from 2% to 3%, and added 0.4–0.7 years of life expectancy.

Medical Costs and the Development of the Health Insurance Market

The fundamental function of any kind of insurance is to reduce financial uncertainty associated with catastrophic events by pooling risk. People pay a fixed amount of money over time, and receive a payout if they experience a loss. The market works because on average, the amount paid in premiums by the group is less than the amounts paid out in benefits. Prior to the late 1920s, health insurance in the U.S. did not develop for two reasons. First, the demand for health insurance was low. When medical care was inexpensive, people did not need health insurance to pay unexpectedly high bills (Thomasson 2002). Instead, wage earners needed to cover wage-loss associated with disability, so they obtained disability insurance (called "sickness" insurance) through their firms, unions, or through fraternal societies (Emery 1996; Emery and Emery 1999; Murray 2007). The lack of need for actual health insurance is reflected in the first failure of an attempt at health insurance reform during the Progressive Era. In 1916, the American Association for Labor Legislation (AALL) published a draft bill in which they proposed comprehensive sickness and medical benefits for low income workers. Under the plan, local mutual insurance companies would manage premium contributions shared by employers, workers, and the state. Employers and workers would each contribute 40% of the plan's premium, while the state would contribute the remaining 20% (Chasse 1994).

Entrenched interests, including insurance companies, druggists, and physicians opposed the bill, and popular support was lacking, as demonstrated by the results of a 1918 referendum in California in which the plan was defeated with 358,324 votes against to 133,858 in favor of the bill (Murray 2007).

The lack of demand for health insurance was joined by a lack of insurance companies willing to underwrite "health." Commercial insurance companies did not view health as an insurable product. In order for insurance markets to function well, two conditions must hold. First, the losses that insurance companies cover must be able to be measured and observed. Insurance companies wondered how they would monitor health and pay claims for ill health. Would they be able to tell who was really sick versus just malingering? If losses are hard to measure, they can result in moral hazard – when having insurance makes it more likely that an insured person would have a claim. Second, insurance companies can only offer insurance if they are able to measure the likelihood of someone having a claim. If policyholders have private information about their likelihood to have a claim, insurance companies will not collect enough money in premiums to be able to pay for losses. This problem, known as adverse selection, can prevent insurance markets from functioning. In the early twentieth century (and even today), insurance companies worried that they might not be able to tell who was likely to have a health claim and who was not, and as a result, they did not feel like they could profitably offer health insurance coverage.

The demand for health insurance did not rise until the overall cost of medical care increased and became more variable as medical treatment shifted to hospitals. By 1929, a nationwide study by the Committee on the Costs of Medical Care (CCMC) showed that the average American family had medical expenses totaling $108, with hospital expenditures contributing to about 14% of medical expenses (Falk et al. 1933). This average disguised significant variation; urban families with incomes between $2,000 and $3,000 per year had mean medical expenses of $67 without hospitalization but expenses of $261 if someone had been admitted to the hospital (Falk et al. 1933).

As the costs associated with hospitalization increased, some families had difficulty paying their bills. In response, hospitals began to organize payment plans, and in so doing, unwittingly mitigated the problems of adverse selection and moral hazard, and set the stage for the broad sale of health insurance. In 1929, Justin Ford Kimball, a former superintendent of schools, was an administrator at Baylor University Hospital. He worked with Dallas teachers to develop a plan, later known as Blue Cross, based on the principles of insurance to help them pay their bills: Baylor would provide each teacher with 21 days of hospital care for an annual fee of $6.00. These plans reduced adverse selection by selling insurance to groups of workers healthy enough to work. They reduced moral hazard because Blue Cross plans reimbursed hospitals directly and patients generally could not admit themselves to hospitals. During the Great Depression, more and more hospitals began to develop these plans as hospital occupancy rates – and revenues – fell.

Blue Cross plans also benefited from state-level legislation called "enabling laws" that allowed them to form as nonprofit corporations and enjoy tax-exempt status, as

well as being exempt from certain insurance regulations such as reserve requirements and assessment liability. Thomasson (2002) shows that these laws increased the amount of health insurance at the state level.

Although physicians were somewhat slower than hospitals to develop prepaid plans, the American Medical Association (AMA) feared national health insurance and attempted to forestall its development by encouraging state and local medical societies to form their own prepayment plans, which later became known as Blue Shield (Thomasson 2002). By 1940, 9% of the U.S. population was covered by health insurance, largely through the Blue Cross and Blue Shield plans (Thomasson 2003). After Blue Cross demonstrated, it had overcome the problems of adverse selection and moral hazard, commercial, for-profit companies began to move rapidly into the market.

In the 1940s, a series of factors led to the rapid expansion of health insurance. The National War Labor Board limited the ability of firms to raise wages to secure labor, even as the U.S. entry into World War II led to a shortage of workers. Health insurance was exempted from the ruling, and firms began to offer health benefits to attract workers. An administrative tax court ruling in 1943 (later codified under the 1954 Internal Revenue Code) exempted employer contributions to worker health insurance premiums from employee income taxes. Thomasson (2003) finds that the tax change increased the likelihood that a household would have coverage by 9%, and increased the amount of coverage purchased by 9.5%. By 1957, about 76% of the population held some form of private health insurance coverage.

The Impact of War on Poverty on Health Insurance

The development of employment-based insurance as a means to alleviate adverse selection and its accommodation by U.S. tax policy makes it difficult for people without jobs to receive health insurance coverage. The elderly, disabled, and unemployed often had a difficult time finding health insurance and paying for medical expenses. Their need for financial assistance was recognized in Congress. In 1950, amendments to the Social Security Act allowed the federal government to provide matching funds to states to pay doctors and hospitals for providing medical care to welfare recipients. In 1960, these "vendor payments" were expanded to include the elderly who were not welfare recipients (Moore and Smith 2005).

Despite the effort to support the elderly who could not afford medical care under the Medical Assistance for the Aged Act, Congressional hearings in the early 1960s concluded that ". . .increasing numbers of our older people are confronted with financial catastrophe brought on by illness" (U.S. Congress. Senate. Special Committee on Aging 1964). In 1965, Medicare became law. Under Part A, the elderly were automatically enrolled in the compulsory hospital insurance program upon reaching age 65. Part B provided insurance for physicians' services. Research on Medicare shows both its costs and benefits. David Card et al. (2008) demonstrate that because of Medicare, insurance coverage increases at age 65. By leveraging this discontinuous change, they are able to examine differences in health

care utilization and outcomes across different groups (such as educated whites compared to less educated minorities). They find that health insurance coverage does affect utilization of medical services, and find a small increase in self-reported health status. Looking at the impact of Medicare on mortality, Amy Finkelstein and Robin McKnight (2008) found that while Medicare did not reduce the mortality of elderly individuals, it significantly reduced their out-of-pocket expenses and financial risk.

Medicaid (Title XIX of the Social Security Act) was also enacted with Medicare to provide health insurance coverage to non-elderly populations in need of assistance with medical bills. In contrast to Medicare, which was federally funded and provided uniform benefits to all enrollees, states' participation in Medicaid was voluntary. States that participated in Medicaid received some federal funds to provide means-tested benefits originally to recipients of public assistance, although legislative changes over the years have expanded eligibility. While the federal government specified minimum standards for eligibility and benefits, states have the option to make eligibility for coverage or benefits more generous. Under the Affordable Care Act of 2010, the federal government provides additional funds to states seeking to expand Medicaid eligibility. There is a very large literature on the impact of Medicaid on both physical and financial health of enrollees. Thomas Buchmueller et al. (2016) provide a summary of the program and a very thorough review of the literature.

Directions for Future Research

This chapter enumerates a wide variety of studies that examine the effect of public health initiatives on health and well-being, as well as the development of health care and health insurance markets. Nevertheless, significant gaps in the literature point to areas where future research is needed. For example, we know little about the impact of vendor payments paid to hospitals and doctors on behalf of welfare recipients under the 1950 amendments to the Social Security Act, yet research suggests the effects may have been substantial. For example, work by Martha Bailey and Andrew Goodman-Bacon (2015) shows that Community Health Centers (rolled out as part of the War on Poverty) significantly lowered age-adjusted mortality among older Americans. Also of interest is more work in the area of the impact of both de jure and de facto racial segregation and prejudice. Douglas Almond et al. (2006) demonstrate that the elimination of segregation in Southern hospitals after 1964 reduced infant mortality among blacks. A 2018 study found that underrepresentation by blacks in the medical profession may lead to excess mortality among black men (Alsan et al. 2018), which echoes results from Alsan and Wanamaker (2018), who found that the exploitative Tuskegee Study contributed to racial disparities in health among its victims and those close to them. Both of these studies suggest that more work can be done by economic historians to examine the causes and effects of racial health disparities.

Cross-References

► Nutrition, the Biological Standard of Living, and Cliometrics

References

Almond D, Chay KY, Greenstone M (2006) Civil rights, the war on poverty, and Black-White convergence in infant mortality in the rural South and Mississippi. Social Science Research Network, Rochester

Alsan M, Goldin C (forthcoming) Watersheds in child mortality: the role of effective water and sewerage infrastructure, 1880 to 1920. J Polit Econ

Alsan M, Wanamaker M (2018) Tuskegee and the health of Black men. Q J Econ 133:407–455. https://doi.org/10.1093/qje/qjx029

Alsan M, Garrick O, Graziani GC (2018) Does diversity matter for health? Experimental evidence from Oakland. National Bureau of Economic Research, Cambridge, MA

American Medical Association, Council on Medical Education (1922) Medical education in the United States. J Am Med Assoc 79

Anderson DM, Brown R, Charles KK, Rees DI (2016) The effect of occupational licensing on consumer welfare: early midwifery laws and maternal mortality. National Bureau of Economic Research, Cambridge, MA

Anderson DM, Charles KK, Las Heras Olivares C, Rees DI (forthcoming) Was the first public health campaign successful? The Tuberculosis Movement and its effect on mortality. Am Econ J Appl Econ. https://doi.org/10.1257/app.20170411

Bailey MJ, Goodman-Bacon A (2015) The war on poverty's experiment in public medicine: Community Health Centers and the mortality of older Americans. Am Econ Rev 105: 1067–1104. https://doi.org/10.1257/aer.20120070

Barreca AI, Fishback PV, Kantor S (2012) Agricultural policy, migration, and malaria in the United States in the 1930s. Explor Econ Hist 49:381–398. https://doi.org/10.1016/j.eeh.2012.05.003

Bleakley H (2007) Disease and development: evidence from hookworm eradication in the American South. Q J Econ 122:73–117. https://doi.org/10.1162/qjec.121.1.73

Bleakley H (2010) Malaria eradication in the Americas: a retrospective analysis of childhood exposure. Am Econ J Appl Econ 2:1. https://doi.org/10.1257/app.2.2.1

Brinkley GL (1997) The decline in southern agricultural output, 1860–1880. J Econ Hist 57: 116–138

Buchmueller T, Ham JC, Shore-Sheppard LD (2016) Economics of means-tested transfer programs in the United States, vol 1. University of Chicago Press, Chicago

Cain LP, Rotella EJ (2001) Death and spending: urban mortality and municipal expenditure on sanitation. Ann Demogr Hist 1:139

Card D, Dobkin C, Maestas N (2008) The impact of nearly universal insurance coverage on health care utilization: evidence from Medicare. Am Econ Rev 98:2242–2258. https://doi.org/10.1257/aer.98.5.2242

Centers for Medicare & Medicaid Services (2018) NHE-Fact-Sheet. In: NHE Fact Sheet. https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html. Accessed 20 Aug 2018

Chasse JD (1994) The American Association for Labor Legislation and the institutionalist tradition in national health insurance. J Econ Issues 28:1063–1090

Clay K, Troesken W, Haines M (2014) Lead and mortality. Rev Econ Stat 96:458–470. https://doi.org/10.1162/REST_a_00396

Clay K, Egedesø P, Hansen CW, Jensen PS (2018) Controlling tuberculosis? Evidence from the mother of all community-wide health experiments. SSRN Electron J. https://doi.org/10.2139/ssrn.3144355

Clay K, Schnick E, Troesken W (2018). The rise and fall of pellagra in the American South. National Bureau of Economic Research Working Paper 23730
Condron G, Crimmins-Gardner E (1978) Public health measures and mortality in U.S. cities in the late nineteenth century. Hum Ecol 6:27–54
Craig L (2006) Consumer expenditures. In: Carter SB, Gartner SS, Haines MR, Olmstead AL, Sutch R, Wright G (eds) Historical statistics of the United States: millennial edition, vol 3. Cambridge University Press, Cambridge
Cutler D, Miller G (2005) The role of public health improvements in health advances: the twentieth-century United States. Demography 42:1–22. https://doi.org/10.1353/dem.2005.0002
Cutler D, Deaton A, Lleras-Muney A (2006) The determinants of mortality. J Econ Perspect 20:97–120
Duffy J (1992) The sanitarians. University of Illinois Press, Urbana
Emery JCH (1996) Risky business? Nonactuarial pricing practices and the financial viability of fraternal sickness insurers. Explor Econ Hist 33:195–226
Emery G, Emery JCH (1999) A young man's benefit: the Independent Order of Odd Fellows and sickness insurance in the United States and Canada, 1860–1929. McGill-Queen's University Press, Montreal & Kingston
Falk ISC, Rorem R, Ring MD (1933) The cost of medical care. The University of Chicago Press, Chicago
Ferrie JP, Troesken W (2008) Water and Chicago's mortality transition, 1850–1925. Explor Econ Hist 45:1–16. https://doi.org/10.1016/j.eeh.2007.06.001
Feyrer J, Politi D, Weil DN (2017) The cognitive effects of micronutrient deficiency: evidence from salt iodization in the United States. J Eur Econ Assoc 15:355–387. https://doi.org/10.1093/jeea/jvw002
Finkelstein A, McKnight R (2008) What did Medicare do? The initial impact of Medicare on mortality and out of pocket medical spending. J Public Econ 92:1644–1668. https://doi.org/10.1016/j.jpubeco.2007.10.005
Flexner A, Carnegie Foundation for the Advancement of Teaching, Pritchett HS (1910) Medical education in the United States and Canada; a report to the Carnegie Foundation for the Advancement of Teaching. New York City
Floud R, Fogel RW, Harris B, Hong SC (2011) The changing body: health, nutrition, and human development in the western world since 1700. Cambridge University Press, Cambridge
Fogel RW (1994) Economic growth, population theory, and physiology: the bearing of long-term processes on the making of economic policy. Am Econ Rev 84:369–395
Haines MR (2001) The urban mortality transition in the United States, 1800–1940. National Bureau of Economic Research historical paper 134. National Bureau of Economic Research, Cambridge, MA
Higgs R (1971) The transformation of the American economy, 1865–1914: an essay in interpretation. Wiley, New York
Hollingsworth A (2013) The impact of sanitaria on pulmonary tuberculosis mortality: evidence from North Carolina, 1932–1940. Unpublished working paper
Hong, Sok Chul (2007) The Health and Economic Burdens of Malaria: The American Case. Ph.D. diss., The University of Chicago
Humphreys M (2001) Malaria: poverty, race, and public health in the United States. The Johns Hopkins University Press, Baltimore
Jayachandran S, Lleras-Muney A, Smith KV (2010) Modern medicine and the twentieth century decline in mortality: evidence on the impact of sulfa drugs. Am Econ J Appl Econ 2:118–146. https://doi.org/10.1257/app.2.2.118
Kitchens C (2013a) The effects of the Works Progress Administration's anti-malaria programs in Georgia 1932–1947. Explor Econ Hist 50:567–581
Kitchens C (2013b) A dam problem: TVA's fight against malaria, 1926–1951. J Econ Hist 73:694–724. https://doi.org/10.1017/S0022050713000582

Law MT, Kim S (2005) Specialization and regulation: the rise of professionals and the emergence of occupational licensing regulation. J Econ Hist 65:723–756
Law MT, Marks MS (2009) Effects of occupational licensing laws on minorities: evidence from the Progressive Era. J Law Econ 52:351–366. https://doi.org/10.1086/596714
Ludmerer KR (1985) Learning to heal: the development of American medical education. Basic Books, New York
Meckel RA (1990) Save the babies: American public health reform and the prevention of infant mortality, 1850–1929. The Johns Hopkins University Press, Baltimore
Meeker E (1972) The improving health of the United States, 1850–1915. Explor Econ Hist 9:353–374
Miller G (2008) Women's suffrage, political responsiveness, and child survival in American history. Q J Econ 123:1287–1327. https://doi.org/10.1162/qjec.2008.123.3.1287
Moehling CM, Thomasson MA (2014) Saving babies: the impact of public education programs on infant mortality. Demography 51:367–386
Moehling CM, Niemesh GT, Thomasson MA, Treber J (2018) Medical education reforms and the origins of the rural physician shortage
Moehling CM, Niemesh GT, Thomasson MA (2018) Shut Down and Shut Out: Women Physicians in the Era of Medical Education Reform. Unpublished working paper
Mokyr J, Stein R (1996) Science, health and household technology: the effect of the Pasteur revolution on consumer demand. In: Bresnahan TF, Gordon RJ (eds) The economics of new goods. The University of Chicago Press, Chicago
Moore JD, Smith DG (2005) Legislating Medicaid: considering Medicaid and its origins. Health Care Financ Rev 27:8
Murray JE (2007) Origins of American health insurance: a history of industrial sickness funds. Yale University Press, New Haven
Niemesh GT (2015) Ironing out deficiencies: evidence from the United States on the economic effects of iron deficiency. J Hum Resour 50:910–958. https://doi.org/10.3368/jhr.50.4.910
Preston S, Haines MR (1991) Fatal years: child mortality in late nineteenth-century America. Princeton University Press, Princeton
Steckel RH (1995) Stature and the standard of living. J Econ Lit 33:1903–1940
Sutch R (2006) National income and product. In: Carter SB, Gartner SS, Haines MR, Olmstead AL, Sutch R, Wright G (eds) Historical statistics of the United States: millennial edition, vol 3. Cambridge University Press, Cambridge
Sylla RE, Legler JB, Wallis J (1993) Sources and Uses of Funds in State and Local Governments, 1790–1915: [United States]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR09728.v1
Thomasson MA (2002) From sickness to health: the twentieth century development of U.S. health insurance. Explor Econ Hist 39:233–253
Thomasson MA (2003) The importance of group coverage: how tax policy shaped U.S. health insurance. Am Econ Rev 93:1373–1384
Thomasson MA, Treber J (2008) From home to hospital: the evolution of childbirth in the United States, 1928–1940. Explor Econ Hist 45:76–99
Troesken W (2004) Water, race, and disease. The MIT Press, Cambridge, MA
Troesken W (2006) The great lead water pipe disaster. The MIT Press, Cambridge, MA
Troesken W (2008) Lead water pipes and infant mortality at the turn of the twentieth century. J Hum Resour 43:553–575. https://doi.org/10.3368/jhr.43.3.553
U.S. Congress. Senate. Special Committee on Aging (1964) Blue Cross and private health insurance coverage of older Americans. Government Printing Office, 88th Congress, 2d. Sess
United States Bureau of Labor Statistics (1899) Bureau of Labor Statistics bulletin, #24, 30, 36, and 42
United States Bureau of the Census (1904) Census bulletin, #20
United States Bureau of the Census (1907) Statistics of cities having a population of over 30,000, 1905 (and through 1908 annually). United States Government Printing Office, Washington, DC

United States Bureau of the Census (1909) Financial statistics of cities having a population of over 30,000, 1909 (and 1910–1913, 1915–1919, 1921–1930). United States. Government Printing Office, Washington, DC

United States Commissioner of Education (1895–1900) Report of the Commissioner of Education. United States Government Printing Office, Washington, DC

Wertz RW, Wertz DC (1989) Lying-in: a history of childbirth in America. Yale University Press, New Haven/London

Cliometrics and the Great Depression

Price Fishback

Contents

Abstract

The Great Depression was the worst economic disaster in American history. There were plenty of factors that helped cause the Depression, but there is still ample disagreement in the large literature on the topic as to how much weight to give each cause. In the early 1930s, the Hoover administration and congress nearly doubled federal government outlays, offered a wide range of loans, and sought voluntary efforts to combat the Depression. The economy continued to slide, and increases in tax rates in 1932 contributed to the slide. The economy finally began to grow again in 1933 as Roosevelt and a Democratic Congress developed the New Deal, a large number of new regulatory and spending

Price Fishback is the Thomas R. Brown Professor of Economics at the University of Arizona. I owe a great debt to the scholars, including numerous coauthors and students, who produced the valuable cliometric research that I survey here. Also special thanks are due to Michael Haupert and Claude Diebolt for their help in editing the chapter.

P. Fishback (✉)
Economics Department, University of Arizona, Tucson, AZ, USA
e-mail: pfishback@eller.arizona.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_6

programs. The 1933 trough was so deep that unemployment rates remained high throughout the decade and real GDP per person did not reach its 1929 level again until around 1939 or 1940 despite rapid growth rates. A growing literature has been evaluating the impact of the New Deal programs, and the effects of several major programs are discussed here.

Keywords
Great Depression · New Deal · Government policy · Fiscal policy · Monetary policy · Unemployment · Regulation

The Great Contraction

The Great Contraction was the worst economic disaster in American history. The unemployment rate (Table 1) skyrocketed from 2.9% in 1929 to nearly 16% in 1931 and then rose above 20% in 1932 and 1933. Except for during the 1930s, the annual unemployment rate has only been higher than 10% in one other year, 1921. A significant share of the population was not counted as unemployed because they became discouraged and stopped looking for work. People who kept their jobs often saw their average weekly hours (Table 2) decline by as much as one-fourth as companies tried to share work among more employees.

If anything, the output statistics were worse. In 1930, Americans produced almost 10% fewer final goods and services per person than in 1929 (Table 1 and Fig. 1). Outside of the 1930s, there were only two worse years in American history, during the 1907 Panic and during the military demobilization in 1946. Yet that was only the first year of the Great Depression. In 1931, the USA produced 16% less per person than in 1929, in 1932 27% less, and in 1933 roughly 29% less. It is hard to conceptualize such a drop in GDP. In 1932 and 1933, the drops were the equivalent of shutting down the entire economy west of the Mississippi River. The annual real GDP per capita did not reach its 1929 level again until 1939.

Meanwhile, the price level dropped like a stone. The price level fell 26% over 4 years. Some saw this "deflation" as good news. Workers who kept their jobs at the old wage could now purchase 26% more. But those who owed money, on homes or on the relatively new credit accounts, suddenly saw the values of the dollars that they had to pay back rise markedly. Lenders might have fared better with the more valuable repayments if so many people had not been forced to default on their loans. After taking into account the depreciation of buildings and equipment, the net investment in the USA grounded to a complete halt and then turned negative in one year. Total corporate profits were negative for the years 1932 and 1933. The small percentage of the population owning stocks saw the Dow Jones Stock Index (Fig. 2) fall by roughly 90% over the 4-year period. If you could sell your house in whatever market was left, you were likely to get 40–60% less in nominal terms than in the late 1920s in some cities.[1]

[1]See the new estimates developed by Fishback and Kollmann (2014).

Table 1 Economic statistics from the 1920s and 1930s

	Per capita estimates in 2013 dollars									
		Federal government			Unemployment rate			Growth rate in		
Year	Real GDP	Receipts	Outlays	Surplus/ deficit(−)	Including relief workers	Excluding relief workers	Inflation/deflation (−) rate	Money supply (M2)	Real GDP	Velocity
1920	7,267	555	531	24	5.2	5.2				
1921	7,100	535	486	49	11.3	11.3	−14.7	−5.6	−0.4	−10.0
1922	7,303	404	330	74	8.6	8.6	−5.6	2.6	4.3	−4.1
1923	8,144	370	301	68	4.3	4.3	2.8	8.5	13.4	7.4
1924	8,289	369	277	92	5.3	5.3	−1.2	5.4	3.7	−2.8
1925	8,417	335	269	66	4.7	4.7	1.8	9.0	3.1	−3.7
1926	8,825	344	265	78	2.9	2.9	0.4	3.9	6.3	2.7
1927	8,857	367	261	106	3.9	3.9	−2.4	2.4	1.8	−3.0
1928	8,678	350	265	84	4.7	4.7	0.8	3.8	−0.8	−3.6
1929	9,173	342	277	65	2.9	2.9	0.2	0.4	6.9	6.7
1930	8,292	369	302	67	8.9	8.9	−3.6	−1.9	−8.6	−10.3
1931	7,702	313	360	−46	15.7	15.7	−10.4	−6.6	−6.4	−10.1
1932	6,652	218	527	−309	23.5	22.9	−11.7	−15.6	−13.1	−9.1
1933	6,516	231	531	−300	22.1	20.9	−2.6	−10.6	−1.5	7.3
1934	7,186	328	724	−395	20.4	16.2	5.4	6.6	11.0	9.7
1935	7,766	393	688	−296	20.1	14.4	2.1	13.7	8.8	−2.3
1936	8,701	415	875	−460	15.8	10.0	1.3	11.3	12.8	2.6
1937	9,120	492	767	−276	16.1	9.2	4.1	5.1	5.5	4.5
1938	8,718	566	685	−119	17.5	12.5	−2.8	−0.4	−3.7	−6.0
1939	9,321	504	896	−391	17.8	11.3	−0.9	8.3	7.8	−1.3

Sources and notes: Sources for per capita estimates in 2013 dollars for real GDP and federal government outlays, receipts, and surplus/deficit are described in notes to Fig. 1. The real GDP used to calculate the real GDP growth rates was constructed from the real GDP used to calculate per capita GDP. The unemployment rates are calculated from David Weir's estimates, series Ba473, Ba474, and Ba477 on pp. 2–82 and 2–83. The growth rate of the M2 money supply was calculated from series Cj45, pp. 3–605 and 3–606. The velocity was calculated by dividing the money supply by nominal GDP, and then the growth rate was calculated from the level which ranged from a high of 2.5 in 1920 to a low of 1.6 in 1932. The series are from Carter et al. (2006). The inflation/deflation rate is calculated from a series downloaded from Williamson and Officer (2014) at http://www.measuringworth.com/

Table 2 Labor and building statistics from the 1920s and 1930s

			Manufacturing			
	Share of civilian labor force		Earnings in 2013 dollars			
Year	Union members	Workers involved in strikes	Hourly	Weekly	Hours per week	Building permits (thousands)
1920	12.2	3.5	5.39	261.18	48.5	247
1921	11.2	2.6	5.46	247.75	45.4	449
1922	9.3	3.8	5.45	268.09	49.2	716
1923	8.4	1.7	5.81	285.72	49.2	871
1924	8.0	1.5	6.11	287.26	47.0	893
1925	7.9	0.9	5.99	289.02	48.3	937
1926	7.9	0.7	6.04	291.49	48.3	849
1927	7.8	0.7	6.27	299.82	47.8	810
1928	7.6	0.7	6.25	300.31	48.0	753
1929	7.6	0.6	6.36	307.79	48.4	509
1930	7.5	0.4	6.59	289.09	43.9	330
1931	7.1	0.7	7.04	282.37	40.1	254
1932	6.4	0.6	7.04	241.03	34.2	134
1933	5.6	2.3	7.13	257.17	36.1	93
1934	6.3	2.8	7.99	276.26	34.6	126
1935	7.1	2.1	8.08	299.95	37.1	216
1936	7.9	1.5	8.25	324.99	39.4	304
1937	13.4	3.5	8.89	342.97	38.6	332
1938	15.2	1.3	9.42	321.51	34.1	399
1939	16.3	2.1	9.56	359.09	37.6	458

Sources and notes. The number of union members is series Ba4783 (pp. 2-336, 2-337), civilian labor force is series Ba475 (pp. 2-82, 2-83), and workers involved in strikes is series Ba4955 (p. 2-354), hourly and weekly earnings are the National Industrial Conference Board estimates in series Ba4381 and Ba4382 (p. 2-279), and weekly hours were calculated by dividing the weekly earnings by hourly earnings. Building permits are series Dc510 (4-481). All series are from Carter et al. (2006). Earnings were converted to 2013 dollars using the GDP price deflator data downloaded from Williamson and Officer (2014) at http://www.measuringworth.com/

The statistics cannot do justice to how bad the economy was. Families ran through their savings and then still had to find ways to survive. As 2–3% of the nonfarm population lost their homes to mortgage foreclosures each year, some people moved in with extended family. A group of dispossessed lived in tent colonies or slept under newspapers renamed "Hoover blankets." Others wandered the countryside looking for work and food. The feature of American society that hit hardest was the optimistic spirit associated with the Horatio Alger stories. In the past, the watchword was to work hard and success would come your way. People who had worked hard all of their lives suddenly found themselves unemployed for long stretches of time. Pessimism about the future soon took hold, making it that much harder to spur a recovery.

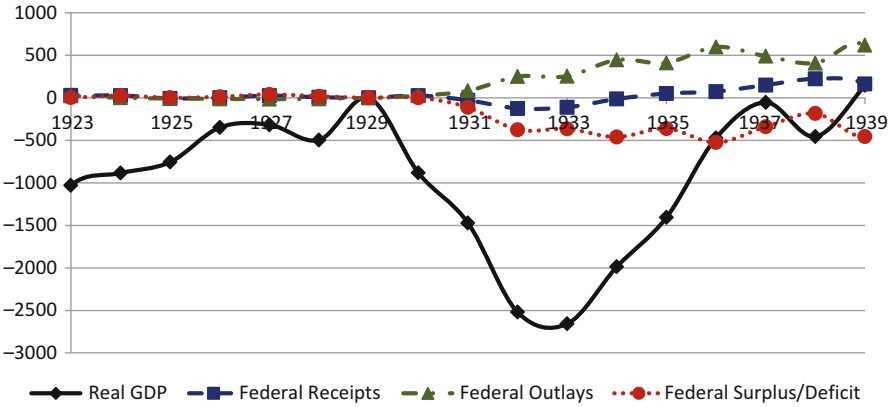


Fig. 1 Per capita gross domestic product and federal government revenue, outlay, and surplus/deficit value minus the 1929 value, (in 2013 Dollars), 1923–1939 (Source: Fiscal years ran from July 1 through June 30, such that the 1930 value covers the period July 1, 1929, through June 30, 1930. Federal government outlays receipts and the budget deficit/surplus are series Ea584, Ea585, and Ea586, and nominal GDP is series Ca10 from Carter et.al. (2006, pp. 5–80, 5–81, and 3–25). The deflator used to calculate the values for 2013 and resident population used to calculate per capita measures was downloaded from the Williamson and Officer (2014) website on October 10, 2014)

Fig. 2 Dow Jones Industrial Average Closing Price, October 1, 1928–December 31, 1933 (Source: Data downloaded from Dow Jones (2010) Historical Data on 4 June 2010)

Why?

Economists have plenty of answers but they do not all agree on how much weight to give each cause. The start of the Depression may have been a natural outcome of the boom and bust cycle in the economy. Investment expanded rapidly in the 1920s with the development and diffusion of new technologies like the auto, electricity, radios, and many new appliances. The boom in the stock market caused the Dow Jones Stock Index to rise from a low of 63 in 1921 to a peak of 381 in 1929. Construction of all types exploded. The number of building permits for housing in urban areas between 1922 and 1928 nearly doubled previous highs in the economy (Carter et al. 2006, pp. 4–481). The optimism that led to the investment boom was matched by the willingness of banks, insurance companies, and building and loans to lend. Stocks were sold on margin, consumers could buy new autos and appliances on installment plans, and mortgage loans expanded rapidly (Olney 1991). Some lenders packaged mortgages for resale to investors in mortgage-backed bonds. In many ways, the 1920s boom and the recession that started in 1929 match Joseph Schumpeter's (1939) description of increasing enthusiasm leading to overbuilding and over-investment followed by corrections that lead to recessions until the actual demand for goods catches up.

But such explanations can only really explain the start of the Depression. It makes sense, for example, that after the number of building permits (Table 2) peaked in 1925 at 937 thousand, nearly double the amount from any year before the 1920s, that number would fall back. By 1929, the number had nearly halved to a level that was still much higher than at the beginning of the 1920s. But what explains a decline to a low of 93 thousand in 1933? What explains an all-time high unemployment rate over 20% and the largest drop in output in history?

Since the Wall Street Crash occurred in late 1929, the timing seems to imply that the stock crash was a major cause of the Great Depression. The Dow Jones Stock Index (Fig. 2) peaked that year when it closed at 381 on September 3, 1929, soon after the recession had started in August. The most spectacular loss occurred when the Dow declined 24% from its previous Friday's close at 301 over Monday and Tuesday, October 28 and 29. It then dropped to a low of 198 on November 11. Stocks then recovered. By April 1930, the Dow was challenging the levels it had reached just before Black Tuesday. It then sunk in fits and starts to a low of 41.2 on July 8, 1932.

Most economists do not focus on the Stock Crash as a major cause of the Great Depression. One reason is that stock market values have fallen sharply on numerous occasions without consequent declines in the real economy. The spectacular drop in 1987 barely impacted the real economy and even the most recent decline of nearly 40% in the stock market in 2008 was followed by only one negative year of real output growth. Economists who assign the highest weight to the stock crash talk about how the crash led to greater uncertainty that caused consumers to cut back on purchases of durable goods like autos and refrigerators. Additionally, it created problems for stockholders who struggled to repay their loans or could no longer borrow, which in turn made it more difficult for banks to have enough funds to lend

for new investments. But a relatively small share of the population owned stocks at the time, so the vast majority had little invested in the stock market crash. One study of bond ratings found that investor confidence held up relatively well for a couple of years after the crash. The long length of the stock market's fall to its extremely low 1933 depth suggests that the market might well have been responding to the changes in the economy rather than being a cause of the decline in the economy.[2]

Explanations of the Depression's causes often lead to answers that still leave much to be explained. Speculations about a decline in consumption as a prime cause just pushed the question back one level because scholars know little about why consumers bought less. Others have focused on a series of negative productivity shocks on the supply side of the economy as causes. Yet, no one has had much success in identifying the exact nature of these shocks. Others point to increased uncertainty.[3]

Whatever was happening in the private economy, it was not helped by the economic policies chosen by the Congress, President Hoover, and the Federal Reserve Board. Nearly everybody agrees that the Federal Reserve Board's monetary policy helped turn a recession into a major Depression. The primary disagreements center on how much blame the Fed deserved and why they followed such an inadequate policy. The economy of the early 1930s was the Federal Reserve's first great test. Sadly, the Fed failed it.[4]

Through 1935, the Fed had two major tools for influencing the money supply. They could buy and sell existing bonds in "open market operations." They could also adjust the "discount rate" at which member banks borrowed funds from the Fed to meet reserve requirements. In response to bank failures in a panic or a general downturn, the Fed could increase the money supply and stimulate the economy by buying bonds and/or lowering the discount rate.

In making policy, the Fed also had to pay attention to the international gold standard. To remain on the gold standard, the Federal Reserve and US banks had to

[2]Peter Temin (1976) uses bond ratings to show that investor confidence held up well for a couple of years after the crash. Christina Romer (1990) and Frederic Mishkin (1978) of modern economists assign the largest role to the stock crash. For other readily readable discussions of the causes of the Great Depression, see Smiley (2002) and Randall Parker's (2002, 2007) volumes of incisive interviews with many of the leading economists who have written on the Depression.

[3]For the consumption arguments, see Temin (1976) and Romer (1990). For negative productivity shocks, see the work of Ohanian (2001) and Cole et al. (2005) and sources cited there. For uncertainty, see Flacco and Parker (1992).

[4]Friedman and Schwartz (1963) led the way in developing this monetarist argument. Bernanke (2000) provides additional arguments based on breakdowns in lending channels as the money supply shrunk. The argument was debated heavily in the 1970s and 1980s, and the debate is nicely summarized in Atack and Passell (1994). The monetarist analysis received support using dynamic general equilibrium analysis from a study by Bordo et al. (2000). Some challengers became more accepting of the argument when it was tied to reliance on the gold standard. Real business cycle economists tend to give less weight to the monetarist argument, but real business cycle economists Cole et al. (2005) assign as much as 33 % of the blame for the Depression internationally to monetary shocks. See also Chari et al. (2002). For a summary of more recent works from the 1990s and 2000s, see Fishback (2010).

stand ready to pay an ounce of gold for every $20.67 in Federal Reserve notes. This meant holding adequate US gold reserves to make the promise believable. If changes in the relative attractiveness of the dollar led the US supply of gold to fall below the appropriate level, the Fed was expected to take actions to make the dollar more attractive. At the time, the standard policies in response to gold outflows included raising the discount rate and selling (or at least reducing purchases of) existing bonds.

In an attempt to slow the speculative boom in stocks, the Federal Reserve policy in 1928 and 1929 aimed at slowing the growth of the money supply (Hamilton 1987). Over the next 4 years, there were a series of negative shocks to the money supply, including the stock market crash in 1929; banking crises in 1930–1931, 1931, and 1932–1933; and Britain's abandonment of the gold standard in September 1931. The Federal Reserve's response to these crises might best be described as "too little, too late," as it allowed the money supply to fall by 30%.

The policy makers at the Fed thought they had applied a great deal of stimulus to the money supply when they cut the nominal discount rate in eleven steps from 6% in October 1929 to 1.5% in 1931. The rates seem low but had little effect because rapid deflation raised the value of dollars that borrowers had to repay in ways that caused the real discount interest rate to rise as high as 10.5%. In the minutes of their meetings, the policy makers did not mention the impact of deflation on real interest rates as a concern (Meltzer 2003).

When Britain left the gold standard in 1931, the Fed felt that it had to stop an outflow of gold to Britain by raising the discount rate back to 3.5% in late 1931. Adjusted for deflation, the discount rate in real terms reached 14%. Even though the rate was lowered again, deflation left the real discount rate at 15.2% at one point in 1932. This is nearly triple the next highest real discount rate of 5.8% that has been reached since 1933. The deflation, itself a result of the fall in the money supply, made the discount rate a nearly useless tool for the Federal Reserve.

The Fed's other alternative was to make "open market purchases" of bonds. Its boldest move was an open market purchase of $1 billion in bonds over several months in the spring of 1932. By that time, output in the economy had sunk by 30% and the unemployment rate was over 20%. Had the Fed offset the banking crises in 1930–31 with a $1 billion purchase, the damage likely would have been limited and the murkiest depths of the Depression avoided.[5]

Why was the Federal Reserve so recalcitrant? In part, the Fed was following the same policies toward banks that they followed in the 1920s. Between 1920 and 1929, the Federal Reserve and state bank regulators allowed an average of 630 banks per year to suspend operations. Most of the banks were small and a good case could be made that they had made bad loans and investments that could not be salvaged. As the economy deteriorated between 1930 and 1933, banking problems worsened dramatically as bank runs led to a reduction in the number of banks from 25 thousand to 17.8 thousand. In

[5]This is the essence of Friedman and Schwartz's (1963) argument. See Christiano et al. (2003) for an analysis that emphasizes the role of monetary policy but not necessarily Friedman and Schwartz' "too little, too late" hypothesis.

many cases, the suspended banks looked as bad as the banks that failed in the 1920s. Fed policy was not uniform across the country. The Atlanta Fed had some success at staving off bank failures and declines in the southern economy by quickly providing large amounts of reserves to banks that faced runs. The quick backing allowed the banks to reassure all of their depositors, who then redeposited their money.[6]

The Fed's focus on keeping the USA on the gold standard also created significant problems. To combat the downturn, the Fed wanted to stimulate the money supply and thus the economy, but changes in international markets were forcing them to do the opposite to remain on the gold standard. Once the USA left the gold standard in 1933, the economy began to improve. The slow reaction to the bank panics during the 1930s was grouped with a series of policy blunders in other areas. The Hawley-Smoot Tariff Act of 1930 in the USA raised tariffs to levels that restricted US imports. Other countries responded by erecting their own barriers to trade, especially after Britain went off the gold standard in September 1931. The result was a downward spiral in world trade. In the USA, per capita exports and imports (Table 2) were cut in half by 1933. Tariffs led to a variety of inefficiencies in the economy, but the impact on real GDP per capita during the Depression was relatively small because exports were only about 6% of GDP in good times, while net exports were typically less than a half percent.[7]

Until the 1930s, wages and prices had typically declined during downturns. The declines in the past often had contributed to a surge in purchases of consumer goods and in hiring workers that helped turn the economy around. President Hoover held conferences in 1929 to ask leading manufacturers in the country to hold wages stable and run a job-sharing policy that cut weekly hours, so that workers would not lose their jobs. A number of large firms followed his dictate, waiting until the middle of 1931 before cutting hourly wages. In asking manufacturers to follow the job-sharing model, the Hoover administration hoped that keeping more workers employed with stable wages would leave them with enough buying power to keep the economy moving, despite their losses in weekly earnings. In 1932, he signed the Norris-LaGuardia Act, which outlawed several antiunion practices and gave unions more power to organize and hold the line on wages. Following their introduction, the economy continued to slide, and even union membership fell 11% over the next year.[8]

[6]Wheelock (1991) uses econometric methods to show that the statistical relationships between Fed policy instruments and the economic factors on which policy makers focused did not change between the 1920s and early 1930s. Wheelock (1991) and Richardson and Troost (2009) talk about differences in policies at the regional Federal Reserve Banks. For discussions of the declines in asset quality and bank suspensions, see Calomiris and Mason (2003).

[7]Irwin (2011) provides a detailed description of the political economy of the tariff and argues that the size of the tariff increase was not as large as many have stated.

[8]Ohanian (2009) argues that Hoover's jawboning was a major contributor to the Great Depression, arguing that employers followed the policies in part due to fears of the strength of unions. This is somewhat puzzling because union membership declined in the early 1930s, as seen in Table 2. See also Rose (2010) and Neumann et al. (2013) for more specifics about the Hoover policy.

Faced with increasing unemployment, President Hoover did not press for the federal government to start providing new welfare programs. Instead, he followed the path set by the long-term federal structure of governments. Since colonial times, responsibilities for caring for the poor and the disabled had resided in local governments. States began playing a bigger role after 1909 by establishing mothers' pensions, workers' compensation, aid to the blind, and old-age assistance.

Hoover also strongly believed in "voluntarism," as can be seen in his efforts to jawbone manufacturers into paying high wage rates. He thought the federal government should help organize efforts by others to resolve the problems. The government might help through loans. When faced by struggling farmers who demanded subsidies to control production and raise prices, the Hoover administration instead supported the provision of $500 million in loans through a Federal Farm Board. To aid the unemployed, he formed the President's Emergency Committee on Employment in 1930 to aid private organizations as they sought to help the poor. This group morphed into the President's Organization for Unemployment Relief in 1931. Eventually, in the summer of 1932, he signed a legislation to offer $300 million in loans to local governments to aid them in poor relief.

Faced with large-scale bank failures, Hoover convinced bankers to set up the National Credit Corporation (NCC) to aid troubled banks. When the NCC fell short, Hoover and the Republican Congress mimicked loan programs from World War I and created the Reconstruction Finance Corporation (RFC) in February 1932 to make loans to troubled banks, industries, and the farm sector. The bank loans were not effective at preventing suspensions because banks had to hold assets as collateral for the RFC loans and thus could not sell them to pay depositors if trouble arose. The RFC was more successful at saving the banks when it began taking short-run ownership stakes in the banks (Mason 2001; Calomiris et al. 2013). RFC ownership stakes set precedents for the moves by Treasury Secretary Henry Paulson and Fed Chair Ben Bernanke to take ownership positions in banks in the fall of 2008.

Hoover is often seen as a fiscal policy conservative. Compared to his Republican predecessors, however, he looks like a rabid spender. From 1921 through 1929, the Harding and Coolidge administrations had run budget surpluses (Fig. 1). They followed the standard pattern of repaying the debt run-up during World War I. In response to the worsening economy, Hoover and the Republican Congress increased real federal outlays per capita (Fig. 1 and Table 1) by 91% between 1929 and 1932. Herbert Hoover gets less credit for this rise than Franklin Roosevelt does for the New Deal because Hoover did not create new spending agencies, he just expanded existing programs by doubling federal highway spending and increasing the Army Corps of Engineers' river and harbors and flood control spending by over 40%.

Herbert Hoover believed in balanced budgets, as did Franklin Roosevelt during the New Deal. The debates between Hoover and the Congress in 1932 over how to raise taxes to balance the budget led to two major types of tax increases. The first was a "soak the rich" effort. Less than 10% of households earned enough to pay income taxes in the early 1930s. In the Revenue Act of 1932, the tax rate was raised for individuals earning more than $2,000 from 0.1% to 2%, and the rate rose from 0.9%

to 6% for incomes from $10 to $15 thousand. People with income over $1 million saw their tax rate rise from 23.1% to 57% (Carter et al. 2006, pp. 5–114).

The higher income tax rates did little to stem the drop in tax revenues between fiscal years 1932 and 1933 because receipts from income taxes and estate taxes fell to 37%.[9] Some of the fall was due to tax avoidance by the very rich and the rest was due to the continued deterioration in the economy. New excise taxes helped make up the shortfall in income tax revenue, so that per capita federal revenue stayed roughly the same in 1932 and 1933 (Fig. 1 and Table 1). The Revenue Act tacked on new excise taxes on oil pipeline transfers, electricity, bank checks, communications, and manufacturers – particularly autos, tires, oil, and gasoline 0020 (Commissioner of Internal Revenue 1933, pp. 14–15). Unfortunately, these new taxes contributed to retarding the development of some of the new industries that might have led a recovery.

As the economy dove toward the depths of the Depression, Herbert Hoover and the Republican Congress offered a variety of new policies to combat the problems. Some, like the Hawley-Smoot Tariff and the Federal Reserve's inaction, were disastrous, not only for the USA but for the world economy. The job-sharing policies were probably misguided and the efforts at voluntary organization floundered in the face of such a deep Depression. The Hoover administration offered a wide range of subsidized loans and even ramped up federal spending to shares of GDP not seen in peacetime. No matter what Hoover threw at the Depression, nothing stemmed the tide. In consequence, he and the Republican Congress lost power to Franklin Delano Roosevelt and the Democrats in a landslide during the 1932 Election.

Between Roosevelt's November landslide victory and his inauguration on March 4, 1933, the US economy's tailspin deepened. Industrial production, prices received by farmers and producers, real weekly manufacturing wages, and the share of corporations earning profits all bottomed out. Industrial production reached a low that had not been seen since the sharp recession in the spring of 1921 and since 1915 before that. The unemployment rate hovered around 25%, while average weekly work hours fell below 35 for the second time.

The banking sector went through another wave of failures as 633 banks suspended payments from December 1932 through February 1933. Roosevelt and Hoover disagreed over how to deal with the suspensions during the winter. Hoover pressed Roosevelt to agree to Hoover's recommended policies but would not act without Roosevelt's approval. Not wanting to be saddled with Hoover's policies, Roosevelt refused consent and decided to wait and set his own policies after the inauguration. Meanwhile, state governments had begun declaring bank holidays and restrictions on deposits paid out. By March 4, every state and Washington, DC, had imposed some type of restriction (Wicker 1966, p. 153). It remained to be seen what the new administration could do to turn the tide.

[9]Ellen McGrattan (2012) develops a model that shows the negative consequences of taxes on corporate income and dividends in the early 1930s.

The New Deal and Partial Recovery

"This Nation calls for action, and action now," Franklin Roosevelt declared during his inaugural speech on March 4, 1933.[10] Two days later, he announced the National Banking Holiday. Within 100 days, Roosevelt and the Democratic Congress had established a "New Deal for the American public" that developed into the largest peacetime expansion of federal government activity in American history.

Over the next 7 years, Roosevelt and the Democratic Congress tried government solutions to dozens of problems in the American economy. When they saw a problem, they tried to fix it with more spending or new government regulation. But in many cases, a policy designed to fix one problem contradicted the fix for another problem. For example, when they tried to raise prices in the farm sector by limiting production, they contributed to increased unemployment among farm workers while also raising prices for food for workers and the unemployed, leading to reductions in their standard of living.

Measuring the Recovery

The trough for the economy was so deep that growth rates coming out of the recovery were relatively rapid until a second-dip recession occurred in 1937–1938. In 1933, real GDP per person (Table 1 and Fig. 2) was about 29% less than its 1929 figure. Real GDP per person neared its 1929 level again in 1937 but fell back in 1938 before finally surpassing the 1929 level in 1939. An entire decade was spent with less output per person than in 1929. The shortfall was even worse when the long-run growth path is considered. Had real GDP per person risen at its long-run average growth rate of 1.6% per year, GDP per person in 2009 dollars would have been more than $1,000 higher than the $9,112 level reached in 1939.

One factor that might have taken its toll on the private economy was the high degree of uncertainty about what the government planned to do. The New Deal went through multiple phases, as the activities of the Agricultural Adjustment Administration (AAA) and the National Recovery Administration (NRA) were struck down by the courts in 1935 and the Roosevelt administration tinkered with the new regulations, new taxes were introduced then removed, and temporary agencies received new extensions. Such uncertainty can wreak havoc when businesses and workers are making longer-term decisions (Higgs 1997).

The unemployment rate did not recover as well as real GDP did. Table 1 shows two measures of the unemployment rate for the 1930s that differ based on whether people on work relief are categorized as unemployed or employed. Either measure shows unemployment rates during the New Deal that are among the highest in America's economic history. Since 1890, the unemployment rate has risen higher

[10]See the Inaugural Address of Franklin Delano Roosevelt (1933), downloaded on 10 June 2010, from http://www2.bartleby.com/124/pres49.html.

than 10% only during the Depression and in two other years, 1921 and 1983 (Carter, et al. 2006, 2–82 and 2–83).[11] Between 1933 and 1937, the unemployment rate dropped, but the Fed's increase of reserve requirements, the balancing of the federal budget deficit (Fig. 2), and a variety of other shocks to the economy contributed to a sharp spike in the unemployment rate in 1938. Not until 1942 did the unemployment rate fall into a more normal range.

One bright spot while digging out of the Depression was a surprisingly rapid rise in "productivity," a measure of output relative to the inputs put into the process. Alex Field (2003, 2011) has described the 1930s as "the most technologically progressive decade of the century." Partly, the rise in productivity came from the large public investments in roads, dams for electricity, highways, sanitation works, and airports that had begun earlier and been expanded under the New Deal. Many of these investments set the stage for greater productivity during World War II and the postwar boom. Some of the improvements came from firms faced with high wage rates and limited working time, finding new ways to organize their workers and raise output per man-hour. Some of the progress came from investments in research and development laboratories by businesses in the 1920s that bore fruit in the 1930s. Much of the research was based on basic science in chemistry and engineering that led to new uses of electricity, new fabrics, and new household appliances. The television, which dominated communications in the last half of the century, was being readied for commercial applications, but the War slowed the diffusion (Field 2003). In agriculture, new hybrid seeds, tractors, autos, trucks, and fertilizers began to diffuse widely. The usage was likely stimulated in part by farmers who sought to raise productivity on the acres that they had not taken out of production in response to the AAA payments.

In 1940 and 1941, the economy continued to recover. The USA may have benefited some from the disastrous war that had begun raging in Europe. Demand rose for US production of military hardware, food, clothing, and other necessities, as European production of many items was stunted by the Nazi invasions and the bombings in Britain (Gordon and Krenn 2010). Consumption in the USA was not growing as fast as output in part because the USA had begun shifting a number of factories to the production of munitions. The goal was to aid the Allies through Lend Lease and to prepare for the eventuality that the USA might enter the war. Once

[11]One of the thorniest issues for studying unemployment rates in 1930s is whether to define as employed or unemployed the people on emergency work relief programs, like the Federal Emergency Relief Administration from 1933 through 1935 and the Works Progress Administration from 1935 through 1942. The relief workers worked for their relief payments at hourly wages that were roughly half to two-thirds of the norm paid by other government projects. Since 1940, the unemployment statistics have treated people receiving unemployment benefits of similar size as unemployed and they have no work requirement. This suggests to me that the New Deal relief workers were worse off than modern people on unemployment insurance because they received the same benefits as the modern people but had to work for them. Thus, I believe the relief workers should be considered unemployed in comparison with modern unemployment rates. See Darby (1976) and Neumann et al. (2010) for more discussion of this issue.

Japan bombed Pearl Harbor, however, the American economy shifted swiftly to a wartime command economy.

Measuring the Success of the New Deal Policies

Roosevelt's New Deal led to enormous institutional changes that have carried into the twenty first century. The Federal Government took responsibility for insuring against a wide variety of potential crises, several emergency programs, new regulations, and social insurance programs. Economists and economic historians have been studying how much the economic policies contributed to the recovery and how they changed the structure of the economy.

Monetary Policies

Within 2 months of taking office, Roosevelt and the Federal Reserve had completely reversed the monetary policies of the early 1930s. Their goal was to shift expectations from continued deflation to anticipated inflation. Roosevelt announced the goal of higher prices for farmers and producers and higher wages for workers on numerous occasions. The National Bank Holiday closed all banks and thrift institutions temporarily, while auditors examined the banks. Banks declared sound were soon reopened. Insolvent banks were reorganized, some with Reconstruction Finance Corporation backing. These seals of approval for the reopened banks helped change expectations about the solvency of the bank system.[12]

By June, Roosevelt had taken the USA off of the gold standard and appointed Eugene Black from the Atlanta Fed as the Chair of the Federal Reserve Board. Under Black, who had saved many southern banks facing bank runs by flooding them with cash, the Fed began to focus on monetary expansion (Richardson and Troost 2009). Between April and May, the discount rate was cut from 3.5% to 2.5%. It then fell to 2% by the end of the year, to 1.5% in 1934, and then to 1% in 1937. The return of inflation meant that the real discount rate through 1937 was negative, a sharp contrast to the double-digit positive real discount rates during the Hoover era.

The move off of the gold standard and the devaluation of the dollar to $35 dollars per ounce of gold combined with political events in Europe to cause a flow of gold into America. The economy began to recover. This same pattern was repeated throughout the world. In country after country, as central banks sought to maintain the gold standard, their domestic economies continued to sink. As each left the gold standard, their economies rebounded.[13]

[12]For a description of the National Bank Holiday, see Mason (2001). For discussions of the reversal of policy to fight deflation, see Temin and Wigmore (1990) and Eggertsson (2008).

[13]Eichengreen (1992), Temin (1989), Temin and Wigmore (1990), and Kindleberger (1986) talk extensively about the move off of the gold standard.

The Fed's emphasis on raising the money supply lasted through three years of recovery, as real GDP per capita neared its 1929 level in 1937 and the unemployment rate fell toward 14% (Fig. 1 and Table 1). In 1935, the Fed gained control of an additional policy tool, "reserve requirements," the share of deposits banks were required to hold in reserve. Unfortunately, by August 1936 the Fed had begun to fear that banks were holding large excess reserves above and beyond the required reserves. Fearing that the banks would begin lending the excess reserves, raise the money supply, and create rapid inflation, the Fed doubled the long-standing reserve requirements in three steps on August 16, 1936; March 1, 1937; and May 1, 1937. The Fed had failed to recognize that the banks were holding so many excess reserves to protect themselves against bank runs. Their experience of the past decade had given them little confidence that the Fed would act as a lender of last resort. Therefore, the banks increased their reserves to make sure that they retained some excess reserves as a cushion above the newly doubled requirements. These changes were followed by a spike in unemployment to 19% in 1938 (Fig. 2) and a decline in real GDP per person back to its 1936 level (Fig. 1).[14]

Fiscal Policy

The myth that seemingly will not die is the idea that Roosevelt followed the doctrines of John Maynard Keynes in using government spending to stimulate the economy. In *The General Theory of Employment, Interest and Money* from 1935, Keynes (1964) argued that the economy can settle into an equilibrium at less than full employment, particularly when there are factors blocking wage and price adjustments. Increases in government spending and reductions in taxation that lead to larger budget deficits are methods for pushing the economy toward full employment. Because the New Deal ramped up government spending under the New Deal, people have mistakenly presumed that Roosevelt followed a Keynesian policy.

Although by 1939 the Roosevelt administration had raised real per capita government outlays by nearly 70% (Fig. 1), the per capita tax revenues rose at roughly the same pace. As a result, the budget deficits in Fig. 1 do not look much different from the deficits the Hoover administration ran in fiscal years 1932 and 1933. Economists and economic historians, including Keynes himself, have long known that Roosevelt did not follow Keynes' dictates. Keynes even wrote an open letter to President Roosevelt published in newspapers in late December 1933 saying that the increased spending was good but the increase in tax revenues was reducing the stimulative effect.[15]

[14]This description is based on Friedman and Schwartz (1963). For a view that puts less emphasis on the Fed's role, see Romer (1992). Calomiris et al. (2011) use data for individual banks to challenge the idea that the reserve requirement rise had a strong impact on the 1937–1938 recession. For a dynamic model of the 1937–1938 recession, see Eggertsson and Pugsley (2006).

[15]See Brown (1956) and Peppers (1973) for more sophisticated analysis showing that the New Deal did not follow the Keynesian policy. Barber (1996) describes the economic thinking of many of the New Deal advisors.

Both federal outlays and the size of deficits fell well short of the Keynesian recommendation, given the size of the shortfall in real GDP. The bottom line in Fig. 1 is the difference in real dollars between GDP per person in the year marked on the graph and in 1929. The GDP per person shortfall was $1,987 in 2013 dollars in 1934. Government outlays in that fiscal year were only about $436 per person more than they had been in 1929 and the deficit was only $455 per person more negative than it had been in 1929. To even begin to be close to the size of a Keynesian stimulative policy designed to get to full employment, the deficit would have needed to be at least 3 times as large and probably larger.

Roosevelt's tax rate policies made matters worse by chilling incentives for investors. After some minor tinkering in 1934, income tax rates were raised again for people earning over $100,000 in 1936. The top rate went from 57.2% to 68% for people earning over $1 million. The National Industrial Recovery Act of 1933 instituted a tax on capital stock, dividends, and excess profits that was collected through the rest of the 1930s. In 1936, a surtax was added on profits that were not distributed as dividends. None of these new tax rates generated a great deal of tax revenue, but they did create the wrong incentives for investment. Sadly, the companies least able to avoid the highest marginal rate of 27% on undistributed profits were smaller, faster-growing firms that were not yet able to obtain external financing (Calomiris and Hubbard 1993). Most of the growth in tax revenues came from natural increases in tax revenues as the economy recovered, a temporary processing tax on agricultural goods that ended in 1935, and the renewed collections of alcohol taxes after the end of Prohibition.

The Roosevelt administration's best tax policy was its relaxation of some of the tariff barriers imposed in 1930. The Reciprocal Trade Agreement Act of 1934 freed Roosevelt to sign a series of tariff reduction agreements with Canada, several South American countries, Britain, and key European trading partners. As a result, American imports rose from a 20-year low in 1932–1933 to an all-time high by 1940.[16]

Alphabet Soup

The New Deal combined expansions in federal spending and regulatory roles in a proliferation of acronyms for new agencies. Some were temporary, like the Federal Emergency Relief Administration (FERA), Civil Works Administration (CWA), and the Works Progress Administration (WPA) relief agencies, but the majority became permanent parts of the economic landscape.

The agencies that distributed the most grant money provided relief to the poor, built public works, and paid farmers to take land out of production. When the FERA was established during the first 100 days, the federal government took responsibility

[16]For historical comparisons of the impacts of tariff rates, see Irwin (1998). Kindleberger (1986, p. 170) and Atack and Passell (1994, p. 602) describe the international trade developments in the 1930s.

for the first time for aiding the poor and the unemployed.[17] The FERA provided both direct relief payments and work relief jobs through 1935, while the CWA lasted 4 months in the winter of 1933–1934. In 1935, the responsibility for "unemployables" was returned to the state and local governments and the WPA took over the provision of work relief. Meanwhile, the Public Works Administration (PWA), Public Roads Administration (PRA), and Public Building Administration (PBA) were new agencies that continued the federal government's role in funding the building of large dams, federal highways, federal buildings, and improvements to federal lands while also aiding the state and local governments in building their own projects.

A series of studies in the past 10 years show, on net, that the public works and relief spending were beneficial to the communities where they were built. Nearly all of the studies are based on panel data sets with multiple years of data for each location. The methods for identifying the effects typically examined the impact of changes over time within the same location after controlling for nationwide shocks to the economy. They used methods to avoid negative feedback effects that arose from the government providing more funds in areas where the economy was bad. An additional dollar of public works and relief spending in a state raised income in the state by between 67 cents and $1.09. Areas with more public works and relief projects saw increases in retail sales, drew more internal migrants, experienced less crime, and had lower death rates from infant mortality, suicides, and infectious disease. The one area where public works and relief spending had no positive impact was on raising private employment, which may help explain why the unemployment rate remained so high throughout the decade.[18]

The Social Security Act of 1935 established a new long-run set of social insurance institutions. The new old-age security pension program, what people now call social security, called for taxes on employers and workers to finance pensions for retired workers. Unemployment insurance (UI) required that employers pay into funds that would provide benefits when their workers became unemployed. Although many states had already set up programs to aid widows with children, the poor elderly, and the blind, the Social Security Act helped expand these programs by federal government-provided matching grants that improved benefits and gave states incentives to create the programs if they had not already. The states spent the most on the needs-based old-age assistance programs. These programs encouraged the elderly to live on their own and retire, although they did not have much impact on the death rates of the elderly.[19]

[17]The federal government had long provided benefits and disability payments for its administrative employees, soldiers, and veterans.

[18]See Wallis and Benjamin (1981), Benjamin and Mathews (1992), Fleck (1999), Fishback (2015), Fishback et al. (2007), Fishback et al. (2005, 2006), Fishback and Kachanovskaya (2015), Johnson et al. (2010), Neumann et al. (2010), and Garrett and Wheelock (2006). For a dataset with federal spending by state in the 1930s, see Fishback (2015). For datasets at the state, city, and county level, see Price Fishback's website at the University of Arizona Economics Department http://econ.arizona.edu/faculty/fishback.asp. For a general survey, see Fishback and Wallis (2013).

[19]See Costa (1999), Stoian and Fishback (2010), Parsons (1991) Balan-Cohen (2009), and Friedberg (1999).

In contrast, the AAA farm program was specifically designed to reduce output and raise farm prices with an aim to raise a farmer's incomes from a decade of doldrums. In the final analysis, the AAA led to a significant redistribution of benefits to landowning farmers away from consumers, farm workers, and some farm tenants. Large farmers were the chief beneficiaries of the payments and whatever price boosts occurred. The reduction in land under cultivation generally reduced the demand for labor and thus made it more difficult for farm workers and sharecroppers to find work. Recent estimates of the local effect of the AAA show that counties receiving more AAA spending saw no change or a negative effect on overall economic activity and experienced some out-migration.[20]

The financial disaster led to a variety of new financial regulations. Since the 1930s, the SEC has monitored the stock markets, set reporting requirements for firms issuing stock, combated insider trading, and enforced rules on market trades. To stem the tide of future bank runs on deposits, the FDIC and FSLIC provided federal government insurance of deposits in banks and savings and loans. Limits were set on the types of investments that could be made by commercial banks and savings and loans. Regulation Q prevented payment of interest on checking accounts.[21]

In the moribund housing sector, states had tried to prevent foreclosures with moratoria laws that allowed home and farm owners to delay payments on their mortgages. The laws had the unfortunate effect of raising the risk of making loans because lenders could not be sure the states would not prevent repayments again (Rucker and Alston 1987). The result after the moratoria were eliminated was higher interest rates and much more restricted lending during the recovery. The Home Owners' Loan Corporation (HOLC) bought over one million mortgages that were in danger of foreclosure "through no fault of" the homeowner and then refinanced them on generous terms. The purchases almost fully replaced the bad loans on the lenders' books while helping about 80% of the borrowers remain in their homes. Given the uncertainty about the program, the up-front subsidy to housing markets probably was as high as 20–30% of the value of the loans, although after the fact that the HOLC only had losses equal to about 2% of the loans. The program also helped stave off further drops in housing prices and homeownership.[22] In 1934, the Federal Housing Administration was formed to offer federal insurance for mortgage loans both for new and existing homes and for repair and reconstruction. In 1938, Fannie Mae was established as a government corporation to provide a secondary market for mortgage loans in which banks could sell the loans as assets and then use the funds to make new mortgages.

[20]See Fishback et al. (2005, 2006), Depew et al. (2013), Fishback et al. (2003), and Fishback and Kachanovskaya (2015).

[21]For detailed accounts of the banking regulations, see Mitchener and Richardson (2013), Calomiris (2010), and Mason and Mitchener (2010).

[22]For analysis of the HOLC, see Courtemanche and Snowden (2011), Fishback et al. (2011, 2013), Harriss (1951), and Rose (2011).

The most controversial economic agency created by the New Deal was the National Recovery Administration (NRA). Between 1933 and 1935, when it was declared unconstitutional, the NRA fostered the development of "fair" codes of competition in industry. Industrialists, workers, and consumers in each industry were expected to meet and establish rules for minimum prices, quality standards, and trade practices. The workers were to be protected by minimum wages, limits on work hours, and rules related to working conditions in ways that looked like the job-sharing proposals Hoover had made earlier. Section 7a of the NIRA established a standard language for the codes that gave workers the right to bargain collectively through the agent of their choice. Once the code was approved by the NRA, the codes were to be binding to all firms in the industry, even those not involved in the code writing process. One of several goals was to prevent the "destructive" competition that some of Roosevelt's advisors believed had caused the deflation. The advisors expected that firms allowed to raise prices would sell more. Hour limits were put in place to allow more workers to remain employed, while the wages were raised to help reduce the losses in weekly earnings from the cut in hours.

The NRA might have had a beneficial macroeconomic effect to the extent that it contributed to shifting people's expectations from deflation to inflation.[23] However, from a microeconomic perspective, the NRA was the antithesis of antitrust policy at any other time in American history. The US law had always banned cartels and price-fixing agreements in restraint of trade. Suddenly, the federal government gave industry leaders antitrust exemptions and cartel-like powers to set prices, wages, and output. Many were written by industry trade groups with little input from unions, which were relatively weak at the time. Even worse, the federal government became the enforcer called on to prevent the natural tendency for firms to break away from cartel agreements. A recent study of the timing of the industry codes and their impact on the industries served to raise hourly wages and lowered hours worked in ways that cut the average weekly wage. When the Supreme Court struck down the NRA as unconstitutional in the Schechter Poultry case in 1935, no one was sorry to see the NRA go. Unlike the AAA, which was quickly reintroduced in a revised form after it was declared unconstitutional, there was little support for reenacting the codes of competition from many quarters and the Roosevelt administration let it die.[24]

After the NRA was struck down, the National Labor Relations Act of 1935 reinstituted the section 7A right of labor unions to organize and collectively bargain.

[23]See Temin and Wigmore (1990) and Eggertsson (2008, 2012) for this argument.

[24]Bellush (1975) offers a good administrative history of the NRA. Cole and Ohanian (2004) find that the high-wage policies and retrenchment in antitrust action associated with the NRA and the Roosevelt administration's post-NRA policies significantly slowed the recovery. Alexander (1997), Taylor (2007), and Vickers and Ziebarth (2014) discuss the problems the industries had in establishing the codes of "fair" competition and the reasons why businesses did not press for a new NRA when it was declared unconstitutional. Jason Taylor (2011) studied the impact of the NRA and the President's Reemployment Agreement on hourly wages, weekly wages, weekly hours, total hours employed, and industry output. Alexander and Libecap (2000) describe the different attitudes toward replacing the NRA and the AAA.

An employer was required to negotiate with a union if a majority of its workers voted to unionize. A National Labor Relations Board was established to monitor elections and arbitrate collective bargaining disputes. After the Act was affirmed as being constitutional in the spring of 1937, there was a surge in union recognition strikes, and the number of members rose from 4.2 million in 1936 to 8.3 million in 1938 (see Table 2).[25]

Conclusions

The Great Depression was the worst economic disaster in the American economic history. The annual output fell to 30% below the previous high and unemployment rates topped 10% for most of the decade and 20% in 4 years. Depression studies describe a variety of causes, including mistaken Federal Reserve policies tied to the gold standard and inadequate attention to deflation, uncertainty and damaged balance sheets related to the stock market crash, the Hawley-Smoot tariff, unexplained drops in consumption, negative productivity shocks, and labor market policies. Scholars still disagree on how much weight to give to each. In response to the contraction from 1929 to 1932, President Herbert Hoover and the Republican Congress nearly doubled real government outlays; distributed loans to banks, industry, and local governments; and tried several voluntary measures. Seeking to balance the budget in the fiscal year 1933, they sharply increased income tax rates and did not increase federal outlays any further, and the economy worsened even more.

Inaugurated in March 1933, President Franklin Roosevelt joined a newly elected Democratic majority in Congress to develop a large number of new regulatory and spending programs that they described as a New Deal for the American people. The USA left the gold standard and the reins on the money supply were loosened, although they were tightened again when the Fed doubled reserve requirements in three steps after 1935. The Roosevelt administration also roughly doubled federal government outlays but raised tax revenues at roughly the same pace and never really followed a Keynesian stimulus policy. Real output per person grew rapidly while climbing out of the deep trough hit in 1933, but was slowed by a second recession in 1937–1938. It finally reached the 1929 level after 1939. In spite of the output growth, unemployment rates remained above 9% until 1941.

The New Deal programs had a mixed record of success. RFC loans to banks in 1932 were not very effective at preventing bank suspensions, although more success was met when the RFC took ownership stakes in banks. Public works and relief spending contributed to the increases in economic activity and helped to reduce a variety of death rates and crime rates, but did not stimulate private employment. The AAA farm program that paid farmers to take land out of production aided farm owners, primarily large farm owners, but at the expense of a significant number of farm workers, sharecroppers, and tenants who lost their positions. The National

[25]For an economic overview of the changes in union policy, see Freeman (1998).

Recovery Administration seemed to help reverse deflationary expectations but had strong negative microeconomic effects. The HOLC helped keep about 800,000 people in their homes and helped stave off drops in housing values and home ownership rates at relatively low ex post cost.

The focus here has been on the emergency programs created during the New Deal. The New Deal also created many programs that are still in place today and programs that set precedents for current policy responses.[26] Cliometric research on the New Deal continues and I am willing to bet that the breadth and depth of New Deal research will increase greatly over the next decade.

References

Alexander B (1997) Failed cooperation in heterogeneous industries under the national recovery administration. J Econ Hist 57:322–44

Alexander B, Libecap G (2000) The effect of cost heterogeneity in the success and failure of the New Deal's agricultural and industrial programs. Explor Econ Hist 37:370–400

Atack J, Passell P (1994) A new economic view of American history from colonial times to 1940, 2nd edn. Norton, New York

Balan-Cohen A (2009) The effect on elderly mortality: evidence from the old age assistance programs in the United States. Unpublished working paper. Tufts University

Barber WJ (1996) Designs within disorder: Franklin D. Roosevelt, the economists, and the shaping of american economic policy, 1933–1945. Cambridge University Press, New York

Bellush B (1975) The failure of the NRA. Norton, New York

Benjamin D, Mathews K (1992) U.S. and U.K. unemployment between the wars: a doleful story. Institute for Economic Affairs, London

Bernanke B (2000) Essays on the Great Depression. Princeton University Press, Princeton

Bordo M, Erceg C, Evans C (2000) Money, sticky wages, and the Great Depression. Am Econ Rev 90:1447–1463

Brown EC (1956) Fiscal policy in the 'thirties: a reappraisal. Am Econ Rev 46:857–79

Calomiris C (2010) The political lessons of depression-era banking reform. Oxf Rev Econ Policy 26:540–560

Calomiris C, Hubbard G (1993) Internal finance and investment: evidence from the undistributed profits tax of 1936–1937. NBER working paper no. 4288

Calomiris C, Mason J (2003) Fundamentals, panics, and bank distress during the depression. Am Econ Rev 93:1615–47

Calomiris C, Mason J, Wheelock D (2011) Did doubling reserve requirements cause the recession of 1937–1938? A microeconomic approach. NBER working paper no. 16688

Calomiris C, Mason J, Weidenmier M, Bobroff K (2013) The effects of reconstruction finance corporation assistance on Michigan's banks' survival in the 1930s. Explor Econ Hist 50:526–547

Carter S et al (2006) Millennial edition of the historical statistics of the United States. Cambridge University Press, New York

Chari V, Kehoe P, McGrattan E (2002) Accounting for the Great Depression. Am Econ Rev Pap Proc 92:22–27

Christiano L, Motto R, Rostagno M (2003) The Great Depression and the Friedman-Schwartz hypothesis. J Money Credit Bank 35:1119–1197

[26]For additional discussions of more New Deal programs and long-run changes relative to the past and present, see Fishback and Wallis (2013).

Cole H, Ohanian L, Leung R (2005) Deflation and the international Great Depression: a productivity puzzle. National Bureau of Economic Research working paper no. 11237
Cole H, Ohanian L (2004) New Deal policies and the persistence of the Great Depression: a general equilibrium analysis. J Polit Econ 112:779–816
Commissioner of Internal Revenue (1933) Annual report for the year ending June 30, 1933. GPO, Washington, DC
Costa D (1999) A house of her own: old age assistance and the living arrangements of older nonmarried women. J Public Econ 72:39–59
Courtemanche C, Snowden K (2011) Repairing a mortgage crisis: HOLC lending and its impact on local housing markets. J Econ Hist 71:307–337
Darby M (1976) Three and a half million U.S. employees have been mislaid: or, an explanation of unemployment, 1934–1941. J Polit Econ 84:1–16
Depew B, Fishback P, Rhode P (2013) New Deal or no deal in the cotton south: the effect of the AAA on the labor structure in agriculture. Explor Econ Hist 50:466–486
Dow J (2010) Dow Jones historical data. Data downloaded on 4 June from http://dowjonesdata.blogspot.com/2009/04/historical-dow-jones-data.html
Eggertsson G (2008) Great expectations and the end of the depression. Am Econ Rev 98:1476–1516
Eggertsson G (2012) Was the New Deal contractionary? Am Econ Rev 102:524–555
Eggertsson G, Pugsley B (2006) The mistake of 1937: a general equilibrium analysis. Monetary Econ Stud 24:1–41
Eichengreen B (1992) Golden fetters: the gold standard and the depression 1919–1939. Oxford University Press, New York
Federal Reserve Board of Governors (Various years) Federal reserve bulletin. Government Printing Office, Washington, DC
Field A (2003) The most technologically progressive decade of the century. Am Econ Rev 93(4): 1399–1413
Field A (2011) A great leap forward: 1930s depression and U.S. economic growth. Yale University Press, New Haven
Fishback P (2010) Monetary and fiscal policy during the Great Depression. Oxf Rev Econ Policy 26:385–413
Fishback P (2015) New deal funding: estimates of federal grants and loans across states by year, 1930–1940. Res Econ Hist
Fishback P, Kachanovskaya V (2015) The multiplier for the states in the Great Depression. With Valentina Kachanovskaya. J Econ Hist 75(1):125–162
Fishback P, Kollmann T (2014) New multi-city estimates of the changes in home values 1920–1940. In: White E, Snowden K, Fishback P (eds) Housing and mortgage markets in historical perspective. University of Chicago Press, Chicago, pp 203–244
Fishback P, Wallis J (2013) What was new about the New Deal? In: Crafts N, Fearon P (eds) The Great Depression of the 1930s: lessons for today. Oxford University Press, Oxford, pp 290–327
Fishback P, Kantor S, Wallis J (2003) Can the New Deal's three R's be rehabilitated? A program-by-program county-by-county analysis. Explor Econ Hist 40:278–307
Fishback P, Horrace W, Kantor S (2005) Did New Deal grant programs stimulate local economies? A study of federal grants and retail sales during the Great Depression. J Econ Hist 65:36–71
Fishback P, Horrace W, Kantor S (2006) The impact of New Deal expenditures on mobility during the Great Depression. Explor Econ Hist 43:179–222
Fishback P, Haines M, Kantor S (2007) Births, deaths, and New Deal relief during the Great Depression. Rev Econ Stat 89:1–14
Fishback P, Flores-Lagunes A, Horrace W, Kantor S, Treber J (2011) The influence of the home owners' loan corporation on housing markets during the 1930s. Rev Financ Stud 24:1782–1813
Fishback P, Rose J, Snowden K (2013) Well worth saving: how the New Deal safeguarded home ownership. University of Chicago Press, Chicago

Flacco P, Parker R (1992) Income uncertainty and the onset of the Great Depression. Econ Inq 30:154–171
Fleck R (1999) The marginal effect of New Deal relief work on county-level unemployment statistics. J Econ Hist 59:659–87
Freeman R (1998) Spurts in union growth: defining moments and social processes. In: Bordo M, Goldin C, White E (eds) The defining moment: the Great Depression and the American economy in the twentieth century. University of Chicago Press, Chicago, pp 265–296
Friedberg L (1999) The effect of old age assistance on retirement. J Public Econ 71:213–232
Friedman M, Schwartz A (1963) A monetary history of the United States 1867–1960. Princeton University Press, Princeton
Garrett T, Wheelock D (2006) Why did income growth vary across states during the Great Depression. J Econ Hist 66:456–466
Gordon R, Krenn R (2010) The end of the Great Depression 1939–41: policy contributions and fiscal multipliers. NBER working paper no. 16380
Hamilton J (1987) Monetary factors in the Great Depression. J Monetary Econ 19:145–169
Harriss CL (1951) History and policies of the Home Owners' Loan Corporation. National Bureau of Economic Research, New York
Higgs R (1997) Regime uncertainty: why the Great Depression lasted so long and why prosperity resumed after the war. Indep Rev 1:561–590
Irwin D (1998) Changes in U.S. tariffs: the role of import prices and commercial policies. Am Econ Rev 88:1015–1026
Irwin D (2011) Peddling protectionism: Smoot-Hawley and the great depression. Princeton University Press, Princeton
Johnson R, Fishback P, Kantor S (2010) Striking at the roots of crime: the impact of social welfare spending on crime during the great depression. J Law Econ 53:715–740
Keynes JM (1964) The general theory of employment interest, and money. A Harbinger Book Harcourt Brace and World, New York
Kindleberger C (1986) The World in depression 1929–1939, rev edn. University of California Press, Berkeley
Mason J (2001) Do lenders of last resort policies matter? The effects of the Reconstruction Finance Corporation assistance to banks during the Great Depression. J Financ Service Res 20:77–95
Mason J, Mitchener K (2010) 'Blood and treasure': exiting the great depression and lessons for today. Oxf Rev Econ Policy 26:510–539
McGrattan E (2012) Capital taxation during the U.S. Great Depression. Q J Econ 127:1515–1550
Meltzer A (2003) A history of the federal reserve volume I: 1913–1951. University of Chicago Press, Chicago
Mishkin F (1978) The household balance sheet and the Great Depression. J Econ Hist 38:918–937
Mitchener K, Richardson G (2013) Does "skin in the game" reduce risk taking? Leverage, liability and the long-run consequences of New Deal banking reforms. Explor Econ Hist 50:508–525
Neumann T, Fishback P, Kantor S (2010) The dynamics of relief spending and the private urban labor market during the New Deal. J Econ Hist 70:195–220
Neumann T, Taylor J, Fishback P (2013) Comparisons of weekly hours over the past century and the importance of work sharing policies in the 1930s. Am Econ Rev Pap Proc 102:105–110
Ohanian L (2001) Why did productivity fall so much during the Great Depression? Am Econ Rev Pap Proc 91:34–38
Ohanian L (2009) What—or who–started the Great Depression? J Econ Theory 144:2310–2335
Olney M (1991) Buy now, pay later: advertising, credit, and consumer durables. University of North Carolina Press, Chapel Hill
Parker R (2002) Reflections on the Great Depression. Edward Elgar, Northampton
Parker R (2007) The economics of the Great Depression: a twenty-first century look back at the economics of the interwar era. Edward Elgar, Northampton
Parsons D (1991) Male retirement behavior in the United States 1930–1950. J Econ Hist 51:657–674

Peppers L (1973) Full employment surplus analysis and structural change: the 1930s. Explor Econ Hist 10:197–210
Richardson G, Troost W (2009) Monetary intervention mitigated panics during the Great Depression: quasi-experimental evidence from a Federal Reserve District border 1929–1933. J Polit Econ 119:1031–1073
Romer C (1990) The great crash and the onset of the Great Depression. Q J Econ 105:597–624
Romer C (1992) What ended the Great Depression? J Econ Hist 52:757–84
Roosevelt FD (1933) Inaugural address of Franklin Delano Roosevelt. Downloaded on 10 June 2010 from http://www2.bartleby.com/124/pres49.html
Rose J (2010) Hoover's truce: wage rigidity in the onset of the Great Depression. J Econ Hist 70:843–870
Rose J (2011) The incredible HOLC: mortgage relief during the Great Depression. J Money Credit Bank 43:1073–1107
Rucker R, Alston L (1987) Farm failures and government intervention: a case study of the 1930s. Am Econ Rev 77:724–730
Schumpeter J (1939) Business cycles (abridged). McGraw Hill, New York
Smiley G (2002) Rethinking the Great Depression: a new view of its causes and consequences. Ivan R Dee, Chicago
Stoian A, Fishback P (2010) Welfare spending and mortality rates for the elderly before the social security era. Explor Econ Hist 47:1–27
Taylor J (2007) Cartel codes attributes and cartel performance: an industry-level analysis of the National Industrial Recovery Act. J Law Econ 50:597–624
Taylor J (2011) Work-sharing during the Great Depression: did the President's reemployment agreement promote reemployment? Economica 78:133–158
Temin P (1976) Did monetary forces cause the Great Depression. W.W. Norton, New York
Temin P (1989) Lessons from the Great Depression. MIT Press, Cambridge
Temin P, Wigmore B (1990) The end of one big deflation. Explor Econ Hist 27:483–502
Vickers C, Ziebarth N (2014) Did the National Industrial Recovery Act foster collusion? Evidence from the Macaroni industry. J Econ Hist 74:831–862
Wallis J, Benjamin D (1981) Public relief and private employment in the Great Depression. J Econ Hist 41:97–102
Wheelock D (1991) The strategy and consistency of Federal Reserve monetary policy 1924–1933. Cambridge University Press, New York
Wicker E (1966) Federal reserve monetary policy, 1917–1933. Random House, New York
Williamson S, Officer L (2014) Measuring worth. Data downloaded on 10 October from http://www.measuringworth.com/

Cliometric Approaches to War

Jari Eloranta

Contents

Abstract

This chapter is a review of the many perspectives from history, political science, sociology, and economics that economic historians have applied to the study of war. Here I first review some of the scholarship on the premodern period, especially the formation of European nation states and conflicts. It is fairly clear that Europeans emerged out of this period with a comparative advantage in violence, through technological innovations and repeated warfare. Fiscal innovation and expansion was a key part of this. The period of the revolutions and Napoleonic conflicts represented a change in the nature of warfare and the arrival of total war, as well as the industrial age. The period of the world wars represents perhaps the best represented area of study for economic historians as of late. New data and scholarship has shown the mechanics of mobilization and highlighted the importance of resources in deciding these conflicts. Conversely, the Cold War period has been relatively sparsely studied, at least from the perspective of conflicts or military spending. Given the availability of new data and the opening of many archives, it is highly likely that this state of affairs will change in the near

J. Eloranta (✉)
University of Helsinki, Helsinki, Finland
e-mail: jari.eloranta@helsinki.fi

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_4

future. Economic historians have clearly made an impact in the study of long-run phenomena such as state formation, empires, and democracy. Cliometrics is well suited to the study of such topics, given the new panel and time series techniques, the rapid development of computing power, and the many new online databases.

Keywords
Economic history · Defense Economics · Warfare · Cliometrics · Military spending · State formation

Introduction

The theoretical and methodological impact of the cliometric approach to economic history, embodying the application of economic theories and the use of quantitative methods, has been widely recognized in a number of studies which have brought modeling and quantification into the forefront of economic history analysis. In comparison, quantitative studies of war by economic historians are somewhat rare, although in the last 20 years or so, there have been many more such applications. If we take the world wars as an example, we can see how the approaches to the study of particular conflicts (or times of peace) have differed based on the scholar's theoretical and methodological leanings. Historians in general, especially diplomatic and military historians, have focused on understanding the origins of the world wars and the particular battles that took place – this has also applied to the study of other conflicts, big or small. Most of those studies have not paid adequate attention on the quantitative aspects of war; for example, Paul Kennedy's *The Rise and Fall of the Great Powers: Economic Change and Military Conflict from 1500 to 2000* (1989) contains no quantitative testing of the hypothesis of hegemonic overreach or credible numerical information to back up the findings (Eloranta 2003, 2005).

While economic historical treatments of the costs and impacts of wars have not been abundant, economists have not embraced the study of conflicts wholeheartedly either. For instance, the study of defense economics and military spending patterns is related to the immense expansion of military budgets and military establishments in the Cold War era. It involves the application of the methods and tools of economics to the study of government expansion in the post-Second World War era, and at least three features stand out: (1) the individuals and organizations involved (both private and public spheres of influence, e.g., in contracting); (2) the theoretical challenges introduced by the interaction of different institutional and organizational arrangements, in both the budgeting and the allocation procedures; and (3) the nature of military spending or, more correctly, national defense as a public good as well as its potential for destruction (Sandler and Hartley 1995). Most studies in the rather small field of defense economics have had a limited time span in their analyses, namely, from 1945 onward. The longer-run developments and historical issues have typically fallen outside the interest of defense economists, although many of the tools and theoretical insights are useful for long-run analysis too.

Political and conflict scientists, including peace sciences, often cover similar ground as defense economists, with a longer-run view of history, a focus on the causal factors behind the most destructive conflicts, and the determinants of state formation. One of the most significant efforts in these overlapping fields has been the *Correlates of War* (COW) project, which started in the spring of 1963. This project, and the researchers loosely associated with it, has had a big impact on the study of conflicts, not to mention its importance in producing comparative statistics (Singer 1979, 1981, 1990). Moreover, these contributions have had a lot to offer to the study of long-run dynamics of military spending and warfare. For example, according to Charles Tilly (1990), one of the key contributors in the study of state formation, *coercion* (a monopoly of violence by rulers and an ability to also wield coercion externally), and *capital* (the means of financing warfare) was the key elements in the European ascendancy to world domination in the early modern era. Warfare, state formation, and technological supremacy were all interrelated fundamentals of the same process.

In this chapter I review some of the applications of these interdisciplinary perspectives to the study of war, especially from the point of view of the findings that quantitative methods have brought forth. First, I analyze some of the scholarship on the Middle Ages and the early modern period, especially the formation of European nation states and conflicts. Then I review some of the perspectives on the French Revolution and Napoleonic Wars, in particular the changed nature of conflicts and the arrival of total war. I follow this by examining one of the most fruitful topics of study, especially from the perspective of quantitative approaches: the era of the world wars. Moreover, the Cold War period represents a relatively unexplored frontier for economic historians of conflict, although defense economists have done more work in this area. Finally, I discuss some studies of long-run processes, especially state formation, empires, democracies, and military spending.

Theme 1: Medieval and Early Modern Warfare

The emerging nation states of the early modern period were much better equipped to fight wars. On the one hand, the frequent wars, new gunpowder technologies, and the commercialization of warfare forced them to consolidate resources for the needs of warfare. On the other hand, the rulers eventually had to share some of their sovereignty to be able to secure required credit both domestically and abroad. The Dutch and the British were the best at this, with the latter creating an empire that spanned the globe on the eve of the First World War.

During the Middle Ages, following the collapse of the Roman Empire (or at least the Western half) and barbarian invasions, a system of European feudalism emerged, in which feudal lords provided protection for communities for service or price. After the first millennium ended, the command system was usually used to mobilize human and material resources for large-scale military ventures (France 2001). Most European societies, with the exception of the Byzantine Empire, paled in comparison with the splendor and accomplishment of the empires in China and

the Muslim world. However, it seems that the Christian Western Europe under the feudal system provided more stability and longer leadership tenures, which contributed to Europe's military resurgence (Blaydes and Chaney 2013). Furthermore, it was not until the twelfth century and the Crusades that the feudal kings engaged in larger-scale operations that required supplementing ordinary revenues to finance the conflicts. The political ambitions of medieval kings, however, were a precursor of things to come and led to short-term fiscal deficits, which made long-term credit and prolonged military campaigns difficult (Webber and Wildavsky 1986; Eloranta 2005).

Innovations in the waging of war and technology, especially gunpowder, arrived in Europe with a delay, which in turn permitted armies to attack and defend larger territories. This also made possible a commercialization of warfare in Europe in the fourteenth and fifteenth centuries as feudal armies had to give way to professional mercenary forces. The age of commercialization of warfare was accompanied by the rising importance of sea power as European states began to build their overseas empires. Portugal, the Netherlands, and England, respectively, became the "systemic leaders" due to their extensive fleets and commercial expansion in the period before the Napoleonic Wars. The early winners in the fight for world leadership, such as England, were greatly influenced by the availability of inexpensive credit, enabling them to mobilize limited resources effectively to meet military expenses (Thomas 1983; Modelski and Thompson 1988; Ferguson 2001; Eloranta 2003).

These earlier efforts at expansion of armies, nation states, and credit have been studied more and more recently by economic historians. While efforts such as the project and database constructed by Richard Bonney (see, e.g., Bonney 1999a)[1] and Philip Hoffman and Peter Lindert (see, e.g., Hoffman et al. 2002)[2] are illustrative of broader trends in fiscal development, they typically do not utilize econometric techniques or economic theory in the analysis explicitly. Rather, they offer data for others to use. One of the more interesting efforts that has looked at various episodes in history, including the medieval period, is the book by Brauer and van Tuyll, *Castles, Battles, and Bombs: How Economics Explains Military History* (2008), which argues that the events and outcomes in military history can be explained using logic derived from economic theory. Such explicit use of economic theory to explain military decision making has been rare in the literature so far. One of the issues they tackle is the location and layout of medieval and early modern castles. They use the concept of opportunity costs and sunk costs to explain the use of what were often inefficient armaments, since castles were often expanded outward, especially after the fourteenth century, as a response to the emergence of gunpowder weaponry. Moreover, they argue that diminishing marginal returns set in for bigger castles, although more remains to be studied in this respect.

What the existing literature suggests is that the newly emerging nation states began to develop more centralized and productive revenue–expenditure systems, the

[1]Coordinator of the *European State Finance Database*: http://www.esfdb.org/

[2]Founders of *The Global Price and Income History Group*: http://gpih.ucdavis.edu/

goal of which was to enhance the state's power, especially in the absolutist era. This also embodied a growing cost and scale of warfare. For example, during the 30 Years' War between 100,000 and 200,000, men fought under arms, whereas 20 years later 450,000–500,000 men fought on both sides in the War of the Spanish Succession. The numbers notwithstanding, the 30 Years' War was a conflict directly comparable to the biggest global conflicts in terms of destruction. For example, Charles Tilly (1990) estimated the battle deaths to have exceeded two million. Henry Kamen, in turn, emphasized the mass-scale destruction and economic dislocation this caused in the German lands, especially to the civilian population (Kamen 1968; Tilly 1990). In many ways this was total war, especially considering the high numbers of civilian casualties. Cliometricians have not yet really tackled the scale and scope of this conflict adequately, using modeling and other tools frequently utilized by demographers to assess the economic dimensions of this conflict. For example, did the war have similar impacts on the demand for labor as the Black Death epidemic?

Another issue pertains to the cost of Spain's empire and whether the funding model for the state's expansion was the root cause of long-run economic decline. With the increasing scale of armed conflicts in the seventeenth century, the European participants became more and more dependent on access to long-term credit, because whichever government ran out of money first had to surrender. For example, even though the causes of Spain's supposed decline in the seventeenth century are still disputed, nonetheless it can be said that the lack of royal credit and the poor management of government finances resulted in heavy deficit spending as military exertions followed one after another in the seventeenth century. Therefore, the Spanish Crown defaulted repeatedly during the sixteenth and seventeenth centuries and on several occasions forced Spain to seek an end to its military activities (Kamen 2004, 2008). But is this the whole story?

Douglass C. North (1990, 1993) has argued repeatedly that the institutional framework that was adopted early on in the newly unified Spain (the Reconquista was completed in 1492) became a hindrance in the long run, leading Spain to lose its competitive edge vis-à-vis emerging powers like France, the Netherlands, and Great Britain. However, Henry Kamen (2004) has taken a contrary point of view, emphasizing the unlike early success in Spain becoming an empire. Spain developed military expertise through empire building and lost its edge a lot later than North would contend, due to outsourcing its expansion to foreign banks and capital. Kamen did not accept the thesis of long-term decline, at least as outlined by scholars like North. It is hard to ascertain how large a role the recurring conflicts of the time period and the ensuing debt played in the decline. Cliometricians may be able to provide the most compelling arguments in this debate. The study of Philip II of Spain by Mauricio Drelichman and Hans-Joachim Voth (2011, 2014) shows that while the king defaulted several times, he was able to obtain more debt fairly soon after each default. According to their findings, the lenders were often able to thrive through the turbulent Spanish debt relations and that the contract structure in place offered opportunities for risk sharing between the parties. Therefore, Spain's decline may indeed be linked to its institutional framework, but it probably was not an irrational

pattern of development, but rather a mechanism that enabled the sovereigns to expand the empire in many ways for a long time.

Theme 2: Revolutionary and Napoleonic Wars

In the eighteenth century, with rapid population growth in Europe, armies also grew in size. In Western Europe, a mounting intensity of warfare with the 7 Years' War (1756–1763) finally culminated in the French Revolution and Napoleon's conquests (1792–1815). The new style of warfare brought on by the revolutionary wars, with conscription and war of attrition as new elements, can be seen in the growth of army sizes. For example, the French army grew to 650,000 men in 1793, more than 3.5 times its size in 1789. Similarly, the British army grew from 57,000 in 1783 to 255,000 men in 1816. The Russian army was a massive 800,000 men by 1816, a size they maintained throughout the nineteenth century. However, the actual number of Great Power wars declined in absolute numbers, as did the average duration of these wars. It was the *nature* of war that had changed (Eloranta 2005).

A key question for France, for example, was how to finance all these wars. According to Richard Bonney (1999b, c), the cost of France's armed forces in its era of "national greatness" was stupendous, with expenditure on the army by the period 1708–1714 averaging 218 million livres, whereas during the Dutch War of 1672–1678, it had averaged only 99 million in nominal terms. This was due to both the growth in the size of the armed forces and the decline in the purchasing power of the French currency. The overall burden of war, however, remained roughly similar in this period, at least as measured by budgetary share of military expenditures (see Table 1). Furthermore, as for most European monarchies, it was the expenditure on war that brought fiscal change in France, especially after the Napoleonic Wars. Between 1815 and 1913, there was a 444% increase in French public expenditure and a consolidation of the emerging fiscal state. This also embodied a change in the French credit market structure.

The military spending in the late eighteenth century was fairly consistent, representing the largest budgetary item for many European states (Table 1). And, whereas Prussia's defense share was continuously high, the English defense share went up and down, influenced by the various conflicts during this period. In turn, the revolutionary and Napoleonic Wars were truly total wars based on the methods chosen by the belligerents, which affected countries outside the direct fighting. The effects of the war spilled over to influence the relations between neutrals as well. Due to the fact that these wars had an impact on the trade relations of *all* nations, many countries scrambled to find new outlets and sources for their trade. As a result, the bargaining power of weak (United States) and/or smaller states (Portugal – which was both weak and small) increased albeit only temporarily (Moreira and Eloranta 2011).

Scholars have paid too little attention to the smaller players in times of war, often assuming that they occupied an insignificant role in conflicts. Moreira and Eloranta (2011) have argued this is an erroneous assumption, especially in the context of

Table 1 English, French, and Prussian defense shares in the seventeenth and eighteenth centuries

England		France		Prussia	
Year(s)	Defense share	Year(s)	Defense share	Year(s)	Defense share
1690	82	1620–1629	40	–	–
1700	66	1630–1639	35	–	–
1710	88	1640–1649	33	–	–
1720	68	1650–1659	21	1711–1720	78
1730	63	1660–1669	42	1721–1730	75
1741	77	1670–1679	65	1731–1740	82
1752	62	1680–1689	52	1741–1750	88
1760	88	1690–1699	76	1751–1760	90
1770	64	1726	35	1761–1770	91
1780	89	1751	41	1771–1780	91
1790	63	1775	30	1781–1790	78
1800	85	1788	25	1791–1800	82

Sources: calculated from the various sources in European State Finance Database (2013). See also Eloranta and Land (2011) for further details

major conflicts. For example, the United States was a weak state as well and for the most part remained neutral during this period due to its short existence and limited military power. Nonetheless, the United States played an important role in world trade prior to assuming the mantle of a hegemon after the Second World War. Ultimately, it is completely natural to focus most of the analysis on the great empires such as Great Britain and France. After all, it is staggering to conceptualize the evolution of an empire like Great Britain, from its humble beginnings in the sixteenth century, to the building of the navy and its first major victory against the Spanish Armada, to the multicultural, industrialized empire that ruled the world in the nineteenth century. Perhaps it was the intense nature of these rivalries and the total wars between the Great Powers that explains why they had to rely on alliances and often lesser powers to complement their war efforts and retain access to crucial strategic resources. Therefore, even a Great Power like Great Britain had to tolerate the activities of the neutral states, sometimes to the detriment of their own war efforts.

In recent years many scholars have analyzed the disruptions caused by major conflicts like the world wars. Reuven Glick and Alan Taylor (2010) studied the indirect economic effects of wars, in particular the First and the Second World War, with a large database, using an econometric gravity model. Their analysis, focusing on disruptions of trade and the subsequent economic losses, yielded clear results: trade was disrupted by such massive wars and did not return to prewar levels. Furthermore these economic disruptions affected even those countries that were not directly involved in the conflict. These findings may be applicable to other large-scale conflicts as well, even in the preindustrial era.

Furthermore, many scholars have doubted the efficacy of economic warfare, which can range from fairly benign policy measures and pressure to outright warfare

in the context of total war (O'Leary 1985; Førland 1993; Naylor 2001). Lance Davis and Stanley Engerman (2006) have studied one particular form of economic warfare, naval blockades, spanning several centuries. They also emphasize both the costs and challenges of sustaining a successful blockade. For example, during the Napoleonic Wars, the legalities of blockades were not that clearly agreed upon, especially the issue of neutrality. The success of a blockade, as they point out, is often difficult to assess as well (Crouzet 1964). Periods of actual warfare, even blockades, can bring substantial opportunities, as well as disruptions, for trade. As other scholars have argued, rising relative prices and substantial profits may be the answer to why such risky situations bring forth increases in trade (Thornton and Ekelund 2004). Moreover, recent scholarship, as represented by David Bell (2007), for example, certainly puts the revolutionary wars and the ensuing Napoleonic conflicts into the same category as the world wars. Finally, Kevin O'Rourke (2006) has provided cliometric insights into the revolutionary and Napoleonic Wars by focusing on the contraction of trade in particular. His results show that Great Britain was the least affected of the belligerents, whereas France and the United States suffered more. The welfare losses were around 5–6% for the United States, which could be classified as *substantial*.

Theme 3: World Wars

The last four decades leading to the First World War were a period of an intensifying arms race. As argued by Eloranta (2007), the military burdens incurred by the Great Powers also varied in terms of timing, suggesting different reactions to external and internal pressures. Nonetheless, the aggregate, systemic real military spending of the period showed a clear upward trend for the entire period. Moreover, the impact of the Russo–Japanese War was immense for the total (real) spending of the 16 states examined in Eloranta (2007), due to the fact that they were both Great Powers and Russian military expenditures alone were massive. The unexpected defeat of the Russians, along with the arrival of dreadnoughts, launched an even more intensive arms race (Hobson 1993). The basic parameters of the military spending by the most important participants are listed in Table 2.

In August of 1914, this military capacity was unleashed in Europe with horrible consequences, sparking the First World War, which lasted more than 4 years. Another total war had begun, now taking place in the industrial era. About 9 million combatants and 12 million civilians died during the so-called Great War, with property damage concentrated in France, Belgium, and Poland. There have been many new studies that have analyzed the war, in particular *The Economics of World War I*, edited by Stephen Broadberry and Mark Harrison (2005a). What do we know about the economic damages caused by the war? According to Rondo Cameron and Larry Neal (2003), the direct financial losses arising from the Great War were circa 180–230 billion 1914 US dollars, whereas the indirect losses of property and capital rose to over 150 billion dollars. According to Broadberry and Harrison (2005b), the

Table 2 Military burdens (= military spending as a percentage of GDP) and defense shares (= military spending as a percentage of central government expenditures) of sixteen countries, 1870–1913

Country (or group)	Mean military burden	Military burdens, standard deviation	Mean defense shares	Defense shares standard deviation
*AUT**	3.47	0.98	12.03	3.69
BEL	1.88	0.48	14.54	3.67
DEN	1.89	0.50	29.93	7.47
*FRA**	3.68	0.55	25.91	3.69
*GER**	2.56	0.42	54.12	13.45
*ITA**	2.75	0.68	21.69	5.22
*JAP**	4.99	4.63	32.24	17.59
NED	2.77	0.32	26.18	2.65
NOR	5.54c	1.37	85.33	29.48
POR	1.34	0.14	18.95	2.32
*RUS**	3.87	1.63	27.91	5.56
SPA	2.01	0.64	21.35	6.22
SWE	2.13	0.21	35.93	3.99
SWI	1.12	0.32	60.21	5.66
*UK**	2.63	0.97	37.52	7.89
*USA**	0.74	0.25	29.43	10.50
Great Powers (=)*	3.09	1.26	30.11	8.45
Small and medium powers (=the rest)	2.33	0.50	36.55	7.68

Sources: see Eloranta (2007) for details

economic losses arising from the war could be as high as 692 billion 1938 US dollars. But how much of their resources did they have to mobilize and what were the human costs of the war?

The early phase of the First World War involved mobilizing the troops and the immediate economy for warfare. For Germany, the most viable strategy was one which could bring a victory quickly in the situation of geopolitical encirclement by the hostile powers (Ferguson 1999). The possibility that the conflict would be prolonged and Germany and its ally Austria–Hungary would have to fight a war of attrition on two fronts was not taken into serious consideration (Stevenson 2011; Strachan 2011). Great Britain serves as an example of state-directed industrial mobilization. The primary forms of state involvement were co-optation and coordination, not command and compulsion. By bringing in new manufacturers to war production, building new public factories, and buying munitions abroad, the government inadvertently promoted decentralization of the industry and sponsored competition among manufacturers. In a similar fashion, the market-based mechanism of transactions with raw materials had not disappeared; the governmental control in this sphere was accepted as a forced and a temporary measure. During the first 2 years of the war, the government essentially refrained from intervening in

food matters. Unlike all major continental countries, Britain never established a full monopoly on grain. The rationing of some other foods was introduced only 5 months before the end of war. To sum up, the British industry operated as a capitalist market economy, not an administered economy (Blum and Eloranta 2013).

Many of the features of a market economy fully applied to the wartime French economy. More than any other economic system, the French economy was an example of self-mobilization and self-organization of production. Although industrial mobilization was sponsored by the government, industrialists themselves came to play a primary role in organizing mass production of arms and munitions. French resource-allocation institutions, the consortia, represented a weak and temporary version of corporate organization. For about 3 years during the war, the French authorities limited themselves to very few measures of food regulation. What little regulation they imposed was mostly in the form of price controls on some foods. Because of large imports of food from overseas, more dramatic measures of food control did not seem necessary. The government introduced a grain monopoly and the rationing of bread late in 1917 (Blum and Eloranta 2013).

Allied Great Powers were also able to mobilize their resources more effectively during the war. Even though the Central Powers initially did quite well with the limited resources they had, the Allies were able to mobilize their far superior economic and military resources better both at the home front and on the front lines. Their more democratic institutions supported the demands of the total war effort better than their authoritarian counterparts, especially in terms of being able to mobilize their societies farther than their competitors. Therefore, the richer countries mobilized more men and materiel for the war, and their war industries proved quite capable of adapting to fulfill the needs of the war machine (Broadberry and Harrison 2005b).

Moreover, having a large peasant population turned out to be a hindrance for the production of food under wartime constraints. In poorer countries, and even in affluent Germany, mobilization efforts siphoned off resources from agriculture, and the farmers preferred to hoard food rather than sell it for low prices. As Avner Offer (1989) has argued, food (or the lack thereof) played a crucial part in Germany's collapse. Germany's problem was not so much that it was not able to mobilize resources for the war, but the fact that its main ally, Austria–Hungary, was a poor nation with limited resources and plagued by an inability to mobilize effectively. The collective mobilization of resources by the Allies proved too big an obstacle for Germany to overcome (Blum 2011, 2013; Blum and Eloranta 2013)

Ultimately, resources and mobilization were weaknesses of the Central Powers, and led to their defeat, after the overextension of their supply lines in the spring of 1918. After the ceasefire in 1918 and the Versailles peace treaty of 1919, Germany experienced political instability and economic turmoil. A complex process of interactions between a military defeat, an exhausted economy, the burden of reparations and war debt, and an uncertain political future laid the basis for three more turbulent decades. One consequence of the war and the war economy, which is believed to have contributed to the destruction of the Weimar Republic, was the hyperinflation experienced between 1921 and 1923. By 1923, the average price indices for food

prices and retailer prices were 198 billion and 166 billion, respectively, while wages rose disproportionally. Wage indices (1913 = 1 for all of them) for skilled workers in 1923 were 85 billion, while corresponding figures for unskilled and civil servants were 100 and 56, respectively (Blum 2013; Blum and Eloranta 2013). Despite the fact that hyperinflation was a stimulus for economic impulses, especially for the exporting sectors, and helped delete considerable amounts of internal and non-reparation foreign debt, the inflation's aggregate effect was disastrous; German GDP dropped by one third from 1922 to 1923.

Moreover, Albrecht Ritschl (2005) has argued that conditions for peace were simultaneously too lax and too strict, and the way Germany, especially the German public, perceived the end of this conflict was not realistic. He maintains that Germany's military was technically defeated, but Germany had not yet been invaded by Allied troops. German leaders had not been captured and the capital had not been besieged or conquered; in fact, German territory remained by and large unscathed and the Emperor managed to escape to the neutral Netherlands. Soon legends spread, saying that the lack of morale at the home front was sabotaging military strength – the birth of the *stab-in-the-back* legend. It was difficult to realize the inevitable defeat in the absence of obvious evidence; it took another coalition of Allied forces to complete this task in 1945. Illusions about the "true" reason for the lost war and financial disasters served right-wing propagandists to stir up hatred against ethnic and political minorities.

On the macroeconomic consequences, Feinstein et al. (1997) have identified four direct economic outcomes that arose from the conflict: (1) two immediate exogenous shocks, in terms of disruptions of supply and demand as well as excess mobilized production (and military) capacity after the conflict; (2) a more rigid economic environment, due in part to diminished wage flexibility; (3) a weaker financial structure, since the economies had to carry the new, increased levels of public spending as well as the acquired debts with mainly prewar levels of taxation; and (4) a fragile international monetary system. The "winners" of the war, at least in terms of economic growth effects, seemed to be the neutral states, such as the Nordic countries, who outperformed other Western states.

Ronald Findlay and Kevin O'Rourke (2007) have summarized the challenges brought on by the period of world wars in their aftermath for global trade and politics. First of all, there were three wartime adjustments that had serious consequences in the interwar period: (1) non-European producers who increased their role during the war and the subsequent price competition; (2) industrial expansion during wartime, which was difficult to redirect after the conflict; and (3) the boost in non-European industrialization. In terms of broader effects, they list the following: (1) the increased importance of turbulent domestic politics, (2) the legacy of the war debts, (3) the difficulties in returning to the gold standard, (4) the creation of new nation states, (5) the Communist revolution and state in Russia, and (6) the creation of instability and conditions as a result of the First World War that would lead to the Second World War.

There have been many studies of the interwar period, especially the disarmament and rearmament processes that took place. For example, studies on German

rearmament have typically focused on the war period. However, Max Hantke and Mark Spoerer (2010) have studied the hidden military spending of the 1920s, utilizing classic cliometric techniques such as counterfactual analysis. Their basic aim is to analyze the economic importance of the Versailles Treaty, i.e., reparations, on the German economy. In order to achieve this, they have devised a clever strategy. Their counterfactual framework attempts to see what the impact of "normal," unrestricted military spending would have been on the German economy without the treaty restrictions on their armed forces. In doing so, they conclude that the size of that spending, as well as its impact, would have been roughly equal and that the failure of the Weimar economy was due to domestic policies. At the heart of this failure were fiscal effects and constraints imposed by the loss of territory and industrial capacity. In terms of some of the techniques involved, Hantke and Spoerer showed considerable ingenuity in reconstructing the German budgetary figures for the turbulent 1920s. The discussion of reparations and their impact is also reminiscent of Eugene White's (2001) work on the Franco–Prussian war of 1870–1871 and its aftermath.

Moreover, one could ask why the League of Nations ultimately failed to achieve widespread disarmament, its most fundamental goal. Eloranta (2011) has shown that the failure of the League of Nations had two important aspects: the failure to provide adequate security guarantees for its members (like a credible alliance, e.g., the NATO) and the failure of this organization to achieve the disarmament goals it set out in the 1920s and 1930s. Thus it was doomed from the outset to fail, due to built-in institutional contradictions. In terms of economic theory and econometric analysis, the League of Nations can also be modeled and analyzed as a military alliance. Based on Eloranta's (2011) analysis, the results are fairly conclusive: the League of Nations did not function as a pure public good alliance, which encouraged an arms race in the 1930s.

In the interwar period, many countries maintained fairly high military spending levels, despite tendencies to disarm, particularly in the 1920s. The mid-1930s marked the beginning of intense rearmament, whereas some of the authoritarian regimes had begun earlier in the decade. Germany, under Hitler, increased its military burden from 1.6% in 1933 to 18.9% in 1938, a rearmament program combining creative financing and promising both guns and butter for the Germans. Mussolini was not quite as successful in his efforts to realize the new Roman Empire, with a military burden fluctuating between 4% and 5% in the 1930s (5.0% in 1938). The Japanese rearmament drive was perhaps the most impressive, with a military burden as high as 22.7% and greater than 50% defense share in 1938. For many countries, such as France and Russia, the rapid pace of technological change in the 1930s rendered many of the earlier armaments obsolete only 2 or 3 years later (Eloranta 2002).

Mark Thomas (1983) has analyzed the British rearmament using an adjusted version of input–output tables for the 1930s, which enabled him to examine the causal interdependencies in the economy. He argued that the rearmament helped alleviate the effects of the recession, essentially making a Keynesian argument about the government's fiscal policy. Crafts and Mills (2013) have challenged this recently,

making a case for a lower multiplier effect, in the range of 0.3–0.8. They also discuss at length the previous models of estimating the impact and note that the theoretical findings are, by and large, model dependent. Their choice is to use the more recent time series techniques, which take into account potential problems of endogeneity. This debate is far from resolved, which also applies to the debate over broader 1930s programs like the New Deal.[3]

In the Second World War, the initial phase from 1939 to early 1942 favored the Axis as far as strategic and economic potential was concerned. After that, the war of attrition, with the United States and the USSR joining the Allies, turned the tide in favor of the Allies. For example, in 1943 the Allied total GDP was 2,223 billion international dollars (in 1990 prices), whereas the Axis accounted for only 895 billion. Also, the impact of the Second World War was much more profound for the economies of the participants. For example, Great Britain at the height of the First World War incurred a military burden of circa 27%, whereas the military burden level consistently held throughout the Second World War was over 50% (Harrison 1998; Eloranta 2003).

Other Great Powers experienced similar levels. Only the massive economic resources of the United States made possible its lower military burden. The United Kingdom and the United States also mobilized their central/federal government expenditures efficiently for the military effort. In this sense the Soviet Union fared the worst, and additionally the share of military personnel out of the population was relatively small compared to the other Great Powers. On the other hand, the economic and demographic resources that the Soviet Union possessed ultimately ensured its survival during the German onslaught. On the aggregate, the largest personnel losses were incurred by Germany and the Soviet Union. They were many times those of the other Great Powers. In comparison with the First World War, the second one was even more destructive and lethal, and the aggregate economic losses from the war exceeded 4,000 billion 1938 US dollars. After the war, the European industrial and agricultural production amounted to only half of the 1938 total (Harrison 1996, 1998, 2000).

The Second World War has been extensively covered in recent economic history scholarship. *The Economics of World War II*, edited by Mark Harrison, is a good collection of such efforts. For example, although Britain and Germany had significantly different buildups to war, the war itself brought extreme economic activity where both nations had virtually zero unemployment and economies geared toward the production of armaments. From 1933 to 1944, Germany's GDP and GDP per capita rose every year. Naturally, the end of the war brought economic collapse to Germany, with a serious drop-off in both 1945 and 1946 (33% and 50%, respectively). Similarly, the United Kingdom experienced significant growth in its GDP and GDP per capita. By 1943, however, the United Kingdom's economic growth began to slow down and experienced a drop in both 1944 and 1945 GDP, though not

[3]See Fishback and Kachanovskaya (2010) as an example of this debate.

as devastating as the German economy's deterioration (Broadberry and Howlett 1998; Abelshauser 2000).

Moreover, Germany's total munitions output increased monthly for every year of the war, until September 1944, when production finally began to wane. However, German productivity in machine tools decreased throughout the war, falling to 79% of 1939 levels. Even the production of coal suffered a drop in productivity, falling to 75 t per shift per worker from the 1939 peak of 100 t. Some of this productivity loss was due to labor shortages or disputes. Though coal production in the United Kingdom fell, the agricultural sector experienced a seemingly miraculous increase in efficiency. Though losing a good portion of its experienced workers to the armed services and decreases in food imports of 70% or more, agricultural production was able to increase the calories per employee by nearly 77% from 1937 levels. Employment in the agriculture sector increased even with the loss of its experienced workers by using women, volunteer labor, and eventually prisoners of war. Ultimately, the success or failure of a sector varied wildly from industry to industry (Broadberry and Howlett 1998; Abelshauser 2000).

Finally, the war had immediate and long-term impacts on the economic development of these economies. In Germany, rationing began even before the war began. Furthermore, the cost of living steadily increased for the average family, particularly for food. By 1944, food prices had increased to 113% of 1938 prices, and clothing had risen to 141% of 1938 prices. As a result of these higher prices and rationing, the average caloric intake for a worker's family member fell from 2,435 cal in 1939–1940 to 1,412 by 1945–1946 (Abelshouser 2000). Of course, the dismantling of the German state and economy following the war left little doubt of the destructive power that the Second World War had on the economies of Europe. Even victors, such as the United Kingdom, experienced major losses. Based on physical capital, the British lost 18.6% of its prewar wealth (Broadberry and Howlett 1998). This left many concerned that the state planned to continue nationalizing industries (though that fear never materialized). In the end, every nation in Europe felt the economic pain of war.

Theme 4: Cold War and Beyond

The end of the Second World War brought with it a new global order, with the United States and the Soviet Union as the most dominant players in global security affairs. With the establishment of NATO in 1949, a formidable defense alliance was formed by the Western countries. The USSR, rising to new prominence due to the war, in turn established the Warsaw Pact in 1955. The war also meant a change in the public spending and taxation levels of most Western nations. Military spending levels followed increases in welfare and social spending and peaked during the early Cold War. The American military burden increased above 10% in 1952–1954, and the United States retained a high mean value for the postwar period of 6.7% (up until 1991). Great Britain and France followed the American example after the Korean War (Eloranta 2003).

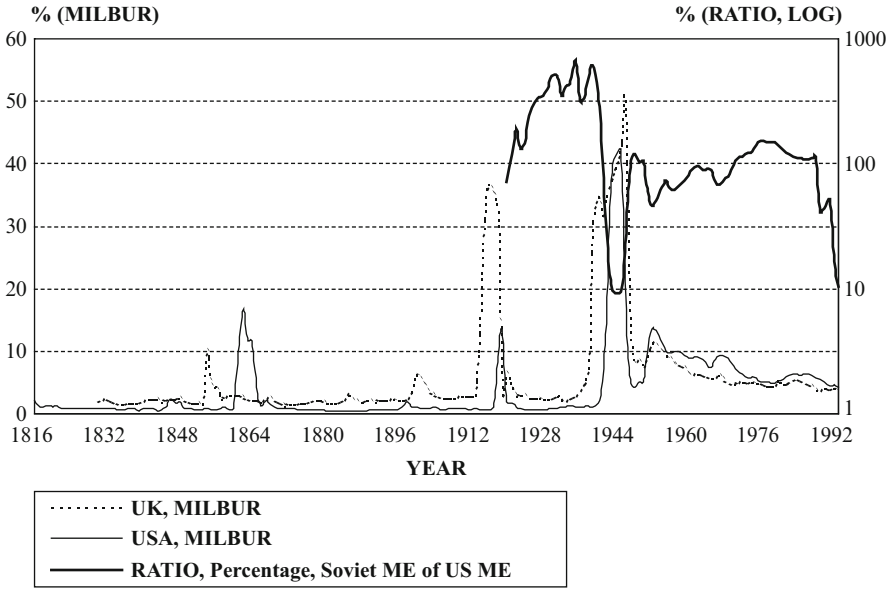


Fig. 1 Military burdens (= MILBUR) of the United States and the United Kingdom and the Soviet military spending as a percentage of the US military spending (*ME*), 1816–1993 (Sources: see Eloranta (2005) for details on the sources and methods of calculation)

The Cold War was tied to an extensive arms race, with nuclear weapons as the main investment item, with the USSR spending circa 60–70% of the American level in the 1950s, and the USSR briefly spending more than the United States in the 1970s (see Fig. 1). Nonetheless, the United States maintained a massive advantage over the Soviets in terms of nuclear warheads. Yet, the comparative spending figures suggest that NATO had a 2-to-1 lead in spending vis-à-vis the Warsaw Pact in the 1970s and early 1980s. Some of this armament race was due to technological advances that led to increases in the cost per soldier – it has been estimated that technological increases have produced a mean annual increase in real costs of circa 5.5% in the postwar period. However, the "bang for the buck" increased drastically with the introduction of nuclear weapons and other new weapons systems (Ferguson 2001; Eloranta 2005).

While the Second World War has been studied at length, the Cold War conflicts and military spending have not. However, some themes have been explored, for example, the so-called Military Industrial Complex (MIC), which refers to the influence that the military and industry would have on each other's policies. The more negative connotation refers to the unduly large influence that military producers might have over the public sector's acquisitions and foreign policy in particular, in such a collusive relationship. In fact, the origins of this type of interaction can be found further back in history. As Paul Koistinen (1980) has emphasized, the First World War was a watershed in business–government relationships, since businessmen were often brought into government to make supply

decisions during this total conflict. Most governments, as a matter of fact, needed the expertise of the core business elites during the world wars. In the United States some form of MIC came into existence before 1940. Similar developments can be seen in other countries before the Second World War, for example, in the Soviet Union. The Cold War simply reinforced these tendencies (Koistinen 1980; Harrison 2003; Eloranta 2009). Findings by example Robert Higgs (Trevino and Higgs 1992; Higgs 1994) establish that the financial performance of the leading defense contracting companies was, on the average, much better than that of comparable large corporations during the period 1948–1989. Nonetheless, his findings do not support the normative conclusion that the profits of defense contractors were abnormal.

The Cold War arms race has not been covered much by economic historians – most of the studies have been conducted by defense economists. Beginning with Mancur Olson and Richard Zeckhauser's path-breaking work on NATO (Olson and Zeckhauser 1966), there have been many testable hypotheses relating to the idea of collective security provision in an alliance and the implications of this provision on military spending. As the logic suggested by Olson and Zeckhauser implies, pure public good alliances are characterized by free riding by the smaller (or poorer) states since they have a reasonable expectation of military assistance from the larger nations under a cohesive defensive arrangement. For example, recent studies have found that the pure public good alliance concept describes NATO until about 1966, when a change in the strategic doctrine forced the members to rely more on their own military provision (Sandler and Hartley 1999).

Of course, as I would argue, the explanation resting on the foundation of the public good theory and the suboptimality of defense provision via the spillover effect has sound theoretical foundations. Indeed, Olson and Zeckhauser (1966) found a significant positive correlation, using Spearman rank correlation tests, between the NATO Allies' GNP and their military burdens in 1964, indicating clear free-riding behavior by the smaller Allies. Later studies specified the pure public good alliance to describe NATO until 1966, when the positive rank correlation between the variables ceased to be statistically significant (Sandler and Murdoch 1990; Sandler and Hartley 1999).

Another interesting field of debate is how the Cold War ended. It is obviously a topic that is fraught with political importance as well as scholarly debate. The list of possible explanatory factors is long indeed (see, e.g., Kegley 1994). Moreover, there is a big debate over the collapse of the Soviet Union. Economic historians have become involved in this debate too. For example, Mark Harrison (2002) has analyzed this issue by developing a simple model of costs and benefits to a dictator and producer of the command system. According to his findings, the command system was not necessarily inherently unstable; rather the stability was conditional, based on the equilibrium surrounding the level and application of coercion. In the case of the Soviet Union, the lessening of the threat of coercion that came with perestroika policies unraveled the system. Harrison used historical statistical and archival data on military spending to prove his assertions.

Nonetheless, despite these analyses, economic historians have not yet been very active in the debate over the scale and scope of the Cold War and the impacts of the

arms race. Economic historians have been more interested in the so-called Golden Age of Economic Growth (1950–1973), the evolution of the European Union, and the Marshall Plan (Maddison 1989, 2001; Berger and Ritschl 1995; Eichengreen 1995; Ritschl 2004). Most of this domain has been occupied by defense economists and political scientists. The opening of various archives and new data sources will certainly change this state of affairs in the near future.

Theme 5: Long-Run Analyses (Military Spending, Societal Structures, and Empires)

Economic historians have typically been more interested than economists in general in the long-run development of societies. Political and conflict scientists, as well as sociologists, have been interested in the same issues and even time periods, but not always from the same angle or using the same methods. However, even though some cycle theorists and conflict scientists have been interested in the formation of modern nation states and the respective system of states since 1648, they have not expressed any real interest in premodern societies and warfare (Wright 1942; Blainey 1973; Levy 1985, 1998; Geller and Singer 1998). Economic historians have continuously extended their reach back to earlier periods, especially since new data is now available to analyze, for example, state formation.

Political scientists are also in search of patterns of development, such as waves of development. According to George Modelski and William R. Thompson (1988, 1996), for proponents of Kondratieff waves and long cycles as explanatory forces in the development of world leadership patterns, the key aspect in a state's ascendancy to prominence is naval power. One of the less explored aspects in most studies of hegemonic patterns is the military expenditure component in the competition between the states for military and economic leadership in the system. It is often argued, for example, that uneven economic growth levels cause nations to compete for economic and military prowess. The leader nation thus has to dedicate increasing resources to armaments in order to maintain its position, while the other states, the so-called followers, can benefit from greater investments in other areas of economic activity. Therefore, the follower states act as free riders in the international system stabilized by the hegemon. A built-in assumption in this hypothesized development pattern is that military spending eventually becomes harmful for economic development, a notion that has often been challenged based on empirical studies (Kennedy 1989; Eloranta 2005).

There have been relatively few credible attempts to model the military (or budgetary) spending behavior of states based on their long-run regime characteristics. Here I will elaborate on three in particular: the Webber–Wildavsky (1986) model of budgeting, the Richard Bonney (1999a) model of fiscal systems, and the Niall Ferguson (2001) model of interaction between public debts and forms of government. Carolyn Webber and Aaron Wildavsky maintain that each political culture generates its characteristic budgetary objectives, namely, productivity in

market regimes, redistribution in sects (specific groups dissenting from an established authority), and more complex procedures in hierarchical regimes.

Their model, however, is essentially a static one. It does not provide clues as to why the behavior of nations may change over time, especially over long time periods. Richard Bonney (1999a) has addressed this problem in his writings on the early modern states. He has emphasized that the states' revenue and tax collection systems, the backbone of any militarily successful nation state, have evolved over time. For example, in most European states, the government became the arbiter of disputes and the defender of certain basic rights in the society by the early modern period. During the Middle Ages, the European fiscal systems were relatively backward and autarchic, with mostly predatory rulers (or roving bandits, as Mancur Olson (1993) has coined them). In his model this would be the stage of the so-called tribute state. Next in the evolution came, respectively, the domain state (with stationary bandits, providing some public goods), the tax state (more reliance on credit and revenue collection), and finally the fiscal state (embodying more complex fiscal and political structures). A superpower like Great Britain in the nineteenth century had to be a fiscal state to be able to dominate the world, due to all the burdens that went with an empire (Ferguson 2003).

While both of the models mentioned above have provided important clues as to how and why nations have prepared fiscally for wars, or survived them, the most complete account of this process has been provided by Niall Ferguson (2001, 2006). He has maintained that wars have shaped all the most relevant institutions of modern economic life: tax-collecting bureaucracies, central banks, bond markets, and stock exchanges. Moreover, he argues that the invention of public debt instruments has gone hand in hand with more democratic forms of government and military supremacy – hence, the so-called Dutch or British model. These types of regimes have also been the most efficient economically, which has in turn reinforced the success of this fiscal regime model. In fact, military expenditures may have been the principal cause of fiscal innovation for most of history. Ferguson's model highlights the importance, for a state's survival among its challengers, of the adoption of the right types of institutions, technology, and a sufficient helping of external ambitions. Typically, however, none of these models utilize extensive quantitative testing or methods in order to prove their assertions, which will make this field of inquiry a fruitful one in the future.

Cliometricians have already brought a lot to the table as well, especially concerning the long-run development of states, regime type, and financial/fiscal evolution. For example, Philip Hoffman (Hoffman and Rosenthal 1997; Hoffman 2011) has shown that it is possible to analyze the military sector and technology over several centuries, in fact before the industrial revolutions. He discovered, mostly based on the analysis of price data, that Western Europe developed a comparative advantage in violence long before 1800. European military industries exhibited tremendous productivity growth in the early modern period, which gave them the edge, particularly in gunpowder technologies. Hoffman (2012) has also introduced an interesting (and testable) model to explain where this comparative advantage came from, namely, the tournament model. In this model winning wars and

technological development to gain the edge with a country's competitors were intricately intertwined. And this led to giving the Europeans an edge in their military development and contributed to the building of empires.

Economic historians have also engaged heavily in the debate over the expansion, functioning, and profitability of empires, for example, the British Empire. Such critiques, of course, go back centuries, to such pivotal figures as Adam Smith (1776) or John Hobson (1965 reprint), who were very skeptical about the profitability of the British Empire and who ultimately benefited from the empire. More recently, several scholars, including Avner Offer (1993), Patrick O'Brien (1988), Lance Davis and Robert Huttenback (1982, 1986), and Niall Ferguson (2003, 2004), have weighed in on this question. The various tools of economic historians have been employed to tackle these questions. Davis and Huttenback maintained, on the basis of econometric exercises and modeling, that the profits from the empire were not sufficient to contribute to the overall economic growth of Britain and that the middle class incurred a larger share of the relative costs than the elites. In a related fashion, O'Brien argued, also along the lines of Smith and Hobson, that the empire was an unnecessary expenditure that dragged Britain to a multitude of conflicts, at high cost, and that the costs were borne unequally, mainly by mainland Britons. On the other hand, Offer has criticized some of these findings, especially the data solutions and the role that military expenditures have played in the building of the empire. He used alliance theories to explain the unequal allocation of expenditures within the empire. Ferguson, in a different fashion, has also attempted to highlight some of the benefits of the empire, for example, the ability to trade within a large area.

Another set of issues to which cliometricians have contributed heavily is the formation of states in the long run. Along with the scholars that I have already discussed, Mark Dincecco (2009, 2010; Dincecco and Prado 2012) has made some significant contributions to the debate over the fiscal revolution of European states. Typically employing newer panel data and related techniques, he found that centralization and limitations on political power led to higher revenues among European states, that premodern war casualties were correlated with fiscal institutions, that improvements in fiscal capacity possibly led to higher economic returns, and that Eastern and Western Europe, measured by the river Rhine, diverged institutionally after 1789. Moreover, in a similar fashion, David Stasavage et al., prominent political scientists, have discovered that between 1600 and 2000 the biggest factors in the rise and fall of the mass army were changes in transportation and communications technology and that mobilization for the needs of total warfare in the twentieth century led to the emergence of substantial redistribution of wealth and progressive taxation (Onorato et al. 2012; Scheve and Stasavage 2010).

In turn, Mark Harrison and Nikolaus Wolf (2011) have examined a different aspect of the development of states, namely, the development of democracies and warfare. The topic of so-called democratic peace, implying that democracies do not fight one another, is vast and very interdisciplinary (see, e.g., Russett 1993; De Mesquita et al. 1999; Choi 2011; Gowa 2011; Dafoe and Russett 2013). Harrison and Wolf (2014) argue that this widely accepted thesis may not always hold,

especially when analyzing the development of states from 1870 onward. They claim that trade and democracy do not always work to prevent conflicts, but can in fact lead to increasing the capacity for war and the frequency of conflicts. This entry has inspired some debate over the legitimacy of the democratic peace concept (Gleditsch and Pickering 2014; Harrison and Wolf 2014).

Conclusion

This chapter is meant as a review of the many of the perspectives from history, political sciences, sociology, and economics that economic historians have applied to the study of war, especially how theoretical and quantitative perspectives achieved in this fashion have enriched the debate over the causes, buildup, costs, and outcomes of conflicts. I would argue that economic historians, using a variety of techniques ranging from simple data tools to more complicated econometric exercises, have often broadened the scope of the debates, enabling both more comprehensive long-run analysis of conflicts and polities as well as deeper economic analysis of particular conflicts and periods.

Here I first reviewed some of the scholarship on medieval and the early modern periods, especially the formation of European nation states and the role conflicts played in these processes. I also returned to these themes toward the end of the review when looking at long-run processes. In sum, it is now fairly clear that Europeans emerged out of this period with a comparative advantage in violence, which they achieved through technological innovations and repeated warfare. Fiscal innovation and expansion was a key part of this. Then, I moved to the discussion of a pivotal period in history, namely, the period of the revolutions and Napoleonic conflicts. This period represented a change in the nature of warfare, the arrival of total war tactics and strategies in earnest, the emergence of new kinds of states, and the advent of the industrial age. The nineteenth century represented a period of globalization and relatively few conflicts, but it also set the stage for the destructive twentieth century.

The period of the world wars is perhaps the best represented area of study for economic historians of late. Many scholars have now delved into the economic dimensions and impacts of the conflicts, as well as the disarmament/rearmament of the interwar period. New data and scholarship has shown the mechanics of mobilization and highlighted the importance of resources in deciding these conflicts. Conversely, the Cold War period has been relatively sparsely studied, at least from the perspective of conflicts or military spending. Given the availability of new data and the opening of many archives, it is highly likely that this state of affairs will change in the near future. Economic historians have clearly made an impact in the study of long-run phenomena like state formation, empires, and democracy. Comparative studies are at the heart of these new scholarly efforts, and cliometrics is well suited to the study of such topics, especially given the new panel and time series techniques available, the rapid development of computing power, and the creation of many new online databases.

References

Abelshauser W (2000) Germany: guns, butter, and economic miracles. In: Harrison M (ed) The economics of World War II. Six great powers in international comparison. Cambridge University Press, Cambridge, pp 122–176

Bell D (2007) The first total war: Napoleon's Europe and the birth of warfare as we know it. Houghton Mifflin Harcourt, New York

Berger H, Ritschl A (1995) Germany and the political economy of the Marshall Plan, 1947–1952: a re-revisionist view. In: Eichengreen B (ed) Europe's postwar recovery. Cambridge University Press, Cambridge, pp 199–245

Blainey G (1973) The causes of war. Free Press, New York

Blaydes L, Chaney E (2013) The Feudal Revolution and Europe's rise: political divergence of the Christian West and the Muslim World before 1500 CE. Am Polit Sci Rev 107(01):16–34

Blum M (2011) Government decisions before and during the First World War and the living standards in Germany during a drastic natural experiment. Explor Econ Hist 48(4):556–567

Blum M (2013) War, food rationing, and socioeconomic inequality in Germany during the First World War. Econ Hist Rev 66(4):1063–1083

Blum M, Eloranta J (2013) War zones, economic challenges, and well-being – perspectives on Germany during the First World War. In: Miller N (ed) War: global assessment, public attitudes and psychological effect. Nova Publishers, New York

Bonney R (ed) (1999a) The rise of the Fiscal State in Europe c. 1200–1815. Oxford University Press, Oxford

Bonney R (1999b) Introduction. In: Bonney R (ed) The rise of the Fiscal State in Europe c. 1200–1815. Oxford University Press, Oxford, pp 1–17

Bonney R (1999c) France, 1494–1815. In: Bonney R (ed) The rise of the Fiscal State in Europe c. 1200–1815. Oxford University Press, Oxford

Brauer J, Tuyll HV (2008) Castles, battles & bombs. How economics explains military history. Chicago University Press, Chicago

Broadberry S, Harrison M (eds) (2005a) The economics of World War I. Cambridge University Press, Cambridge, UK

Broadberry S, Harrison M (2005b) The economics of World War I: an overview. In: Broadberry S, Harrison M (eds) The economics of World War I. The Cambridge University Press, Cambridge, UK

Broadberry S, Howlett P (1998) The United Kingdom: 'Victory at all costs'. In: Harrison M (ed) The economics of World War II. Six great powers in international comparisons. Cambridge University Press, Cambridge, UK

Cameron R, Neal L (2003) A concise economic history of the World. From paleolithic times to the present. Oxford University Press, Oxford

Choi SW (2011) Re-evaluating capitalist and democratic peace models 1. Int Stud Q 55(3):759–769

Crafts N, Mills TC (2013) Rearmament to the rescue? New estimates of the impact of "Keynesian" policies in 1930s' Britain. J Econ Hist 73(04):1077–1104

Crouzet F (1964) Wars, blockade, and economic change in Europe, 1792–1815. J Econ Hist 24 (4):567–588

Dafoe A, Russett B (2013) Does capitalism account for the democratic peace? The evidence still says no. In: Schneider G, Gleditsch NP (eds) Assessing the capitalist peace. Routledge, New York, pp 110–126

Davis LE, Huttenback RA (1982) The political economy of British imperialism: measures of benefits and support. J Econ Hist 42(01):119–130

Davis LE, Huttenback RA (1986) Mammon and the pursuit of Empire: the political economy of British Imperialism, 1860–1912. Cambridge University Press, Cambridge

Davis LE, Engerman SL (2006) Naval blockades in peace and war: an economic history since 1750. Cambridge University Press, Cambridge/New York

De Mesquita BB, Morrow JD, Siverson RM, Smith A (1999) An institutional explanation of the democratic peace. Am Polit Sci Rev 93(4):791–807

Dincecco M (2009) Fiscal centralization, limited government, and public revenues in Europe, 1650–1913. J Econ Hist 69(01):48–103
Dincecco M (2010) Fragmented authority from Ancien Régime to modernity: a quantitative analysis. J Inst Econ 6(3):305
Dincecco M, Prado M (2012) Warfare, fiscal capacity, and performance. J Econ Growth 17 (3):171–203
Drelichman M, Voth HJ (2011) Lending to the borrower from hell: debt and default in the age of Philip II*. Econ J 121(557):1205–1227
Drelichman M, Voth H-J (2014) Lending to the borrower from hell: debt, taxes, and default in the age of Philip II. Princeton University Press, Princeton
Eichengreen B (1995) Europe's postwar recovery. Cambridge University Press, Cambridge
Eloranta J (2002) External security by domestic choices: military spending as an impure public good among eleven European states, 1920–1938. Dissertation, European University Institute
Eloranta J (2003) National defense. In: Mokyr J (ed) The Oxford encyclopedia of economic history. The Oxford University Press, Oxford, pp 30–33
Eloranta J (2005) Military spending patterns in history. EH.Net Encyclopedia. Accessed 1 Mar 2008 http://eh.net/encyclopedia/article/eloranta.military
Eloranta J (2007) From the great illusion to the Great War: military spending behaviour of the Great Powers, 1870–1913. Eur Rev Econ Hist 11(2):255–283
Eloranta J (2009) Rent seeking and collusion in the military allocation decisions of Finland, Sweden, and Great Britain, 1920–381. Econ Hist Rev 62(1):23–44
Eloranta J (2011) Why did the League of Nations fail? Cliometrica 5(1):27–52
Eloranta J, Land J (2011) Hollow victory? Britain's public debt and the seven years' war. Essays Econ Business Hist 29:101–118
European State Finance Database (2013) Online database, managed by Richard Bonney. http://www.esfdb.org/Default.aspx. Accessed 1 Mar 2013
Feinstein CH, Temin P, Toniolo G (1997) The European economy between the wars. Oxford University Press, Oxford/New York
Ferguson N (1999) The Pity of war. Explaining World War I. Basic Books, New York
Ferguson N (2001) The cash nexus: money and power in the modern world, 1700–2000. Basic Books, New York
Ferguson N (2003) Empire: the rise and demise of the British world order and the lessons for global power. Basic Books, New York
Ferguson N (2004) Colossus: the price of America's empire. Penguin Press, New York
Ferguson N (2006) The war of the world: twentieth-century conflict and the descent of the West. Allen Lane, London
Findlay R, O'Rourke K (2007) Power and plenty: trade, war, and the world economy in the second millennium. Princeton University Press, Princeton
Fishback PV, Kachanovskaya V (2010) In search of the multiplier for federal spending in the states during the new deal. National Bureau of Economic Research, Cambridge, MA
Førland TE (1993) The history of economic warfare: international law, effectiveness, strategies. J Peace Res 30:151–162
France J (2001) Recent writing on medieval warfare: from the Fall of Rome to c. 1300. J Mil Hist 65 (2):441–473
Geller DS, Singer JD (1998) Nations at war: a scientific study of international conflict. Cambridge University Press, Cambridge/New York
Gleditsch KS, Pickering S (2014) Wars are becoming less frequent: a response to Harrison and Wolf. Econ Hist Rev 67(1):214–230
Glick R, Taylor AM (2010) Collateral damage: trade disruption and the economic impact of war. Rev Econ Stat 92(1):102–127
Gowa J (2011) The democratic peace after the cold war. Econ Polit 23(2):153–171
Hantke M, Spoerer M (2010) The imposed gift of Versailles: the fiscal effects of restricting the size of Germany's armed forces, 1924–9. Econ Hist Rev 63(4):849–864

Harrison M (1996) Accounting for war: soviet production, employment, and the defence burden, 1940–1945. Cambridge University Press, Cambridge
Harrison M (1998) The economics of World War II: an overview. In: Harrison M (ed) The economics of World War II. Six great powers in international comparisons. Cambridge University Press, Cambridge, UK
Harrison M (2000) The Soviet Union: the defeated victor. In: Harrison M (ed) The economics of World War II. Six great powers in international comparison. Cambridge University Press, Cambridge, pp 268–301
Harrison M (2002) Coercion, compliance, and the collapse of the Soviet command economy. Econ Hist Rev 55(3):397–433
Harrison M (2003) Soviet industry and the red army under Stalin: a military-industrial complex? Les Cahiers du Monde russe 44(2–3):323–342
Harrison M, Wolf N (2011) The frequency of wars. Econ Hist Rev 65(3):1055–1076
Harrison M, Wolf N (2014) The frequency of wars: reply to Gleditsch and Pickering. Econ Hist Rev 67(1):231–239
Higgs R (1994) The cold war economy. Opportunity costs, ideology, and the politics of crisis. Explor Econ Hist 31(3):283–312
Hobson JA (1965 (reprint)) Imperialism. University of Michigan Press, Ann Arbor
Hobson JM (1993) The military-extraction gap and the wary titan: the fiscal sociology of British defence policy 1870–1914. J Eur Econ Hist 22(3):466–507
Hoffman P, Rosenthal JL (1997) The political economy of warfare and taxation in early modern Europe: historical lessons for economic development. In: Drobak J, Nye JV (eds) The frontiers of the new institutional economics. Academic Press, San Diego, pp 31–55
Hoffman PT, Jacks DS, Levin PA, Lindert PH (2002) Real inequality in Europe since 1500. J Econ Hist 62(02):322–355
Hoffman PT (2011) Prices, the military revolution, and western Europe's comparative advantage in violence. Econ Hist Rev 64(s1):39–59
Hoffman PT (2012) Why was it Europeans who conquered the world? J Econ Hist 72(03):601–633
Kamen H (1968) The economic and social consequences of the thirty years' war. Past Present 39 (1):44–61
Kamen H (2004) Empire: how Spain became a world power, 1492–1763. HarperCollins, New York
Kamen H (2008) Imagining Spain: historical myth & national identity. Yale University Press, New Haven/London
Kegley CW Jr (1994) How did the cold war die? Principles for an autopsy. Mershon Int Stud Rev 38:11–41
Kennedy P (1989) The rise and fall of the great powers. Economic change and military conflict from 1500 to 2000. Fontana, London
Koistinen PAC (1980) The military-industrial complex. A historical perspective. Foreword by Congressman Les Aspin. Praeger Publishers, New York
Levy JS (1985) Theories of general war. World Polit 37(3):344–374
Levy JS (1998) The causes of war and the conditions of peace. Ann Rev Polit Sci 1(1):139
Maddison A (1989) The world economy in the 20th century. OECD Publications and Information Center Distributor, Paris
Maddison A (2001) The world economy: a millennial perspective. OECD, Paris
Modelski G, Thompson WR (1988) Seapower in global politics, 1494–1993. Macmillan Press, Houndmills
Modelski G, Thompson WR (1996) Leading sectors and world powers. The coevolution of global politics and economics. University of South Carolina Press, Columbia
Moreira C, Eloranta J (2011) Importance of "weak" states during conflicts: Portuguese trade with the United States during the Revolutionary and Napoleonic wars. Revista de Historia Económica 29(03):393–423
Naylor RT (2001) Economic warfare: sanctions, embargo busting, and their human cost. Northeastern University Press, Boston

North DC (1990) Institutions, institutional change, and economic performance. Cambridge University Press, Cambridge/New York
North DC (1993) Institutions and credible commitment. J Inst Theoretical Econ 149:11–23
O'Brien PK (1988) The costs and benefits of British imperialism, 1846–1914. Past Present 120:163–200
Offer A (1989) The First World War: an agrarian interpretation. Clarendon Press, Oxford
Offer A (1993) The British Empire, 1870–1914: a waste of money? Econ Hist Rev 46(2):215–238
Olson M, Zeckhauser R (1966) An economic theory of alliances. Rev Econ Stat 48(3):266–279
Olson M (1993) Dictatorship, democracy, and development. Am Polit Sci Rev 87(3):567–576
O'Leary JP (1985) Economic warfare and strategic economics. Comp Strategy 5(2):179–206
Onorato MG, Scheve K, Stasavage D (2012) Technology and the era of the mass army. IMT Lucca EIC working papers series. Lucca, IMT Lucca, 5
O'Rourke K (2006) The worldwide economic impact of the French Revolutionary and Napoleonic wars, 1793–1815. J Global Hist 1(1):123–149
Ritschl A (2004) The Marshall Plan, 1948–1951. EH. Net Encyclopedia. Accessed 5 Aug 2009 http://eh.net/encyclopedia/the-marshall-plan-1948-1951/
Ritschl A (2005) The pity of peace: Germany's economy at war, 1914–1918 and beyond. In: Broadberry S, Harrison M (eds) The economics of World War I. Cambridge University Press, Cambridge, p 41
Russett B (1993) Grasping the democratic peace. Principles for a post-cold war world. Princeton University Press, Princeton
Sandler T, Hartley K (1995) The economics of defense. Cambridge University Press, Cambridge
Sandler T, Hartley K (1999) The political economy of NATO. Past, present, and into the 21st century. Cambridge University Press, New York
Sandler T, Murdoch JC (1990) Nash-Cournot or Lindahl behavior? An empirical test for the Nato Allies. Quart J Econ 105(4):875–894
Scheve K, Stasavage D (2010) The conscription of wealth: mass warfare and the demand for progressive taxation. Int Organ 64(4):529–562
Singer JD (1979) The correlates of War I: research origins and rationale. Free Press, New York
Singer JD (1981) Accounting for international war: the state of the discipline. J Peace Res 18 (1, Special Issue on Causes of War):1–18
Singer JD (1990) Variables, indicators, and data. The measurement problem in macropolitical research. In: Singer JD, Diehl P (eds) Measuring the correlates of war. University of Michigan Press, Ann Arbor
Smith A (1776) An inquiry into the nature and causes of the wealth of nations. Edwin Canna, London
Stevenson D (2011) From Balkan conflict to global conflict: the spread of the First World War, 1914–1918. Foreign Policy Anal 7:169–182
Strachan H (2011) Clausewitz and the First World War. J Mil Hist 75:367–391
Thomas M (1983) Rearmament and economic recovery in the late 1930S*. Econ Hist Rev 36 (4):552–579
Thornton M, Ekelund RB (2004) Tariffs, blockades, and inflation: the economics of the civil war. Scholarly Resources Inc., Wilmington, Delaware
Tilly C (1990) Coercion, capital, and European states, AD 990–1990. Basil Blackwell, Cambridge, MA
Trevino R, Higgs R (1992) Profits of US defense contractors. Def Peace Econ 3(3):211–218
Webber C, Wildavsky A (1986) A history of taxation and expenditure in the Western World. Simon and Schuster, New York
White EN (2001) Making the French pay: the costs and consequences of the Napoleonic reparations. Eur Rev Econ Hist 5(3):337–365
Wright Q (1942) A study of war. The University of Chicago Press, Chicago

War and Cliometrics in an Age of Catastrophes

Roger Ransom

Contents

Abstract

In an essay exploring the role of cliometrics and economic history in the context of broader disciplines of history and economics, Claude Diebolt and Michael Haupert point out that:

> Perhaps the biggest challenge facing economic history is that in its attempt to pursue truth, economic history is at the same time too vast and too small. In a historical sense, we try to accurately compile all the facts relevant to a given topic of study. The smaller the topic, the

R. Ransom (✉)
University of California, Riverside, CA, USA
e-mail: roger.ransom@ucr.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_57

easier it becomes to gather and arrange all the relevant facts, and the more rigorous the result is likely to be. . . . But for the historian who aims to create general truths, the economic, like any other conventional division of the subject matter of history, is too narrow a conception. (Diebolt and Haupert 2018, 3)

The search for general truths is compounded by the tendency for scholars to stay within the confines of their academic disciplines. Military historians write military history, economic historians worry about problems of economic history and development, and cliometricians stay focused on the use of economic theory and statistical methods to predict the past with ever greater precision. Nowhere is this challenge more evident than in the study of war and economics in the twentieth century.

Keywords
Catastrophes · Cliometrics · War

Introduction

In an essay exploring the role of cliometrics and economic history in the context of broader disciplines of history and economics, Claude Diebolt and Michael Haupert point out that:

> Perhaps the biggest challenge facing economic history is that in its attempt to pursue truth, economic history is at the same time too vast and too small. In a historical sense, we try to accurately compile all the facts relevant to a given topic of study. The smaller the topic, the easier it becomes to gather and arrange all the relevant facts, and the more rigorous the result is likely to be. . . . But for the historian who aims to create general truths, the economic, like any other conventional division of the subject matter of history, is too narrow a conception. (Diebolt and Haupert 2018, 3)

The search for general truths is compounded by the tendency for scholars to stay within the confines of their academic disciplines. Military historians write military history, economic historians worry about problems of economic history and development, and cliometricians stay focused on the use of economic theory and statistical methods to predict the past with ever greater precision. Nowhere is this challenge more evident than in the study of war and economics in the twentieth century.

An Age of Catastrophes

Everyone agrees that the First World War and its aftermath was a major turning point that fundamentally altered the course of world history in the twentieth century. What makes the changes brought about by the war and its aftermath even more dramatic is the suddenness with which these events occurred. In the space of 5 years, the German Empire, the Habsburg Empire, the Russian Empire, and the Ottoman Empire were overthrown and replaced with a collection of new nation states with very different political regimes. “Battle losses of World War I” writes historian

Michael Clodfelter, "were totally unprecedented in human history. Even the greatest battles of the continental wars of the seventeenth and eighteenth centuries in Europe paled beside those of World War I." (Clodfelter 2002, 479). If one adds the deaths resulting from the Russian Civil War and territorial conflicts immediately following the end of the Great War, at least 12 million men lost their lives. The Second World War was even more deadly. "The toll of World War II," according to Michael Clodfelter, "surpasses 30 million – with 40 million a more likely figure and some estimates going as high 55 million" (Clodfelter 2002, 581). These estimates suggest that, if you include the noncombatants – including the seven million victims of the holocaust who died during the war – as many as 80–100 million people may have died as a direct consequence of the two world wars.

The wars were only a part of the story. Military historians tend to view the two decades following the end of the Great War as a lull before the resumption of fighting in 1939. To economic historians, the interwar years were anything but a "lull"; they were a period of intense economic and social turmoil as people struggled to get back into some sort of "normalcy" amid the destruction and confusion surrounding the end of the First World War and were then confronted with a global depression. Among the most visible signs of the economic uncertainty caused by the war were the instability of international financial markets, the collapse of antebellum trading patterns, widespread unemployment throughout the industrial nations, and episodes of hyperinflation that paralyzed several countries immediately after the war. Historian Eric Hobsbawm has characterized the period from the outbreak of the First World War to the end of the Second World War as an *Age of Catastrophes*. "For forty years," he wrote, "[Western civilization] stumbled from one calamity to another. There were times when even intelligent conservatives would not take bets on its survival." (Hobsbawm 1994, 7). Figure 1 presents a quick view of the economic and military events that comprised the Age of Catastrophes. In addition to the "world wars," there were civil wars in Russia, China, and Spain that attracted attention from major powers throughout the world.

The Economics of the Great War

Cliometricians have provided a large body of quantitative research on the economic impact of these changes, including military expenditures and national income of the five major powers during and after the war. Their estimates show that military expenditures, which had totaled just over £468 million in 1913, jumped to £4.1 billion in 1914, and soared to just over £4 trillion in 1915, which was the first full year of fighting. Military personnel which had totaled just over five million men in 1913 had risen to 25.5 million men in 1915. Figure 2 plots the level of military expenditures and personnel from 1910 to 1940. The huge increase in spending during the war was matched by a dramatic demobilization at the end of the war that posed a whole new set of challenges as men returned from the fronts and government spending fell back to the prewar levels.

The need to mobilize resources for the war also had a pronounced effect on the composition of government spending in belligerent countries from 1913 through

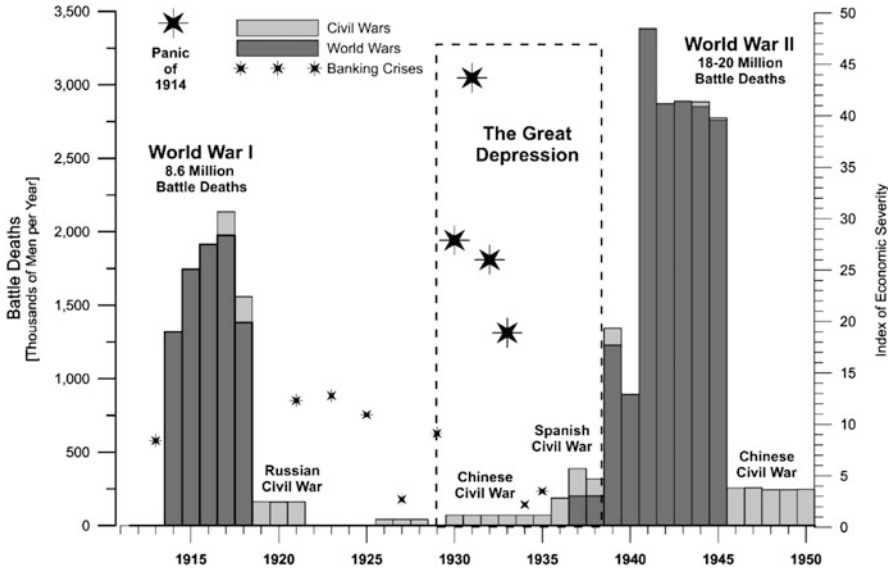


Fig. 1 An Age of Catastrophes, 1914–1945. (Sources: The data on military deaths is from Correlates of War Project (2010) and Clodfelter (2002). The dates and relative severity of banking panics are from Reinhart and Rogoff (2009) and Kindleberger and Aliber (2005). The severity of the crises is measured as the number of countries experiencing banking crises in that year weighted by the share of world GNP for those countries.)

1918. Figure 3 shows the share of national income allocated to military expenses by the five major powers during the war. In 1913, France, Britain, and Germany were all spending about 10% of national income for military expenditures; for France and Germany, military expenditures doubled by the end of 1914 and reached just under 50% by the end of 1915. The share of military expenditures in national income of all three countries remained at or above the 1915 level for the rest of the war. Austria-Hungary also saw a dramatic increase in the share of national income going to military expenditure during the first year and a half of the war. However, unlike the more developed economies of Western Europe, the Austrians were unable to sustain their war effort, and the share of military expenditures in national income actually decreased over the last 3 years of the war. The Russians, who were still recovering from the disastrous losses of their 1905 war with Japan, were already spending a third of their national income on military expenditures in 1913, and the burden of their war effort fell dramatically after the end of 1915. The growing unrest and opposition to the war produced a situation where the defense expenditures decreased as a fraction of income. Both the Austrian and the Russian economies were finding it difficult to finance the war expenditures.

In addition to having to supply the needs of the military, governments had to find a way to pay for the goods and services required by the war. It soon became apparent that the tax systems of peacetime governments would not meet the needs of waging a

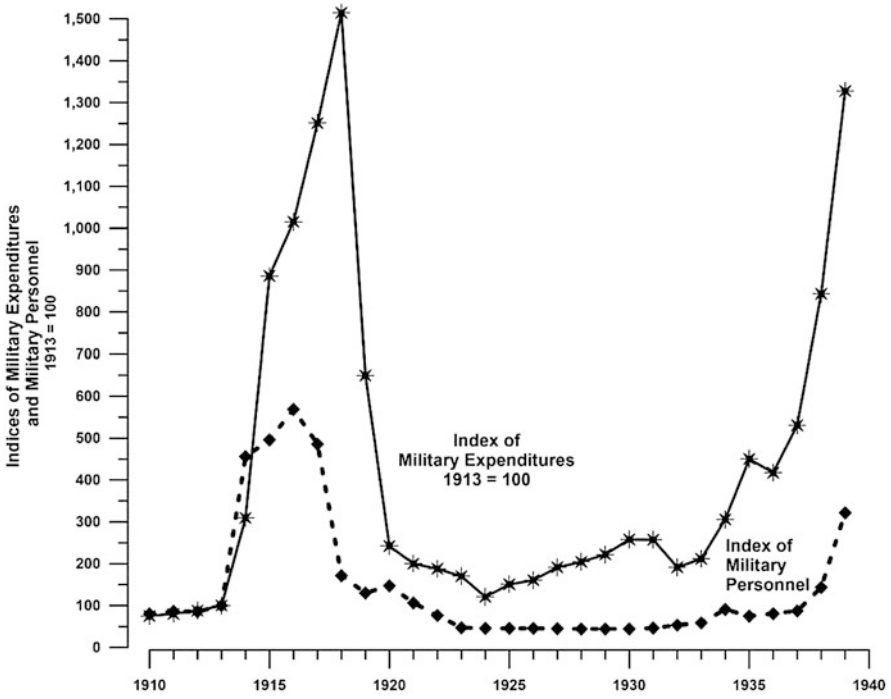


Fig. 2 Indices of military expenditures and military personnel of the major powers, 1900–1939 [1913 = 100]. (Sources: (Correlates of War Project 2010) Data set for military personnel and annual military expenditure in pounds sterling. The countries represented are France, Great Britain, Russia, Germany, Italy, Japan, The United States, and Austria-Hungary through 1918

major war. An obvious alternative to expanding revenues would be issuing government bonds. However, even for countries like Great Britain, which had a well-developed capital market, issuing bonds was not going to cover the costs for the war. In the absence of any other mechanism, governments were forced to simply print money to pay the for the huge budget deficits. The result of this sudden increase in the money supply was that, within months of the outbreak of hostilities, the belligerent countries began to experience significant increases in the level of consumer prices. Figure 4 presents data on consumer prices in Great Britain, Germany, France, Russia, and the Austrian-Hungarian Empire during and immediately following the war.

High rates of inflation, a phenomenon which none of the belligerent countries had experienced during the three decades leading up to the war, persisted for several years in every country. The great advantage of inflation was that it was a tax that nobody could avoid. Consumers paid what amounted to an "inflation tax" in the form of higher prices for the same bundles of goods. The British and French managed to limit the impact of inflation to a doubling of prices over the course of the war, but the other countries were not as successful. In Germany, prices had

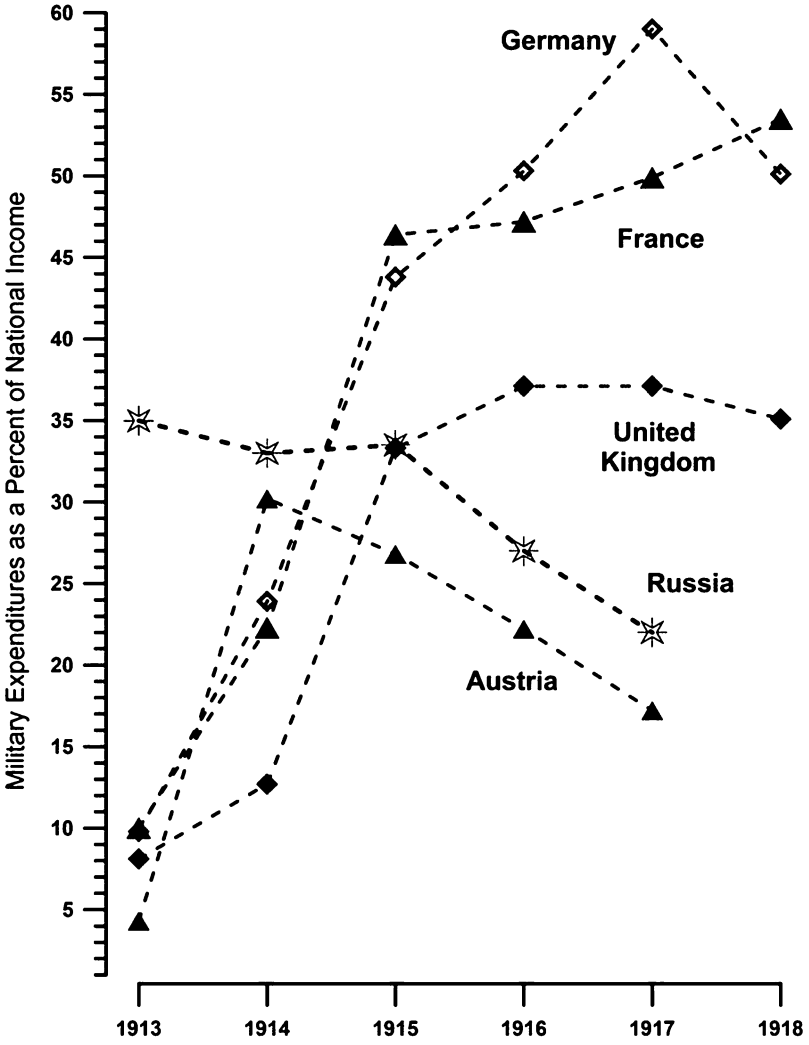


Fig. 3 Military expenditures as a fraction of national income, five European powers, 1913–1918. (Sources: Broadberry and Harrison 2005, 15, Table 1.5.)

quadrupled by the end of the war and the inflation spiral continued to grow out of control. By 1924, the German currency had become worthless. Similar episodes in hyperinflation right after the war were experienced by Russia and Austria-Hungary.

The First World War is often depicted as war where the industrial revolution had produced economies that could provide the weapons, munitions, and skills that armies needed for war in the twentieth century. However, as historian Avner Offer has pointed out, it was shortages of food, not weapons, that ultimately forced Germany and Austria out of the war (Offer 1989, 2000). In both countries,

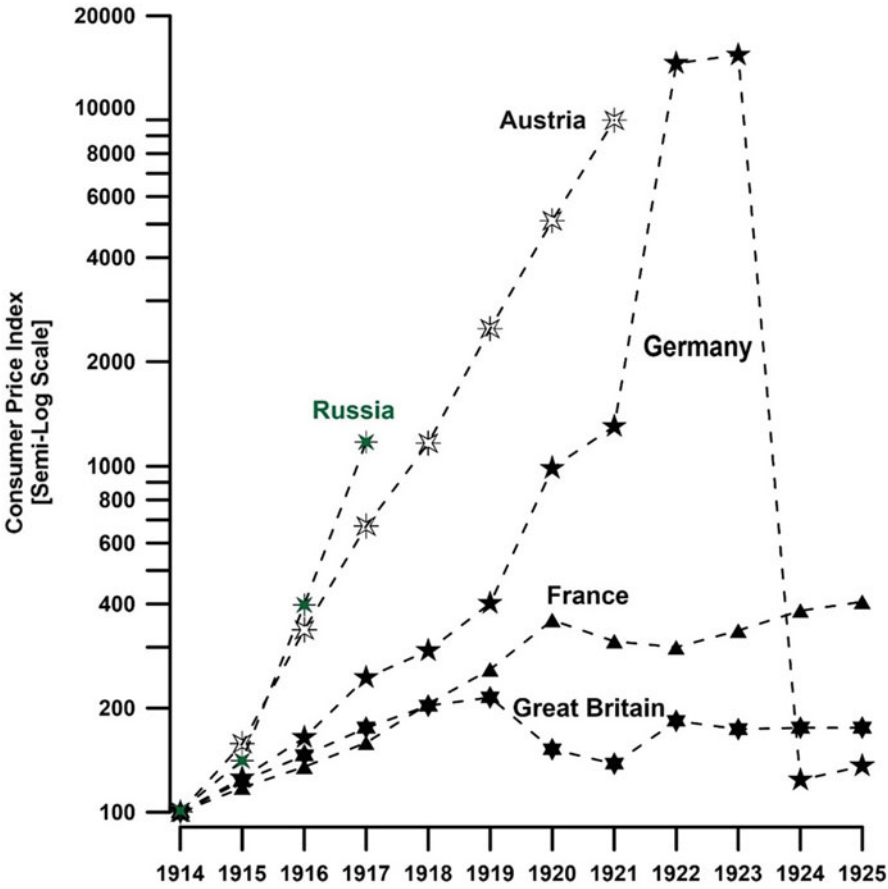


Fig. 4 Consumer price indices for five European powers, 1913–1925. (Sources: (Mitchell 1998, 865–6). All indices have been converted to 1914 = 100)

agriculture was very labor-intensive, and the labor shortage created by men being drafted into the Army placed serious constraints on the production of agricultural output. The shortage of manpower in agriculture was compounded by shortages of imported fertilizers which were no longer available because of the British blockade. The collapse of international trade meant that imported food was no longer available. The result was a constant crisis in the supply of food for the home front throughout the war.

Food was not the only problem posed by the collapse of international trade in 1914. The impact of the war on the international system of trade scrambled the markets for traded goods in the belligerent and neutral countries. The statistics that are most often used to measure international trade are the value of exports and imports. However, the high rates of inflation make it difficult to infer the physical volume of trade from these series. There are, however, enough data for specific

commodities to accept Findlay and O'Rourke's statement that "it is safe to conclude that the volume of trade fell sharply during the conflict, if by an unknown amount. However, such aggregate effects mask a large range of individual country experience." (Findlay and O'Rourke 2007, 435).

The War at Sea

The reliance on trade to supply food and other commodities created an "economic war" in the form of naval blockades to restrict the ability of countries to import or export goods. For the Entente Powers, this involved blocking the western and northern ends of the English Channel. There were however some problems with stopping ships belonging to neutral countries. Historian Mark Frey estimates that Dutch exports to Germany more than doubled between 1914 and 1916, and Dutch foodstuffs amounted to 50% of all German food imports (Frey 2000). For Germany, a blockade of Great Britain involved the use of submarines. In 1915, the Germans announced that any ship entering territory near the British Isles would be subject to attack by German U-boats. The experiment was successful from a military point of view, but the indiscriminate sinking of neutral ships brought forth strong objections from the United States and other neutral countries, and the policy of unrestricted U-boat attacks was rescinded after only a few months. When the Germans reinstated unrestricted submarine warfare in early 1917, U-boats were able to briefly reach a goal of 600,000 tons a month; however, this was not enough to starve the British economy. By the middle of 1918, the allied Navies were able to substantially reduce sinkings through the use of the convoy system to a point where losses from U-Boats no longer threatened food supplies in the United Kingdom.

Was the blockade a major factor in the outcome of the war? Cliometricians Lance Davis and Stanley Engerman presented a cliometric analysis of the German blockade during the First World War and concluded that it "directly accounted for only about a quarter of the decline in German food production. The other three quarters of the fall can be traced to the decline in domestic food production." (Davis and Engerman 2006, 230). The British blockade of Germany was more effective. Historian Albert Ritschl argues that the German reaction to the continuing blockade was an important element in the planning for any future war." (Ritschl 2005, 654–655). Also see, Hardach 1977, 148–150; Davis and Engerman 2006, Chap. 6.

The Chaos of Victory

The treaties that emerged from the Peace Conference in Paris in 1919 paid very little attention to the economic challenges created by the establishment of new national boundaries and governments in Eastern Europe and Russia. The Council of Four had very little interest in a continued involvement with the territories created by the destruction of empires. Britain, France, and the United States substantially reduced the size of their military forces, and they remained at that level until the eve of the

Second World War. The index of military spending also fell dramatically right after the war and did not begin to increase until the mid-1930s.

The British managed to stem the inflationary increase in prices by 1920 and were able to maintain fairly stable prices to 1925. The French elected to pursue a policy that slowed the increase in wartime prices, but they made no attempt to return to the antebellum level of prices. The Austrian and Russian economies eventually collapsed under the weight of hyperinflation by the end of 1917, while the newly created Weimar Republic of Germany inherited a price level four times above the 1913 level and soon found themselves caught up in an explosion of prices that eventually forced the government to revalue the system of currency in 1924. The economic disruption from inflation in every economy after the war, which is depicted in Fig. 4, was a significant factor shaping the political and economic reorganization of the Weimar Republic that emerged in the wake of wartime defeat (Feldman 1997).

Much of the blame for the economic and political problems associated with the economic chaos after the war was initially placed on shortcomings in the Treaty of Versailles. In the summer of 1919, John Maynard Keynes, who had been part of the British delegation to Paris, returned to Britain and wrote a scathing indictment of the treaty and the politicians who framed it. The treaty, according to Keynes, included

> . . . no provisions for the economic rehabilitation of Europe, nothing to make the defeated Central Empires into good neighbors, nothing to stabilize the new States of Europe, nothing to reclaim Russia; nor reached at Paris for restoring the disordered finances of France and Italy, or to adjust the systems of the Old World and the New. (Keynes 1920, 226)

Keynes laid the blame for the shortcomings of the treaty on the narrow attitudes taken by the Council of Four in their deliberations at the Paris Peace Conference. However, the Treaty of Versailles was only the first step in a prolonged process of establishing a peaceful world after 1918. Four additional treaties were signed with the other central powers, and military conflicts involving disputes over borders between the newly formed states of central Europe lasted for another 4 years.[1] Margaret Macmillan points out that the peace makers "could not foresee the future and they certainly could not control it. That was up to their successors. When war came in 1939, it was a result of twenty years of decisions taken or not taken, not of arrangements made in 1919." (MacMillan 2001, 493–494).

By 1925, some semblance of economic stability had returned in the form of stable prices in Germany, France, and Great Britain, but the economic disruptions caused by wartime mobilization had dramatically changed the global marketplace. The antebellum financial stability of world trade had rested on the acceptance of a *gold standard* that was the underlying mechanism for the payment of international credits

[1]The other treaties were: The *Treaty of Saint- Germaine- en- Laye*, signed with Austria on September 10, 1919; The *Treaty of Neuilly- sur Seine* with Bulgaria on November 27, 1919; The *Treaty of Trianon* signed with Hungary on June 4, 1920; and The *Treaty of Sèvres* with the Ottoman Empire on August10, 1920. Also see, Henig 1995; Boemeke et al. 1998; Ransom 2018, Chap. 8).

and debts. Currencies were convertible into gold on demand and linked internationally at fixed rates of exchange. Gold shipments were the ultimate means of balance-of- payments settlement. The demise of the Gold Standard and the resulting instability of exchange rates reflected the loss of confidence and the constant fear of unstable markets. In a world fearful of the unknown, the ties that bound the system of global payments were torn apart.[2] An added element of uncertainty in the international capital markets came from the Allied decision to impose reparation payments upon the defeated Germans. After much deliberation, the amount demanded was set at 132 billion gold marks, which was divided into three "bond issues" over several years. As events played out, the Germans eventually wound up paying only about 20 billion marks. Most economic historians have concluded that that the impact of the reparation payments was more of a political problem than an economic problem (Ferguson 1999; Ritschl 2005; Marks 2013). The reparation payments eventually became a political football that was kicked around the Reichstag and Allied governments until Adolf Hitler repudiated the remaining debt in 1933.

Economic Collapse: The Great Crash

All this economic chaos came to a head with the collapse of stock prices on the New York Stock Exchange on October 29, 1929. Figure 5 presents the monthly index of the closing value of stocks on the New York Stock Exchange from October 1929 to January 1934. The 36-point decline of the stock index on "Black Thursday" 1929 was the beginning of a downward plummet of stock prices that continued for the next 3 years. By July of 1932, the index of stock prices had fallen to only 12.3% of its value in 1929. The decade of the 1930s was a period of falling commodity prices, declining industrial production and gross domestic product, a collapse of world trade, and very high levels of unemployment throughout the world.

The Great Crash had become the Great Depression.

People were still struggling to recover from the chaos of the war when the economic foundations of the global economy crumbled under the pressure of global economic panic. Figure 6 presents four indices that measure the economic performance of the global economy from 1929 to 1938 (Crafts and Fearon 2013b). What is immediately apparent is that all of these indices experienced enormous changes. The index for gross domestic product declined by 15% in 3 years and was only 5% above the 1929 level by the end of the decade. The stagnation of world trade is also evident. The two variables that experienced the greatest fluctuation were the level of employment and the index of consumer prices. After a decade in which the world had

[2]For more on the loss of confidence associated with the problem of exchange rates and the gold standard, see Eichengreen and Temin 1997; Eichengreen 1992; Eichengreen 1991; Ransom 2018; Findlay and O'Rourke 2007, Chap. 3; Wolf 2013.

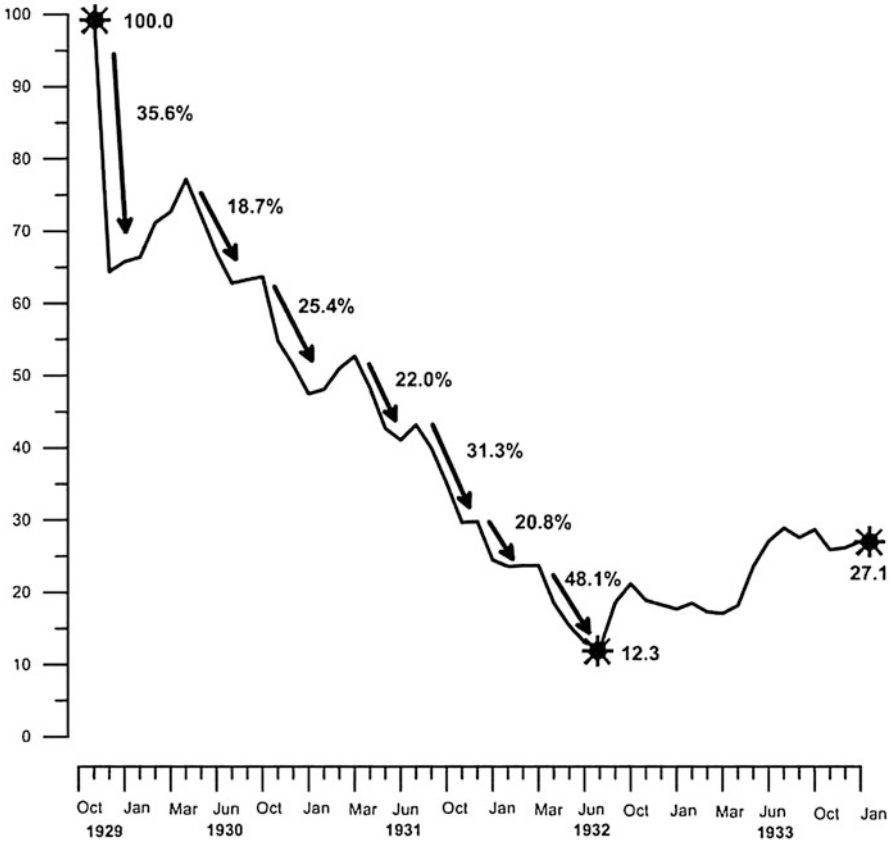


Fig. 5 Average closing value of the New York Times stock index, October 1929–January 1934. (Source: Ransom 1981, 118–19)

experienced runaway inflation, the 1930s saw a 30% decline in prices by 1933. The global index of unemployment reached 30% in 1933 and was still about 15% in 1939.

These phenomena could not be explained by the existing economic theories of how markets were supposed to behave. "Financial instability," economist Hymen Minsky observed, "is a nonevent, something that just cannot happen, insofar as the standard body of today's economic theory is concerned." (Minsky 1982, 13–14). Charles Kindleberger, perhaps the most celebrated chronicler of financial crises, refers to them as "hardy perennials" that are not easily explained. "Econometricians among my friends," he notes, "tell me that rare events such as panics cannot be dealt with by the normal techniques of regression but have to be introduced exogenously as "dummy variables." (Kindleberger 1978, 8).

Faced with a situation where the state of the world did not correspond to the theoretical assumptions of their economic models, economists in the 1930s looked

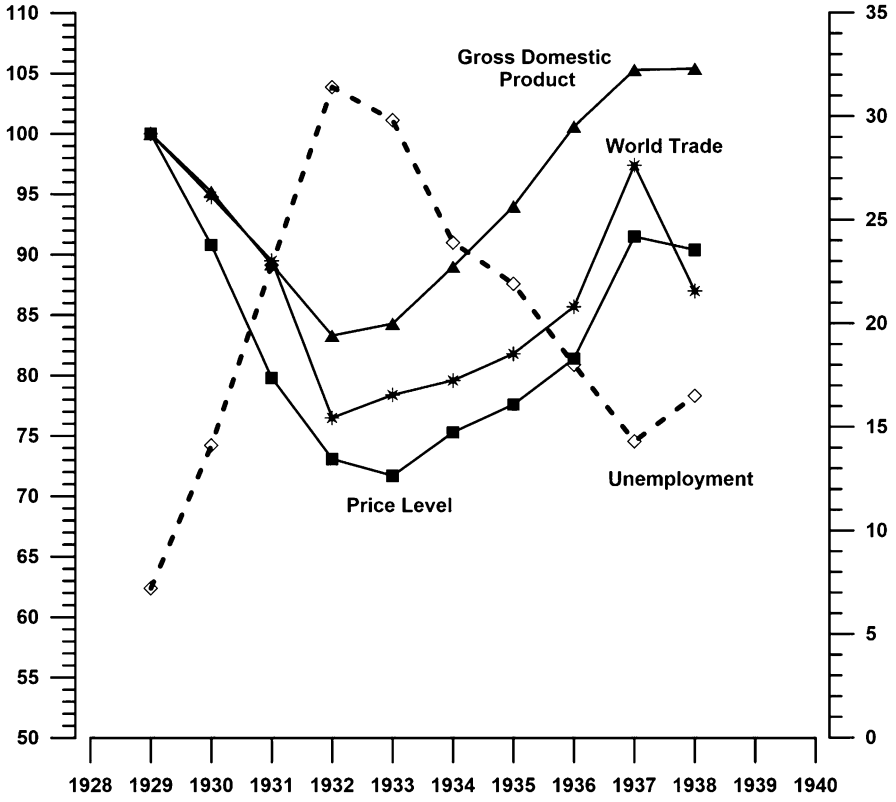


Fig. 6 The Great Depression in Europe, 1929–1939. (Sources: (Crafts and Fearon 2013b, Table 1.1, 1) The authors cite the following sources for each variable: *gross domestic product* (Maddison 2010), Western European and European offshoots; *price level* (Nations 1941) 17 Countries; *unemployment* (Eichengreen and Hatton 1988) 11 countries; and world trade (Maddison 1985) 16 countries.)

for new ways to explain what was happening. In 1936, John Maynard Keynes published *The General Theory of Employment, Interest, and Money* (Keynes 1936). Keynes' macroeconomic approach challenged the existing paradigm of economics markets, which argued that the aggregate level of employment and production would always move towards an equilibrium that guaranteed full employment of resources. Keynes suggested that there would be times when expanding the scope of government spending would be the only effective way of eliminating unemployment.

Keynes's macroeconomic model made it possible to measure the impact that the economic collapse of the 1930s had on the world economy, but by itself, the theoretical framework could not explain the causes of the global economic collapse. Keynes himself admitted as much with his comments at the end of *The General Theory* about the role of "animal spirits" in sparking the boom that led to the Great

Depression. As scholars searched for the key to understanding the Great Depression, they realized that the economic collapse was tied to the political and social changes brought about by the First World War (Crafts and Fearon 2013b; Temin 1981, 1989; Feinstein et al. 2008).

At the root of the problems facing policymakers in the years after 1918 was the inescapable fact that, despite its ferocity, the Great War did little to resolve any of the political issues between the major states of Europe that existed in 1914. The five treaties which emerged from the Peace Conference in Paris in 1920 created a host of new problems in the form of nation states formed from the collapse of German, Austrian, and Russian empires. While there was no resumption of major military conflicts between the larger states of Europe in the interval 1919–1937, there were numerous "border conflicts" and three civil wars, which attracted intervention by third parties at one time or another.

The interwar years were a period when new leaders pursued aggressive policies to maintain or expand their economic and military status in the postwar environment. None of the belligerents were happy with the outcome of the Great War. The British had saved their Empire, but the effort left them with an exhausted economy that would struggle to maintain their economic status as a great power through the next two decades. The French economy was also stretched to the limits of its capacity, and they were fearful that the stipulations of the Versailles Treaty designed to limit Germany's ability to reassert its economic power in central Europe were far too weak. An added element of uncertainty over the next two decades was the frequency with which both Britain and France experienced changes in the government. Italy felt cheated out of their share of the spoils of victory and Italian nationalists claimed this impeded Italy's progress in becoming a "Great Power." In the fractured world of postwar Italian politics, political sentiment moved steadily to the right, culminating with Benito Mussolini's fascist party gaining control of the Italian government in 1925. The Americans showed little interest in European affairs after the war. President Woodrow Wilson was unable to gain support for American participation in the League of Nations and The United States Senate did not approve the Treaty of Versailles. Lacking any previous experience with democratic institutions, and faced with a very uncertain future, postwar governments in the new states moved towards authoritarian rule as the only way to maintain a stable government. By the time the Great Depression had run its course, the only country in central Europe that had managed to maintain democratic institutions was Czechoslovakia.

Russia: The Union of Soviet Socialist Republic

Among the most difficult challenges posed by the chaos after the war were the problems facing the former Russian Empire. In March 1918, the newly formed Bolshevik government had signed the Treaty of Brest-Litovsk with Germany, which ended the fighting between Germany and Russia. The Russians agreed to abandon their claims to Poland, Finland, and the Baltic states, and the Germans continued to occupy part of western Russia that included the Ukraine and Crimea.

The Treaty of Versailles negated this huge transfer of Russian territory to the Germans; however, the territory was not returned to the Russians (Ransom 2018; Herwig 1997; White 1994). The Allied powers did not recognize the Bolshevik government and the Russians were not invited to the Paris Peace Conference in 1919.

Five years of fighting a bitter civil war had produced a command economy in Russia where every aspect of economic activity was tightly controlled by the state. Vladimir Lenin proposed an ambitious set of reforms known as the *New Economic Policy* [NEP]. The strict restrictions on agricultural and industrial prices and production were eased; labor markets were given greater flexibility to accommodate the needs of producers, and individuals could form their own small enterprises. The immediate effect of these changes was a rapid expansion of agricultural output produced on small family farms. However, the industrial sector, which remained largely under the control of the state, experienced only a slight increase in output. The opportunity to form private enterprises created a new group of entrepreneurs – called "NEPmen" – who formed private enterprises in both urban and rural areas (Ball 1987). Lenin regarded the NEP to be a temporary measure that would help shape a Soviet economy that would rely on markets as well as government controls (Fitzpatrick et al. 1991).

On January 21, 1924, the man who masterminded the Bolshevik victory passed away. Vladimir Lenin had been the heart and soul of the Bolshevik victory in the civil war and his death prompted a spirited scramble among the Bolshevik leaders eager to replace him at the head of the Soviet government. The principal contenders were Leon Trotsky, who was the man that Lenin would probably have chosen to be his successor, and Joseph Stalin, who was the general secretary of the Communist Party. Stalin was able to use his political position to eventually emerge as the head of the Soviet government (Kotkin 2014; Service 2006).

Stalin had originally been a strong supporter of Lenin's NEP; however, the two men gradually drifted apart as Stalin become convinced that even the limited role of market capitalism that was embedded in Lenin's NEP was not well suited for the Soviet Union in a world where all of the major powers were likely to be hostile to the Soviet Union. While Lenin had hoped to integrate the Soviet Union in the global economy, Stalin envisioned a centralized command economy characterized by what he called "Socialism in One Country." His primary objective was to provide the economic basis for a military machine to defend the socialist state.

Economic historians have devoted considerable effort to the development of quantitative estimates of economic variables in the Soviet Union in the interwar period. Figure 7 provides data relating to the growth of population and gross national product from 1928 through 1938. There was virtually no growth in either GDP or population before 1932. Stalin's attempt to reorganize agriculture by collectivized farms was an economic disaster that produced a food crisis that reached the level of a famine in many rural areas during the 1930s (Davies 1980; Davies and Wheatcroft 2004). From a global perspective, the main effect of these policies was to isolate the Soviet Union from the economic and political tensions that were sweeping through Western and Central Europe (Davies 1989; Davies

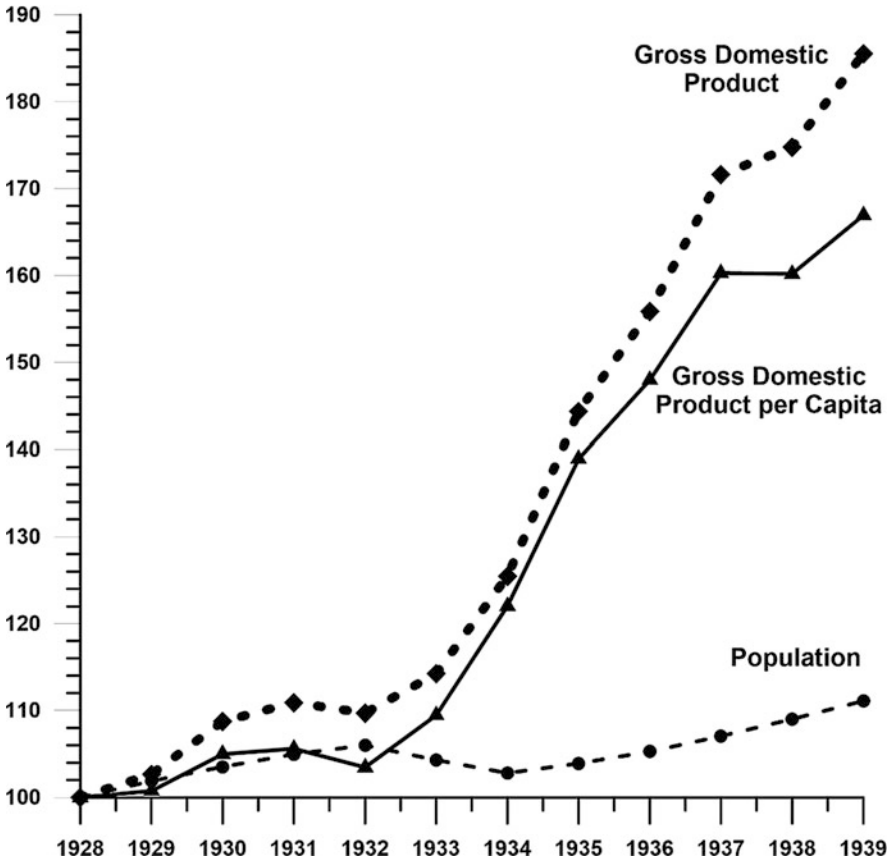


Fig. 7 Gross domestic product and population of the Soviet Union, 1928–1939. (Source: Davies et al. 2018, Table 1, p. 29.)

et al. 2018). For Stalin, all of the other powers were potential enemies by 1937. He was determined that the Soviet Union would be ready wherever the next blow might fall.

Germany: The Rise of the Nazi Party

By 1929, the Weimar Republic had managed to establish some degree of economic stability; however, the political situation remained unsettled. In November 1923, a group of Nationalist Socialist Party members, led by Adolf Hitler, staged an unsuccessful effort in Munich to gain control of the Bavarian government. Hitler and several of his followers were sentenced to 3 years in Landsberg Prison. Hitler made good use of his prison time, spending several hours of each day dictating notes to a

colleague, for a book that was published in 1925 and 1926 (Hitler 1936). In the summer of 1928, Hitler wrote a second manuscript that was apparently intended as a sequel to *Mein Kampf*. The existence of this manuscript, which was not found until 1958, was published in 1961 (Hitler 1961). Hitler's thoughts present a blueprint for the policies which he would implement if he were ever to gain power. At the center of those plans was the need to expand Germany's boundaries to include populations that Hitler regarded as "German." The borders of Bismarck's Reich, according to Hitler, "encompassed only a part of the German nation." The pre-1914 borders, Hitler insisted, "ran straight across German language areas, and even through parts which, at least formerly, had belonged to the German Union, even if in an informal way." (Hitler 1961, 48–49). Hitler intended to see that the German state reclaimed these areas. "No foreign policy aim," he wrote, "could have been more obvious for the strictly formal national state of that time than the annexation of those German areas in Europe which, partly through their former history, had to be an obvious part not only of the German nation but of a German Reich." (Hitler 1961, 56).

The global economic collapse of 1929 had a devastating effect on the German economy. Between 1929 and 1932, gross domestic product fell by 25% and unemployment reached 31% (Crafts and Fearon 2013a, 3). The economic crisis was accompanied by a collapse of the political coalitions in the Reichstag. The Nazi party emerged from the 1932 federal election with 230 seats, which was more than any other party, but not enough to form a government. Paul von Hindenburg defeated Hitler in the presidential elections of 1932, but he was unable to form a government that did not include the Nazis. Eventually, the president realized he had no other choice than to appoint Adolf Hitler as Chancellor of Germany. By the end of July 1933, Hitler had effectively taken control of the German government (Fischer 1995; Abelshauer 1998). As soon as he became Chancellor, Hitler began to expand expenditures of the German government. Figure 8 presents data on German government expenditures between 1932 and 1938. In 1932, government spending was 1.4% of gross domestic product; by 1938, it had grown to 19%. Military expenditures grew from about one-fourth of the total budget to just under three-quarters of all government expenditures.

In January 1935 voters, in the Saar region of Germany, which had been occupied and governed by the United Kingdom and France from 1920 to 1935 under a League of Nations mandate, voted overwhelmingly to return to German control. A year later, German troops marched into the Rhineland without meeting any opposition. Hitler's next target for expansion of the Reich was the German population of Austria. When his initial plan to set up a pro-Nazi government in Austria did not work out, German troops crossed the border into Austria and were warmly greeted by the Austrians. Hitler proclaimed an *Anschluss* which was approved by the Reichstag on March 13 and ratified by a plebiscite on April 10, 1938. There were 3.2 million Germans who lived in the Sudetenland – a region in the western part of Czechoslovakia – which Hitler claimed should be "returned" to Germany. Although the Czech Government was strongly opposed to having them unified into Germany, none of the Western Powers were prepared to challenge Hitler's claim. Benito Mussolini encouraged Hitler to host a Conference just outside Munich on September 30, 1938, to

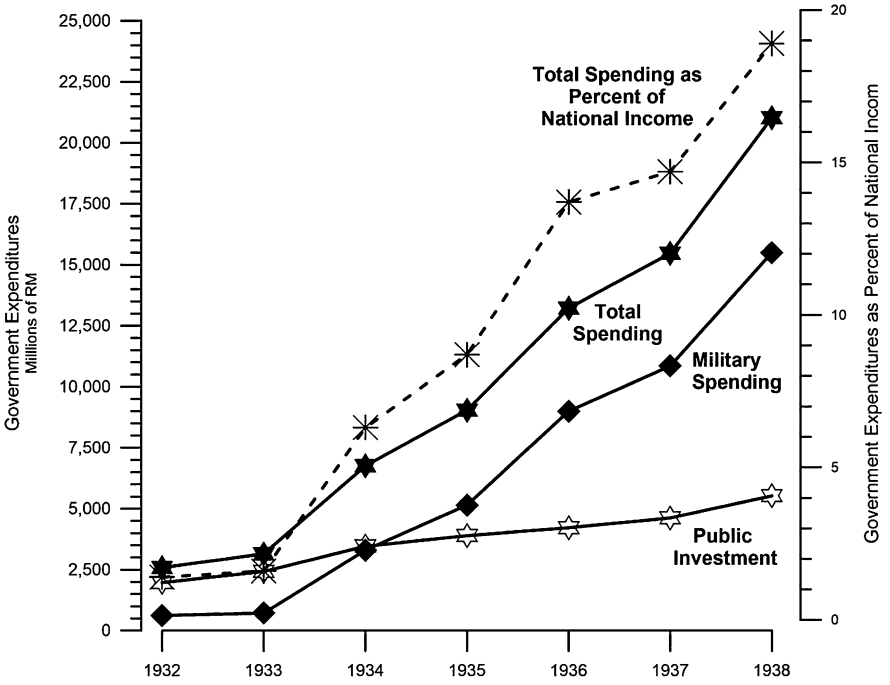


Fig. 8 German government expenditures, 1932–1938. (Source: Abelshauer 1998)

determine the fate of the Sudetenland Germans. The result was an agreement signed by Mussolini, Hitler, Neville Chamberlain of Britain, and Eduoard Daladier of France that the Germans would occupy the Sudetenland. In effect, the agreement allowed Hitler to control all of Czechoslovakia (Weinberg 1995, Chaps. 8 and 9, Ferguson 2006).

Hitler now turned his attention to the Germans living in Poland. Stalin was not likely to quietly watch Germany gobble up all of Poland, so Hitler sent his foreign secretary – Joachim von Ribbentrop – to Moscow to work out a "nonaggression pact" with the Soviets that included a secret agreement to divide Poland. On September 1, 1939, the German troops invaded Poland. This time Britain and France both declared war on Germany.

The Second World War had begun.

Japan: The Empire of the Rising Sun

While the Germans and the Russians were rebuilding their countries after the war, another empire was emerging halfway around the world. The Japanese had been isolated from the rest of the world until the middle of the nineteenth century when a series of visits to Tokyo Bay by the American Admiral Mathew Perry forced them to

open their economy to the world of international trade. In 1868, a group of younger nobles and samurai staged a rebellion that placed a 16-year-old Emperor in charge of the government. Despite their fears of foreign intervention, the leaders of the Meiji Restoration were quickly drawn into the international marketplace. Figure 9 provides data on the Japanese gross domestic product and foreign trade from 1875 to 1940. The index of trade – measured here as the sum of exports and imports – provided a wide range of goods, services, and raw materials. The Japanese economy experienced a steady growth in GDP rate of about 2.5% per year in the years leading up to World War I. International trade was an important part of that growth up to World War I, and then became highly volatile in the two decades after the war.

The Meiji government was willing to use military force to expand their influence in Manchuria, Korea, China, and Taiwan. In 1895, the first Sino Japanese war ended with the Treaty of Shimonoseki. The Chinese were forced to grant the Japanese favorable treaty status similar to that accorded the European powers, and Taiwan became a Japanese province (Paine 2003, 2017). In 1904, Japan declared war against

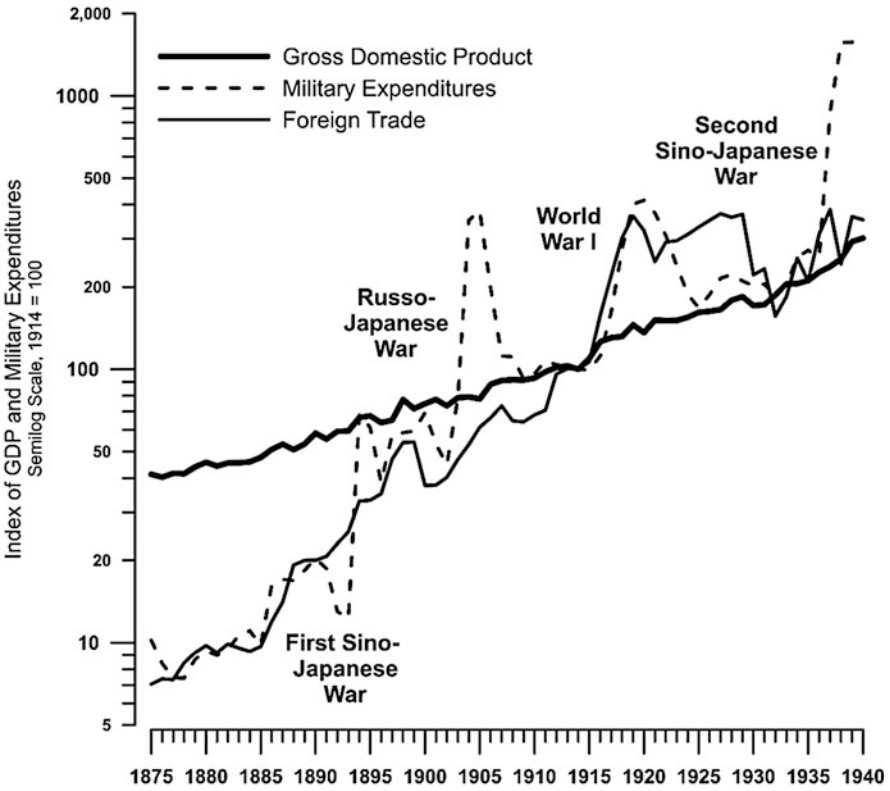


Fig. 9 Japanese gross domestic product, foreign trade, and military expenditures, 1874–1940. (Sources: Gross national product (Maddison 2010); military expenditures (Correlates of War Project 2010).)

Russia to settle a territorial dispute over Manchuria. The war ended in 1905 with a settlement negotiated by American President Teddy Roosevelt (Jukes 2002; Ransom 2018). The Japanese could have stayed neutral in the First World War, but Prime Minister Okuma Shigenobu saw the European war as an opportunity for Japan to further increase its Pacific Empire. Japan declared war on Germany in August of 1914 and proceeded to seize the German naval installation at Tsingtao and the German colonial possessions in the Marshall Islands. By the end of the war, Japan had managed to collect the nucleus of a colonial Empire in Asia. The dramatic impact that each of these wars had on Japanese military expenditures is evident in Fig. 9.

Official Japanese policy throughout the 1920s was to stay out of Chinese affairs; however, they had few qualms about interfering with events in Manchuria. When disagreements between the Kuomintang Party of Chaing Kai-shek and the Communist Party of Mao Zedong erupted into a civil war, the Japanese saw their opportunity to take control of Manchuria. On September 10, 1931, a group of junior army officers staged a minor explosion along the South Manchuria Railway which they blamed on Chinese forces. The "Mukden Incident" served as an excuse for the Japanese Army to occupy the town of Mukden. The generals quickly discovered that the Chinese were not in a position to oppose further Japanese advances, and within several months, the Japanese army occupied all of Manchuria. In February 1932, the Japanese government established Manchukuo as an independent state. By the end of 1937, invading Japanese troops occupied the eastern coast of China (Taylor 2015).

Italy: Too Small to Be a Major Power

Benito Mussolini came to power in 1922 with the promise that he would restore Italy to its former glories. The cornerstone of Mussolini's plan to return Italy to its former glories was the expansion of Italy's colonial foothold in North Africa. In October 1935, the Italians invaded Ethiopia, and they eventually succeeded in subjugating the African nation by the fall of 1936. Adolf Hitler's support for Mussolini's actions in Africa, together with their shared objective of undermining the Treaty of Versailles and expanding their territories provided the basis for what would gradually become a very close military and economic alliance. When Hitler became Chancellor of Germany in 1932, the two countries signed the *Pact of Steel*, an agreement between Fascist Italy and Nazi Germany that eventually led to the creation of the Axis Powers of World War II.

All of this took place in the context of an Italian economy that had been stretched to the limit by their efforts in Ethiopia and Mussolini's support for Francisco Franco's efforts in the Spanish Civil War. Figure 10 presents data on Italian gross domestic product and foreign trade (measured as the sum of imports and exports). GDP was stable through the period 1929 through 1938; however, the index of imports and exports fell dramatically after the great crash and did not recover over the course of the next decade. Vera Zamagni argues that the loss of imports created serious problems in obtaining strategic raw materials – particularly liquid fuels – during the war (Zamagni 1998, 187–88, Table 5.6). Figure 10 also shows that

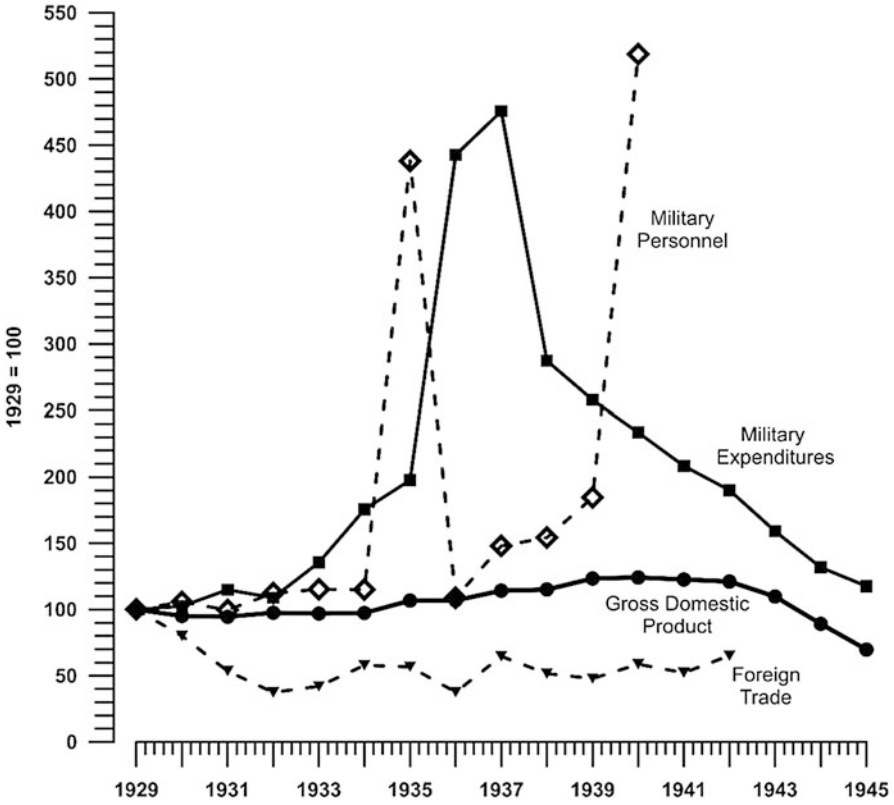


Fig. 10 Indices of gross domestic product, foreign trade, and military personnel and expenditures for Italy, 1929–1945, 1929 = 100. (Sources: Gross domestic product (Maddison 2006); foreign trade, military expenditures, and military personnel (Correlates of War Project 2010).)

military expenditures rose sharply after 1932, and the impact of the Ethiopian and Spanish military adventures is evident in the explosion of military expenditures and personnel in 1935–1937.

Mussolini had one more territorial ambition. Having expanded his colonial holdings in North Africa with the acquisition of Ethiopia, he now shifted his focus on a plan to seize the Suez Canal. In the spring of 1940, Italian troops launched a series of attacks against the British forces in Egypt (Ferris and Mawdsley 2015a; Clodfelter 2002). Looking at the Italian war effort, Vera Zamagni concludes that "If one includes human losses, it is beyond doubt that World War II cost Italy less than World War I, and far less than many other combatants. The Italian economy was unable to engineer any sizable economic expansion during the war." (Zamagni 1998, 212). Simply put, Italy could not afford the Second World War. When the Allied forces successfully invaded Sicily in July 1943, Mussolini's government was overthrown, and a puppet government controlled by Nazis used German troops and continued the fight against the Allied forces in Italy (Moseley 2006).

Blitzkrieg: A New Form of War

In the Spring of 1940, the world was once again on the verge of a World War. The Japanese had already been fighting in China for 2 years. The Germans and the Soviet Union had dealt with the Poland issue, and Hitler was preparing to invade France. It is tempting to say that this was all just a continuation of the Great War of 1914–1918. The same countries were involved in both wars, and at least for Western Europe, the story seems very familiar. But 20 years of political and economic disruption together with new leadership and new technology for fighting wars had produced a very different scenario by 1940. In 1914, the great powers had stumbled into a war they did not want. In 1939, there were three countries – Germany, Japan, and Italy – who were consciously pursuing policies which they knew would eventually escalate into major wars.

The major battles in the First World War had been characterized by prolonged periods of furious fighting that ended without either side being able to claim a victory. Examples include Verdun (1916), the Somme (1916), Passchendaele (1917), and the Ludendorff offensives (1918) on the Western Front; Gallipoli (1915) in the Middle East; and the Brusilov Offensives (1916) on the Eastern Front. All of these battles were titanic struggles involving thousands of casualties, yet none of them created a significant turning point that changed the outcome of the war. World War I was a war of attrition that not only produced horrendous casualties but it also consumed economic resources at a prodigious rate. In the end, it was the economic collapse of the Central Powers, not adverse military outcomes, that forced them to ask for a cease fire (Ransom 2018, Chaps. 7 and 8).

The Second World War, by contrast, was a war of motion and battles whose outcomes fundamentally altered the course of the war. The Blitzkrieg tactics of the German army combined armored vehicles and infantry in Panzer Divisions that historian Karl Heinz Frieser argues "caused a revolution in the image of war in comparison to battles of attrition of the first world war. The principle of physical or not annihilation was replaced by the principle of psychological confusion." (Frieser 2015, 351). In May of 1940, the German Army marched through Belgium and Holland and invaded France. The British BEF was pinned against the English Channel at Dunkirk, where they managed to evacuate over 300,000 men but lost almost all of their equipment and vehicles. The French army collapsed in the face of the German offensive to the South, and on June 22, 1940, an armistice was signed by France and Germany at Compiegne in the same railroad car used for the German surrender to the Allied Powers in 1918. The north and west coasts of France and their hinterlands were occupied by the Germans. What remained of the Third Republic was replaced by a government headed by Philippe Petain, who cooperated with the Nazis. In the space of 6 weeks, Hitler and his generals had managed to accomplish what 5 years of bloody fighting on the Western Front had failed to accomplish 20 years earlier, and they did so at a surprisingly low price.

By the end of 1941, the Third Reich controlled territory from the western coast of occupied France to the Soviet border in Poland. Table 1 summarizes the territorial

Table 1 German expansion, 1938–1941

Country	Date	Population [Million]	Territory [000 Sq. Mi.]	GDP [Million $]
By annexation				
Austria	Mar-1938	6.8	84	24.2
Czechoslovakia	Mar-1939	10.5	140	30.3
		17.3	**224**	**54.5**
By military force				
Poland	Sep-1939	35.1	389	76.6
Norway	Apr-1940	2.9	323	11.6
Denmark	Apr-1940	3.8	43	20.9
Netherlands	May-1940	8.7	33	44.5
Belgium	May-1940	8.4	30	39.6
France	May-1940	42.0	551	185.6
Yugoslavia	Apr-1941	16.1	248	21.9
Greece	Apr-1941	7.1	30	19.3
		124.1	**1,747**	**420.0**
Allied states				
Bulgaria	Mar-1941	6.6	103	10.5
Rumania	Apr-1941	15.6	295	19.4
Hungary	Jun-1941	9.2	117	24.3
		31.4	**515**	**54.2**
Germany	1938	**68.6**	**470**	**351.4**
Total German gains		**172.8**	**2,486**	**528.7**

Sources: Population (Maddison 2006), territory and gross domestic product (Harrison 1998a)

gains to the German Reich from the Anschluss with Austria in 1938 to the invasion of Greece and Yugoslavia in April 1941. In addition to the areas occupied by German troops, three countries – Bulgaria, Romania, and Hungary – allied themselves with the Nazi regime. Including the three allies in the Balkans, the Third Reich had increased its territory from 470,000 square miles in 1938 to 2.5 million square miles in 1941, together with a doubling of the population.

Hitler had hoped that Britain would leave the war after the defeat of France. Instead, Winston Churchill stood defiantly in the House of Commons declaring that Britain would fight on. Hitler and his generals had made no plans for an invasion of Great Britain, and it soon became clear that the British Fleet and the Royal Air Force [RAF] made the success of such a venture doubtful at best. Herrmann Goering insisted that the Luftwaffe could force Britain out of the war by air attacks. The "Battle of Britain" commenced in early July, and by August, German air raids were having an effect on the British defenses. However, the *Luftwaffe* could not sustain the level of losses it was incurring, and Hitler and Goering decided to switch the focus of German raids away from the RAF bases to bomber attacks on major cities – particularly London. By the beginning of October, it was apparent that the air attacks could punish the British, but

they were still in the war. Hitler cancelled Operation Sea Lion in September of 1940 (Ferris and Mawdsley 2015b).

In September 1940, Germany, Japan, and Italy signed the *Tripartite Pact,* which was the last in a series of treaties and agreements that formed the Axis Powers.[3] Hitler hoped that linking the military aims of the three countries would facilitate his plans for German expansion. As things turned out, his new partners proved to be a mixed blessing. The Japanese were preoccupied with a major war in China and unable to offer the Germans any immediate assistance against the Soviet Union. Mussolini hoped to drive the British out of Egypt; however, by the end of February, the Italians had been driven west all the way to Tobruk, and Hitler was forced to come to his ally's rescue. In March of 1941, he dispatched Erwin Rommel in command of the *Afrika Corps* to stabilize the situation there. Rommel was able to drive the British back into Egypt by the end of the fall campaign.

While the Germans were taking control of Europe, the Japanese were expanding their Empire in Asia. By the end of 1941, Japanese forces occupied all of the major cities along the eastern coast of China from Manchuria to Indo-China (Van De Ven 2015). The United States and Britain vigorously objected to the Japanese aggression, but there was not much they could do to check the Japanese military advance in China. In August 1940, Japanese Foreign Minister Matsuoka Yosuke announced the creation of the *Greater East Asia Co-Prosperity Sphere* [GEACS]. This was put forward as a coalition of governments in Japan, Manchukuo, and China that would eventually lead to the construction of a new order in Greater East Asia. A major force behind the creation of the GEACS was the need for the Japanese to secure access to natural resources such as oil and raw materials that were an essential part of the Japanese military-industrial complex. The United States had been a major supplier of scrap iron and petroleum products, and American financial institutions controlled the payments for Japanese international trade. On July 25, 1941, President Roosevelt issued an executive order freezing all Japanese assets in the United States and placing an embargo on oil and gasoline exports to Japan. The effect of the President's order was dramatic. The Japanese were gripped by a paranoia, which convinced them that the United States would the able to choke off the resources Japan needed to survive (Miller 2007). The American economic sanctions gave added credence to those in Japan who believed that a war against the United States was the only way for Japan to accomplish its aims for an Empire in Asia. Neither the Americans nor the British military forces were in a position to seriously intervene with the battles in China. The only real threat to the Japanese military aggressions in China or Southeast Asia was the American Navy. Plans were already underway for an IJN strike against Pearl Harbor.

[3]Hungary (November 1940), Bulgaria (March 1941), and Romania (November 1941) subsequently joined the Tripartite Agreement.

Barbarossa: The German Invasion of Russia

On the eve of his greatest gamble, Hitler wrote a letter to Benito Mussolini explaining why the invasion of the Soviet Union was necessary. The elimination of the Russians, he assured Mussolini, would not only provide security in the east for Germany and Italy; it would also represent "a tremendous relief for Japan in East Asia, and thereby the possibility of a much stronger threat to American activities through Japanese intervention." (Hitler 1941).

At 3:00 am on the morning of June 22, 1941, the massed German guns along a 1,800-mile front in western Russia erupted with terrifying fury. Operation Barbarossa, the German invasion of the Soviet Union, was underway. The invasion consisted of three army groups. Army Group North had as its objective the capture of Leningrad. Army Group Center, which was the largest of the three armies, had as its objective the city of Moscow. Army Group South was aimed at the city of Kiev and the oil fields of the Caucasus.

Despite warnings from the British and his own intelligence people that the Germans were planning an attack on the Soviet Union, Stalin was taken completely by surprise. By the end of September, the front lines stretched in a fairly straight line from Lake Ladoga in the North to the top of the Crimean Peninsula in the South. German forces occupied a huge part of the Soviet Union, but they had not yet captured either Leningrad or Moscow. Summer was coming to a halt, and the Red Army, though severely battered, was still in the field (Glantz 2012). In October, the first rains complicated army maneuvers for both sides, and the situation would get worse as fall turned into winter. Operation Typhoon, the final push organized to capture Moscow before the winter weather struck, was launched at the beginning of October. The Germans managed to encircle three Soviet armies within 2 weeks, and the spires of Moscow were in the sights of advance units by the beginning of December. But the Soviet line held fast. The German drive on Moscow had been halted.[4]

Every historian's account of Operation Barbarossa mentions the Russian winter as a significant factor in the German failure to occupy Moscow at the end of 1941. However, the weather affected both sides. The huge Soviet casualties and the rapid advance of the German forces in the summer months tend to overshadow the extent to which the Germans casualty rates were also huge, and Germany was not able to absorb these losses without moving troops from some other front. The Russians, by contrast were able to bring fresh divisions from Siberia along with newly developed rocket launchers and T-34 Tanks (Clodfelter 2002). Stalin's emphasis on developing the Soviet military-industrial complex in the late 1930s was paying huge dividends. In 1941, the Russians were already out-producing the German economy in terms of tanks (Mawdsley 2009).

[4]The account of the German invasion of Russia in 1941 draws on: Nagorski 2007; Weinberg 1995, Chap. 5, 1994, Stone 2015; Citino 1987, Chap. 1, Clodfelter 2002; Glantz 2012.

What all this meant was that the momentum of the battle for Moscow had shifted dramatically in favor of the Soviets. On December 5 and 6, the Russians launched a major counteroffensive that pushed the *Wehrmacht* back to their November positions. For the moment, at least, the Soviet Capital had been saved, and on December 8, Hitler issued a directive ordering his overextended units hold their positions at all costs. Looking ahead to the spring, Hitler and his generals were confronted with a difficult choice. Should they reinforce the troops in front of Moscow to make a final effort to take the city, or should they redirect their efforts towards the successes they had enjoyed in the South.

Either option involved another risky gamble, but they had to do something.

Tora, Tora, Tora: The Japanese Attack on Pearl Harbor

While the Germans were struggling to capture Moscow, the Japanese had their ambitions focused on Southeast Asia. Following the outbreak of the war in China in 1937, relations between the United States and Japan grew steadily worse. Isoroku Yamamoto, the commander in chief of the IJN, was working on a plan to protect the Japanese forces in Southeast Asia from the American intervention. The existing plan was that the IJN would sit behind a screen of islands in the Western Pacific and lure the Americans into a huge defensive battle. Yamamoto argued that this would not protect the forces in Southeast Asia. "The presence of the US fleet in Hawaii," he pointed out, "is a dagger pointed at our throats. Should war be declared the length and breadth of our southern operation would immediately be exposed to serious threat orders point" (Potter 1967, 84). In August 1940, when Yamamoto presented his plan to have the IJN attack the American fleet at Pearl Harbor to the Naval High Command, the idea was not well-received. However, Yamamoto's arguments that the IJN must take the offensive received a boost from the course of diplomatic and economic events. Back in Washington, Secretary of State Cordell Hull and President Roosevelt continued to tighten the embargo against American exports to Japan. Dwindling supplies of gasoline and oil were becoming critical, and in early November, Hideki Tojo presented Yamamoto's plan to the Emperor for the first time. On November 5, before the Emperor had even heard of the plan, Yamamoto was making preparations for the "Combined Fleet" (as the Pearl Harbor task force was called) to rendezvous at Tankan Bay, a remote spot in the Kurile Islands, to refuel and prepare to launch an attack on Hawaii. By the time Hirohito had signed the order authorizing the attack, the IJN carrier force had already set sail for its target (Kuehn 2015; Prange 1981, Part I).

It was an impressive armada: a total of 31 ships in all, including six carriers, two battleships, two heavy cruisers, and nine of the navy's newest destroyers. What the Japanese Navy had created with their combined fleet task force was equivalent to what the Germans had created with their Panzer Corps on land: a naval strike force with offensive power that was unmatched by any fleet in the world. While diplomats on both sides debated with each other in an effort to gain their concessions, the carriers sped through the North Pacific. Only a handful of men in the IJN knew the

full scope of the plan; even Prime Minister Tojo had been briefed on only the most general aspects of the plan. At 6 o'clock on the morning of December 7, 1941, 183 aircraft took off from the Japanese carriers situated 230 miles north of Oahu and headed for Pearl Harbor, guided by the signal of a radio station in Honolulu. In the space of 2 h, Pearl Harbor was turned into a burning inferno. The Japanese lost 55 airmen in the attack; the Americans lost 10 times that number of personnel along with 8 battleships; 3 cruisers, and 3 destroyers either sunk or damaged, and more than half the 231 planes stationed at the base damaged beyond repair.

The Pacific War between the United States and Japan had begun. While it has generally been regarded as a masterpiece of military planning and a skillful tactical blow against the Americans, "Pearl Harbor," as John Keegan noted, "was no Trafalgar" (Keegan 1989, 255). It was certainly a "wake-up" call that brought the Americans into a war they otherwise seemed determined to avoid – and in which Japan was not likely to win once it was started. The most serious shortcoming of the raid was that none of the American aircraft carriers were docked at Pearl Harbor – which meant that the US navy still had the potential to disrupt Japanese plans in Southeast Asia. Yamamoto would have to come up with another plan to destroy the American Fleet (Prange 1981).

The End of the Beginning: Midway and Stalingrad

Yamamoto's Pearl Harbor gamble did pay significant dividends for the Japanese Army's efforts in Southeast Asia. By the end of April 1942, the Japanese had effectively taken over the British, Dutch, and American colonies in Southeast Asia. All of this was accomplished without the loss of a single fleet carrier, battleship, or heavy cruiser. Despite the shortcomings of the Pearl Harbor raid, the attack on Pearl Harbor meant that the American Fleet was not in a position to quickly interfere with Japanese moves in Southeast Asia. Japan's control now extended as far south as Australia, eastward across Northern New Guinea to the Solomon Islands.

The Japanese high command had several options at this point. One possibility would be to return to the original plan put forward by the IJN in 1940 and build a powerful defensive perimeter to the east of the newly conquered areas in Southeast Asia. As they had during the arguments over Pearl Harbor, proponents of this strategy favored letting the American fleet come west so that the Japanese fleet could engage the enemy in home waters. Once again, Yamamoto's challenge was to convince his peers that the IJN could not sit still and wait for the Americans to come after them. Pearl Harbor had seemingly taken America out of picture, and the naval operations in the waters of Southeast Asia and the Indian Ocean had neutralized the fleets of the other imperial powers. The successes of the Japanese Army in February and March of 1942 meant that the land operations in China had reached a point where further advances would not yield major gains. This meant that the Army was looking to the South for its next move – a proposal that the Navy was quite content to go along with, since the area just north of Australia and eastward along the Solomon Islands was at that moment the weakest link in the defensive perimeter.

The plan that eventually emerged from these considerations was a joint naval and army operation whose principal objective was the capture of Port Moresby, New Guinea, and the fortification of strongholds in the Solomons The Americans hoped to block the Japanese threat to Port Moresby. The result of this action was the battle of the Coral Sea, a naval battle in which the ships engaged were never within sight of each other because the attacks were carried out entirely by aircraft. On paper, it would appear that the Japanese had the better of it at the Battle of Coral Sea. They had one fleet carrier damaged and lost an escort carrier. The Americans lost their biggest carrier and had another carrier seriously damaged. Yet most historians have concluded that it was the Americans who gained the greatest strategic advantages from the battle. The Japanese called off the invasion of Port Moresby, which disrupted their plans to reinforce the defenses in the Southern area of their newly created empire.

The Japanese decision to retreat represented the first significant check to the Japanese army in the south pacific. Although the Japanese still maintained the initiative in the Pacific war, the Battle of the Coral Sea was the first small step leading to the Americans taking the initiative. Yamamoto wanted to draw the Americans into a major fleet action where the superiority of the IJN carriers would overwhelm and destroy the American carriers. The IJN assembled a task force, which included four of the six aircraft carriers in the Japanese Navy. Unbeknownst to the Japanese, American code breakers had succeeded in breaking the Japanese naval codes to the point where they could read virtually all the Japanese messages. With the element of surprise on his side, USN admiral Chester Nimitz committed all three of his fleet carriers to attack the Japanese fleet approaching Wake Island, Nimitz' plan was simple; the Americans would position their forces north and slightly east of Midway and wait for the Japanese to arrive. It was Yamamoto and the IJN, not the USN, who were sailing into a trap.

The most significant sea battle of the Second World War began just as the sun rose on the morning of June 5, and it had run its course by the time the sun had set that evening. Both sides were caught up in a desperate combat in which neither side could see what the other was doing – and indeed for most of the battle, they did not even know where their foes were. Each side sent planes to find and destroy the enemy, with limited success. Then in a stroke of luck, two squadrons of dive bombers from *USS Yorktown* found the Japanese fleet in the process of preparing planes for an attack on Midway. In less than 15 min, the American planes managed to destroy one half of the enemy's aircraft carriers together with a large share of its airplanes. By the morning of June 6 the full extent of the carnage was apparent to both sides. The Japanese had lost all four of their fleet carriers together with more than 250 planes. They had lost over 3,000 men – including 250 of the elite *Kidō Butai* pilot corps. The Americans lost 1 fleet carrier, 100 planes, and 300 men killed in action (Parshall and Tully 1962; Thomas 2006).

The Battle of Midway was one of the most decisive victories in naval history. The Japanese lost four aircraft carriers that could not be replaced. Even under the best of circumstances, any new ships would not be ready in time to counter the growing might of the USN. In contrast to this bleak situation facing the IJN, the United States

already had six new Essex-class fleet carriers under construction at the time of the battle of Midway – four of which would be in service by the end of 1942. Until Midway, the Japanese had been able to trade carrier losses with the USN without losing parity in the number of fleet carriers available for action. After Midway, the United States lost only two carriers from enemy action for the rest of the war. There would be more naval battles in the coming months, but the eventual outcome of the naval war in the Pacific was no longer in doubt. The ability of the Americans to continue to expand their fleet while systematically destroying IJN forces led to the eventual defeat of the Japanese forces. The strength of the American economy would determine the outcome of the war.

While the Japanese struggled to hold on to the southern limits of their empire in Asia, the Germans were considering their options in Russia. As they assessed their situation, Hitler decided not to continue the battle around Moscow. A new strategy, *Operation Blue*, had German forces push southward towards the Oilfields at Maikop and Stavropol and then move south along the Donetz Corridor toward Stalingrad. The economic arguments behind Operation Blue made a great deal of sense. Both the German and the Soviet economies needed the oil from the Caucuses. It was, however, a very ambitious plan considering the circumstances confronting German troops in the Soviet Union at the beginning of 1942. The *Wehrmacht* was an understaffed, banged-up army that was only a shell of the finely tuned machine that had raced into Russia the previous summer. Bolstered by more than 50 divisions of Italian, Hungarian, and Romanian units from the Adriatic, the German forces pushed their territorial gains eastward to the Don River and as far south as Stalingrad and the Caucasus Mountains. By the end of summer 1942, the limits of German territory in the Soviet Union ran in an irregular arc from the outskirts of Leningrad southward to a front just west of Moscow, south to Voronozeh, and the west bank of the Don at Stalingrad. It looked as if Hitler might still be able win his great gamble of June 1941. If his armies could anchor the southern end of the eastern from by capturing the city of Stalingrad, Hitler believed he could complete the conquest of the Soviet Union by finishing the attacks on Leningrad and Moscow. (Weinberg 2015).

General Friedrich Von Paulus' sixth army reached the city limits of Stalingrad at the end of August 1942. Slowly but surely the Germans forced themselves further into the city until, on November 11, von Paulus mounted a final assault that actually got to the river. That would be as far as they would get. The Germans were exhausted, and the Russians had finally regrouped enough to mount a massive counterattack. On November 23, the Soviet armies linked up west of the city and more than a quarter of a million German soldiers were trapped in the resulting "pocket." On February 2, 1943, von Paulus surrendered his troops.

Stalingrad was a huge victory for the Soviets. The cost of the battle for both sides totaled more than 750,000 men, including prisoners. In addition to the losses inflicted on the Germans, the surrender of an entire army was a huge psychological blow to the Nazis and had an equally large positive effect on Soviet morale. In addition to relieving the pressure on Stalingrad, the Russian offensive forced the Germans to abandon the Caucuses – which had initially been the main focus of the summer offensives.

The tide of war had inexorably turned against the Axis Powers.

The Economics of the Second World War

If there was one lesson that generals, politicians, and planners learned from the experience of the First World War, it was that waging war was an economic as well as a military challenge. Economic historians have compiled an abundance of economic data documenting the economic performance of the major powers. A cliometric picture of the war closely mirrors the military events outlined above. Figures 11 and 12 present data on gross domestic product and military manpower of the major powers during the war. The data show that for the first 2 years of the war, the military successes enjoyed by the Axis Powers provided economic resources that substantially exceeded those of the Allied Powers. However, when

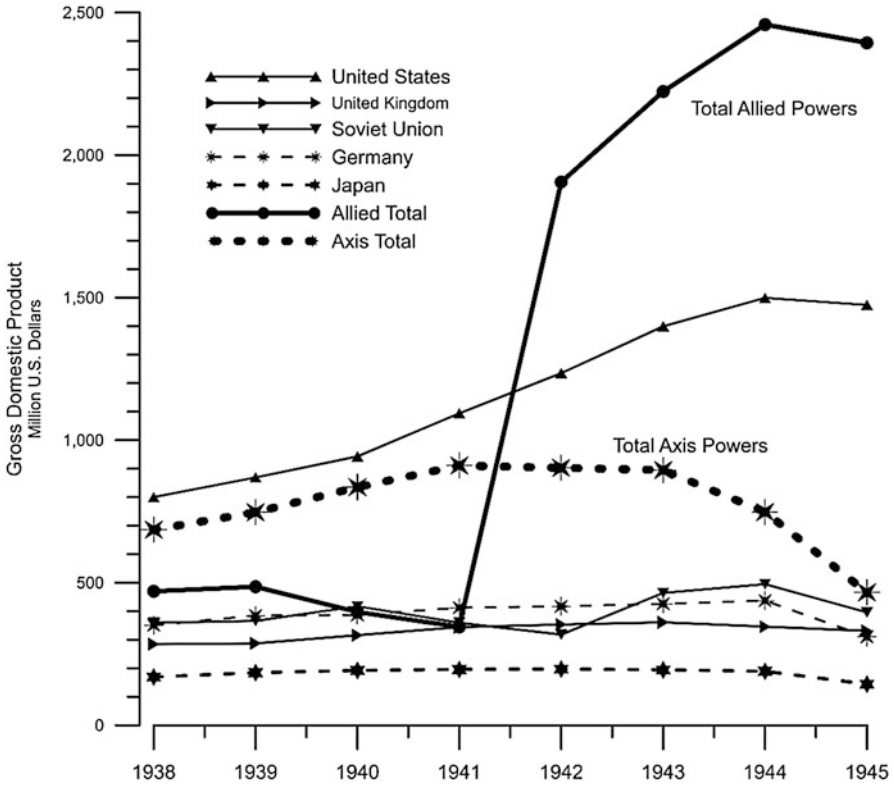


Fig. 11 Gross domestic product of the major powers, 1938–1945. (Sources: USA (Rockoff 1998, 86, Table 3.2); UK (Broadberry and Howlett 1998, 51 Table 2.1); USSR (Harrison 1998b, 278, Table 7.5); Germany (Abelshauer 1998, 135, Table 4.12); and Japan (Hara 1998, 261, Table 6.13))

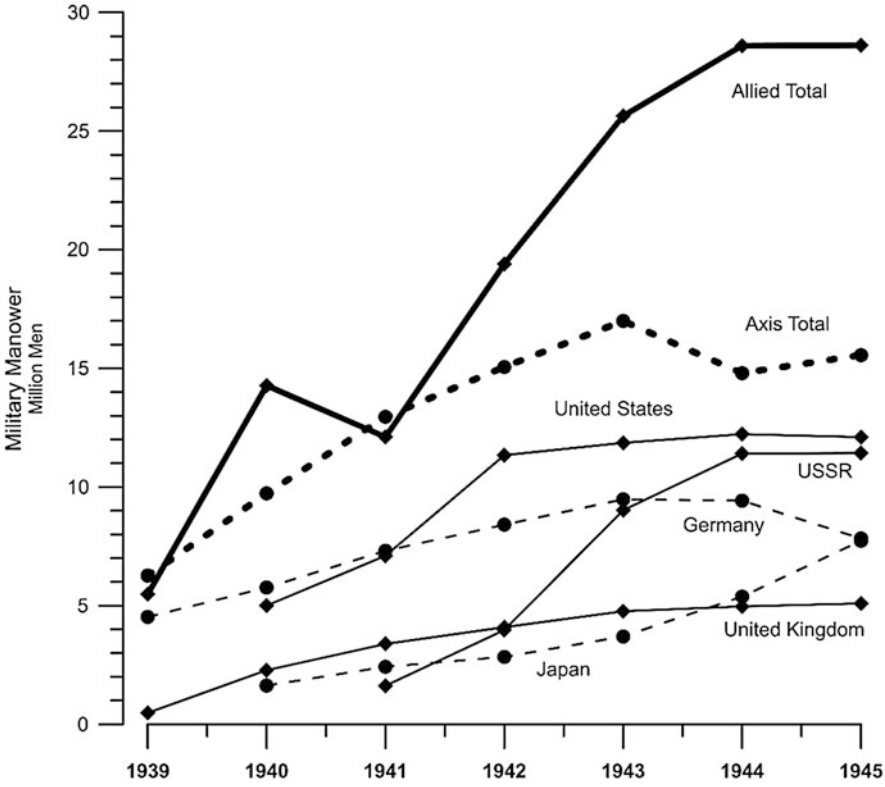


Fig. 12 Military manpower of the major powers, 1939–1945. (Sources: Correlates of War Project 2010)

the United States and the Soviet Union joined the fight in 1942, the economic balance of power shifted dramatically in favor of the Allies.

In 1918, the British were still trying to figure out how to best use their tanks. The British tanks were developed for trench warfare. They were large and unwieldy and easy targets for antitank guns. By 1940, the Germans had developed tanks that were much smaller and mobile. When combined with specially trained infantry units, these tanks comprised the backbone of the Panzer units that were the heart of their blitzkrieg strategy. Airplanes, which were also still in their infancy at the end of the Great War had evolved by 1940 into valuable weapons against both enemy troops and enemy populations. In a war where new weaponry involved thousands of armored vehicles and combat aircraft, the ability to produce these weapons played a significant role in the outcome of battles. Table 2 presents data on the production of these weapons for the United States, Soviet Union, United Kingdom, Germany, and Japan during the war. The figures reinforce the picture provided by the data in Figs. 10 and 11. As might be expected given their preparations for war in the late 1930s, the Germans and the Japanese had more

Table 2 Weapon Production of the Major Powers

Tanks and Mobile Artillery								
	USA	USSR	UK	Allies	Germany	Japan	Axis	Allied/Axis
1939			300	300	700	200	900	0.33
1940			1,400	1,400	2,200	1,000	3,200	0.44
1941	900	4,800	4,800	10,500	3,800	1,000	4,800	2.19
1942	27,000	24,400	8,600	60,000	6,200	1,200	7,400	8.11
1943	38,500	24,100	7,500	70,100	10,700	800	11,500	6.10
1944	20,500	29,000	4,600	54,100	18,300	400	18,700	2.89
1945	12,600	20,500	2,100	35,200	4,400	200	4,600	7.65
Total	**99,500**	**102,800**	**29,300**	**231,600**	**46,300**	**4,800**	**51,100**	**4.53**
Combat Aircraft								
	USA	USSR	UK	Allies	Germany	Japan	Axis	Allied/Axis
1939			1,300	1,300	2,300	700	3,000	0.43
1940			8,600	8,600	6,600	2,200	8,800	0.98
1941	1,400	8,200	13,200	22,800	8,400	3,200	11,600	1.97
1942	24,900	21,700	17,700	64,300	11,600	6,300	17,900	3.59
1943	54,100	29,900	21,200	105,200	19,300	13,400	32,700	3.22
1944	74,100	33,200	22,700	130,000	34,100	8,300	42,400	3.07
1945	37,500	19,100	9,900	66,500	7,200	8,300	15,500	4.29
Total	**192,000**	**112,100**	**94,600**	**398,700**	**89,500**	**55,100**	**144,600**	**2.76**
Naval Vessels								
	USA	USSR	UK	Allies	Germany	Japan	Axis	Allied/Axis
1939			57	57	15	21	36	1.58
1940			148	148	40	30	70	2.11
1941	544	62	236	842	196	49	245	3.44
1942	1,854	19	239	2,112	244	68	312	6.77
1943	2,654	13	224	2,891	270	122	392	7.38
1944	2,247	23	188	2,458	189	248	437	5.62
1945	1,513	11	64	1,588		51	51	31.14
Total	**8,812**	**161**	**1,156**	**10,129**	**954**	**589**	**1,543**	**6.56**

tanks and aircraft than the Allied forces at the outbreak of hostilities in 1939. However, the ability of the Americans and Russians to produce these weapons soon turned the balance of military power in favor of the allies. Over the course of the war, the United States, the Soviet Union, and the United Kingdom produced almost five times as many tanks and three times as many combat aircraft as Germany and Japan. In 1941, the Russians were already producing as many tanks and combat aircraft as the Germans. Because of the disorganization in the leadership of the Soviet forces after the purges of the 1930s, there was a tendency to assume that the Russians were a much weaker military power than the German invaders. That was clearly not the case. As Victor Davis Hansen points out,

> Germany revolutionized armored warfare, yet Russia built the most and best tanks while the British and Americans deployed the most effective ground-support fighter-bombers. So great were the Allied material advantages in armor and artillery that the superiority of German tank crews and Panzer generals ultimately was for naught. (Hanson 2017, 504)

Perhaps the greatest mistake Hitler and his generals made in 1941 was their failure to recognize the economic might of the Soviet Union.

We have seen how the Germans tried unsuccessfully to "starve" the British into submission through a policy of unrestricted submarine warfare as part of their economic war against the Allies in 1917–1918. In 1939, they once again waged an economic war against Allied ships in an effort to curb imports into the British Isles. Figure 13 presents data on Allied shipping sunk by the Germans from 1939 to 1945. This time they had the advantage of ports along the western French coast which extended the range of protection for the U-boats and allowed air attacks by planes and surface raiders. The results were impressive for the first 3 years. By 1942, losses from submarines reached 6.2 million tons of shipping a year and airplanes accounted for another 700,000. Losses for the duration of the war totaled more than 14 million tons of shipping, which is an impressive total, but it was not enough to seriously threaten the flow of men and materials from the United States to Great Britain. The ability of the United States to build "liberty ships" largely canceled the losses of tonnage sunk by German U-Boats. Davis and Engerman argue that "the Allied navies and air forces had effectively broken the blockade before the end of 1943." (Davis and Engerman 2006, 287).

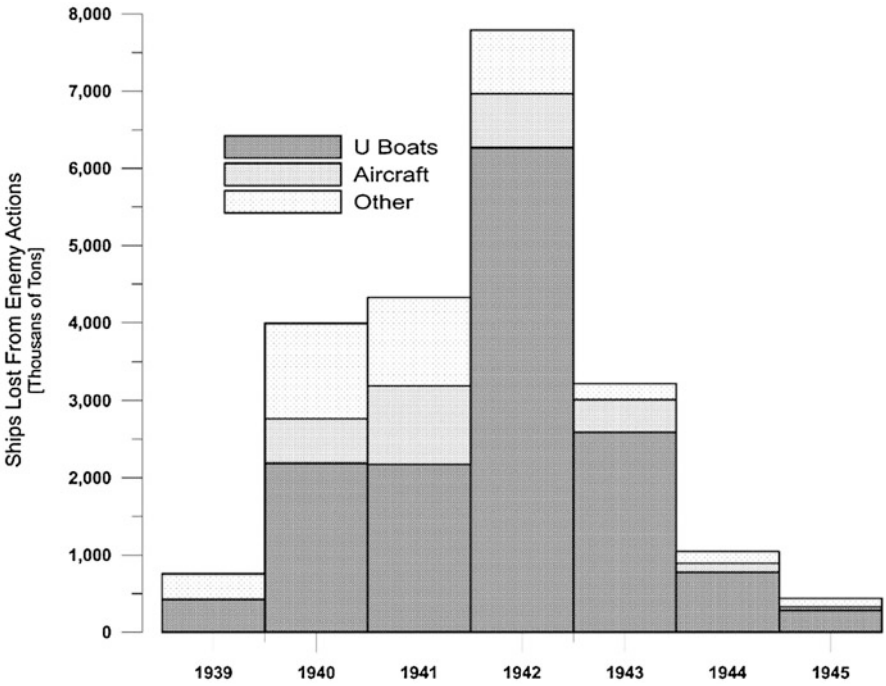


Fig. 13 Allied ships lost to enemy action, 1939–1945. (Sources: (Davis and Engerman 2006, 288, Table 6.3) "Other" includes mines and armed surface vessels acting as raiders and "unknown causes.")

While the German blockade in the Atlantic was ultimately a failure, the American naval blockade and air attacks against the Japanese homeland was a stunning success. "The war against shipping" note Davis and Engerman, "was perhaps the most decisive single factor in the collapse of the Japanese economy and logistic support of military and naval power." The impact of the blockade can be seen in the 50% fall in the consumption of staple goods over the course of the war (Davis and Engerman 2006, 378). The naval blockade was accompanied by massive air attacks made possible by the development of the B-29 bomber. According to Vincent Hanson:

> The bombers may have damaged Japanese industry as much by shutting down its transportation, ports, docks, and factory supplies as by the firebombing of industrial plants. Over 650,000 tons of Japanese merchant shipping was destroyed and another 1.5 million tons rendered useless, given the inaccessibility of ports and Allied control of the air and sea. (Hanson 2017, 114)

These attacks eventually culminated with the decision to drop atomic bombs on Hiroshima and Nagasaki on August 6 and 9, effectively ending the Pacific War with Japan. Whether or not the use of the atomic bomb was necessary, given the damages to Japanese cities already victimized by firebombing, is a matter of some debate. What is clear is that the ability of the American war effort to spend billions of dollars to develop super weapons such as the atomic bomb and the B-29 bomber provides additional support for the argument that the role of economic efforts was a pivotal factor in the outcome of the war.

Going "All In": Gambling on War in an Age of Catastrophe

A basic conclusion to emerge from this essay is that by the beginning of the twentieth century, war and economics were inexorably joined together so that economic strength was as important as military might in determining the outcome of a major war. Victories on the battlefield were a *necessary* condition to winning a major war, but they are not a *sufficient* condition to guarantee victory. In December 1914, Erich von Falkenhayn, the commander-in-chief of the German Army went to the Kaiser and urged him to negotiate a peace treaty with the Entente states following the First Battle of the Marne. The Schlieffen plan, he explained, had come up short of its objective, and there was no way that Germany would be able to win a protracted war of attrition. (Ritschl 2005). Our cliometric narrative supports Falkenhayn's early assessment of the large picture of the war. In the spring of 1918, the German army managed once again to get to within 40 miles of Paris, but the inability of the German economy to support both the troops in the field and the population on the home front meant that the Germans were unable to turn their battlefield successes into a final victory in the First World War. In 1942, Operation Barbarossa brought the *Wehrmacht* within a few miles of Moscow, but that proved to be the high watermark of their offensive capabilities and they were eventually forced to abandon their attack

on the Soviet capital in the face of Soviet counter attacks. The German defeat at Stalingrad ended any hopes the Nazis had of winning the war. In both world wars, the Germans were unable to sustain the level of economic activity necessary to support their military prowess. A similar argument can be made for the Japanese effort against American forces at Pearl Harbor and Wake Island. The military defeat of the IJN at Midway was an economic as well as a military disaster because the losses of men and ships could not be replaced.

While cliometric research provides a guide to the larger picture of why the Allies prevailed in the two world wars, the reasoning behind the decisions by leaders to persistently choose risky gambles on war over other policy options remains something of an enigma. Economists and cliometricians are used to dealing with outcomes based on "rational" thinking on the part of participants. However, in choosing to gamble on war, rational thinking was seemingly trumped by what John Maynard Keynes called "animal spirits" – decisions that were based on an overconfidence on the part of generals that their plans for the conduct of the war would succeed, combined with a fear of losing the war if they did not "do something."[5] Generals, in other words, displayed a strong propensity to gamble on fighting a battle – even if the outcome is very risky – rather than accepting what they viewed as a gamble on a peaceful settlement might have a lower risk of failure. The tendency of countries to choose war over peace was reinforced by the propensity of political leaders to leave the final decisions with regard to military strategy to generals.

Conclusion

Cliometricians have tended to avoid the question of what "causes" wars and follow the example of Ronald Findlay and Kevin O'Rourke, two cliometricians who reluctantly admit that the First World War "appears as somewhat of a *diabolus ex machina*" in their analysis of the causes of the war (Findlay and O'Rourke 2007, xxv).[6] For cliometricians, wars are exogenous events that must be built into the assumptions of the theoretic model. While such models do not explain the causes of wars, they have produced an impressive body of empirical data that allows us to examine the economic consequences of war in an Age of Catastrophes.

[5]Keynes' use of the term "animal spirits" was applied to his analysis of investor speculation during the stock market boom leading to the Great Depression (Keynes 1936, 129). For more on the use of animal spirits to explain decisions involving war and economics, see Ransom 2018, Chap. 1, 2016.

[6]They go on to explain that "There is of course no shortage of authorities who have argued that the way in which the late-nineteenth-century world economy operated helps explain the eruption of World War I, but the causes of this disaster remain controversial." (Findlay and O'Rourke 2007, xxv).

References

Abelshauer W (1998) Germany: guns, butter and economic miracles. In: Harrison M (ed) The economics of World War II: six great powers in international comparison. Cambridge University Press, Cambridge, pp 122–176

Ball A (1987) Russia's last capitalists: the NEPmen 1921–1929. University of California Press, Berkley and Los Angeles

Boemeke M, Feldman G, Glaser E (eds) (1998) The treaty of Versailles: a reassessment after 75 years. Cambridge University Press, New York

Broadberry S, Harrison M (eds) (2005) The economics of World War I. Cambridge University Press, New York

Broadberry S, Howlett P (1998) The United Kingdom, 'Victory at all costs. In: Harrison M (ed) The economics of World War II: six great powers in international comparison. Cambridge University Press, Cambridge, pp 43–72

Citino R (1987) The German way of making War: from the thirty years' War to the third reich. University Press of Kansas, Lawerence

Clodfelter M (2002) Warfare and armed conflicts: a statistical reference to casualty and other figures, 1500–2000, 2nd edn. McFarland & Company, Inc, Jefferson

Correlates of War Project (2010) National Material Capabilities Dataset (V 5.0). http://www.correlatesofwar.org/data-sets/national-material-capabilities

Crafts N, Fearon P (2013a) Depression and recovery in the 1930s: an overview. In: Crafts N, Fearon P (eds) The Great depression of the 1930s. Oxford University Press, Oxford, UK, pp 1–44

Crafts N, Fearon P (eds) (2013b) The great depression of the 1930s: lessons for today. Oxford University Press, Oxford, UK

Davies RW (1980) The socialist offensive: the collectivization of soviet agriculture. Vol. 1, The industrialisation of Soviet Russia. Cambridge University Press, Cambridge

Davies RW (1989) The Soviet economy in turmoil. Vol. 4, The industrialisation of Soviet Russia. Cambridge University Press, Cambridge

Davies RW, Wheatcroft SG (2004) The years of hunger: soviet agriculture, 1931–36. Vol. 5, The Industrialisation of Soviet Russia. Cambridge University Press, Cambridge

Davies RW, Harrison M, Khlevniuk O, Wheatcroft SG (2018) The soviet economy: The late 1930's in historical perspective. Working paper series, #363. Warrick

Davis LE, Engerman SL (2006) Naval blockades in peace and war: an economic history since 1750. Cambridge University Press, New York

Diebolt C, Haupert M (2018) We are all Ninjas: how economic history has infiltrated the economics discipline, with Claude Diebolt, *forthcoming Sartoniana vol 32, 2019*

Eichengreen B (1991) The origins and nature of the great slump, revisited. Working paper. University of California, Berkeley Department of economics, Berkeley

Eichengreen BJ (1992) Golden fetters: the gold standard and the great depression, 1919–1939. Oxford University Press, New York

Eichengreen B, Hatton T (1988) Interwar unemployment in international perspective: an overview. Kluwer Academic Publishers, London

Eichengreen B, Temin P (1997) The gold standard and the great depression. Working paper 6060. National Bureau of Economic Research, Cambridge, MA

Feinstein C, Temin P, Toniolo G (2008) The world economy between the world wars. Oxford University Press, New York

Feldman G (1997) The great disorder: politics, economics, and Society in the German Inflation, 1914–1924. Oxford University Press, New York

Ferguson N (1999) The pity of war: explaining World War I. Basic Books, New York

Ferguson N (2006) The war of the world: twentieth century conflict and the descent of the west. The Penguin Press, New York

Ferris J, Mawdsley E (2015a) Introduction to Part I. In: Ferris J, Mawdsley E (eds) Fighting the war. Cambridge University Press, Cambridge, pp 21–27

Ferris J, Mawdsley E (2015b) The war in the west, 1939–1940: the Battle of Britain? In: Ferris J, Mawdsley E (eds) Fighting the war. Cambridge University Press, Cambridge, pp 315–335
Findlay R, O'Rourke KH (2007) Power and plenty: trade, war, and the world economy in the second millennium. Princeton University Press, Princeton
Fischer KB (1995) Nazi Germany: a new history. Continuum Publishing Company, New York
Fitzpatrick S, Rabinowitch A, Stites R (eds) (1991) Russia in the era of NEP: explorations in soviet society and culture. Indiana University Press, Bloomington/Indianapolis
Frey M (2000) Bullying the neutrals: the case of the Netherlands. In: Chickering R, Forster S (eds) Great war, Total war: combat and mobilization on the Western front, 1914–1918. Cambridge University Press, New York, pp 247–264
Frieser KH (2015) The war in the west, 1940: an unplanned blitzkrieg. In: Ferris J, Mawdsley E (eds) Fighting the war. Cambridge University Press, Cambridge, pp 287–314
Glantz D (2012) Barbarossa derailed: the Battle for Smolensk 10 July–10 September 1941. Helion & Company, Solihull
Hanson VD (2017) The second World Wars: how the first global conflict was fought and won, Kindle edn. Basic Books, New York
Hara A (1998) Japan: guns before rice. In: Harrison M (ed) The economics of World War II: six great powers in international comparison. Cambridge University Press, Cambridge, pp 224–267
Hardach G (1977) The First World War, 1914–1918. University of California Press, Berkeley
Harrison M (1998a) The Soviet Union: the defeated victor. In: Harrison M (ed) The economics of World War II: six great powers in international comparison. Cambridge University Press, Cambridge, pp 268–296
Harrison M (ed) (1998b) The economics of World War II: six great powers in international comparison. Cambridge University Press, New York
Henig R (1995) Versailles and after, 1919–1933, 2nd edn. Routledge, New York
Herwig H (1997) The first world war: Germany and Austria-Hungary, 1914–1918. Oxford University Press, New York
Hitler A (1936) My Battle. Paternoster Library, London
Hitler A (1941) Letter to Benito Mussolini, June 21 1941. https://en.wikisource.org/wiki/Adolf_Hitler%27s_Letter_to_Benito_Mussolini
Hitler A (1961) Hitler's secret book (trans: Attanasio S). Grove Press, New York
Hobsbawm E (1994) The age of extremes: a history of the World, 1914–1990. Pantheon, New York
Jukes G (2002) In: O'Niell R (ed) The Russo-Japanese War, 1904–1905. Essential histories, vol 31. Osprey Publishing, Oxford
Keegan J (1989) The second World War. Viking, New York
Keynes JM (1920) The Economic consequences of the peace. Penguin Books, New York. Original edition, 1920. Reprint, 1988
Keynes JM (1936) The general theory of employment, interest, and money. Harcourt Brace and World, New York
Kindleberger CP (1978) Manias, panics and crashes: a history of financial crises, 1st edn. Basic Books, New York. Original edition, 1975
Kindleberger C, Aliber R (2005) Manias, panics and crashes: a history of financial crises, 5th edn. Wiley, Hoboken
Kotkin S (2014) Stalin: Paradoxes of Power, 1878–1928, Kindle edn. Penguin, New York
Kuehn JT (2015) The war in the pacific, 1941–1945. In: Ferris J, Mawdsley E (eds) Fighting the war. Cambridge University Press, Cambridge, pp 420–454
MacMillan M (2001) Paris 1919: six months that changed the world. Random House, New York
Maddison A (1985) Two crises: Latin America and Asia, 1929–38 and 1973–83. OECD, Paris
Maddison A (2006) The World economy. 2 vols. Vol. 2: Historical statistics. OECD, Paris
Maddison A (2010) Historical statistics of the world economy 1–2000. www.ggdc.net/maddison
Marks S (2013) Mistakes and myths: the allies, Germans and the versailles treaty, 1918–1921. J Mod Hist 85(2):632
Mawdsley E (2009) World War II: a new history. Cambridge University Press, Cambridge

Miller ES (2007) Bankrupting the enemy: the U.S. Financial Siege of Japan before Pearl Harbor. Naval Institute Press, Annapolis
Minsky H (1982) The financial instability hypothesis: capitalist processes and the behavior of the economy. In: Kindleberger C (ed) Financial crises: theory, history, and policy. Cambridge University Press, New York, pp 13–47
Mitchell BR (1998) European historical statistics, 1750–1993, 4th edn. Stockton Press, New York
Moseley R (2006) The last days of mussolini. Sutton Publishing, Gloucestershire
Nagorski A (2007) The greatest Battle: Stalin, Hitler and the desperate struggle for Moscow that changed the course of World War II. Simon and Schuster, New York
Nations, League of (1941) Statistical yearbook, 1940/1. Secretariat of the League of Nations, Geneva
Offer A (1989) The first world war: an agrarian interpretation. Oxford University Press, New York
Offer A (2000) The blockade of Germany and the strategy of starvation, 1914–1918: an agency perspective. Chap. 9. In: Chickering R, Forster S (eds) Great War, Total War: combat and Mobilization on the Western Front, 1914–1918. Cambridge University Press, New York, pp 169–188
Paine SCM (2003) The first Sino-Japanese War of 1894–1895: perceptions, power, and primacy. Cambridge University Press, Cambridge
Paine SCM (2017) The Japanese empire: grand strategy from the Meiji restoration to the Pacific war. Cambridge University Press, Cambridge
Parshall J, Tully A (1962) Shattered Sword: the untold story of the battle of midway. Potomac Books, Washington, DC
Potter JD (1967) Yamamoto: the man who menaced America. Paperback Library, New York
Prange GW (1981) At dawn we slept: the untold story of pearl harbor. McGraw-Hill Book Company, New York
Ransom RL (1981) Coping with capitalism: the economic transformation of the United States, 1776–1980. Prentice-Hall, Englewood Cliffs
Ransom RL (2018) Gambling on war: confidence, fear, and the tragedy of the first word war. Cambridge University Press, Cambridge
Reinhart CM, Rogoff KS (2009) This time is different: eight centuries of financial folly. Princeton University Press, Princeton
Ritschl A (2005) The pity of peace: Germany's economy at war, 1914–1918 and beyond. In: Broadberry S, Harrison M (eds) The economics of World War I. Cambridge University Press, New York, pp 71–76
Rockoff H (1998) The United States; from ploughshares to swords. In: Harrison M (ed) The economics of World War II: six great powers in international comparison. Cambridge University Press, Cambridge, pp 81–117
Service R (2006) Stalin: a biography, Kindle edn. Belnap Press, New York
Stone D (2015) Operations on the eastern front, 1941-45. In: Ferris J, Mawdsley E (eds) Fighting the war. Cambridge University Press, Cambridge, pp 331–356
Taylor J (2015) China's long war with Japan. In: Ferris J, Mawdsley E (eds) Fighting the war. Cambridge University Press, Cambridge, pp 51–77
Temin P (1981) Notes on the causes of the great depression. In: Brunner K (ed) The great depression revisited. Martinus-Nijhoff, Boston
Temin P (1989) Lessons from the great depression. The MIT Press, Cambridge, MA
Thomas E (2006) Sea of thunder: four commanders and the last great naval campaign, 1941–45. Simon and Schuster, New York
Van De Ven H (2015) Campaigns in China, 1937–1945. In: Ferris J, Mawdsley E (eds) Fighting the war. Cambridge University Press, Cambridge, pp 256–286
Weinberg GL (1995) Germany, Hitler and World War II. Cambridge University Press, New York
Weinberg G (2015) German strategy, 1939–45. In: Ferris J, Mawdsley E (eds) Fighting the war. Cambridge University Press, Cambridge, pp 107–131
White J (1994) The Russian revolution: a short history. Edward Arnold, New York

Wolf N (2013) Europe's great depression: coordination failure after the first world war. In: Crafts N, Fearon P (eds) The great depression of the 1930s. Oxford University Press, Oxford, UK, pp 74–109

Zamagni V (1998) Italy: how to lose a war and win the peace. In: Harrison M (ed) The economics of World War II: six great powers in international comparison. Cambridge University Press, Cambridge, pp 177–224

Part VII

Innovation, Transportation, and Travel

Innovation in Historical Perspective

Stanley L. Engerman and Nathan Rosenberg

Contents

Abstract

It is necessary to study the historical record concerning the economic nature of technological change, the constraints it confronts, and the complementarities with other sectors of the economy in order to fully understand the nature of innovation. Consideration must be given to the market environment, the available production facilities, the existing body of knowledge, and the social and organizational contexts of the innovation, in addition to the series of required changes within other sectors, not just to the limited aspects of a narrowly-defined specific innovation. Since theoretical models cannot deal with the full complexity of the process of invention, innovation, and the utilization of new devices, some historical study is required to develop a full understanding of these processes. Without consideration of past events, it is difficult to understand either the present

S. L. Engerman (✉)
Department of Economics, University of Rochester, Rochester, NY, USA
e-mail: s.engerman@rochester.edu

N. Rosenberg
Department of Economics, Stanford University, Emeritus, Stanford, CA, USA

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_26

or the future. Consideration of these factors will not only increase our historical knowledge but also serve to enrich our theorizing about these questions.

Introduction

In a conversation with Nathan Rosenberg on the topic of innovation, Kenneth Arrow pointed out (to paraphrase) that theoretical models do not provide a complete depiction of the process of innovation, in part because of the impossibility of having "a theory of the unexpected."[1] Such theoretical modeling has tended to be unsuccessful both in providing guides to understanding the past and in pointing to future changes. The implication is that it is necessary to study the historical record concerning the economic nature of technological change, the constraints it confronts, and the complementarities with other sectors of the economy to fully understand the nature of innovation. Consideration must be given to the market environment, the available production facilities, the existing body of knowledge, and the social and organizational contexts of the innovation, in addition to the series of required changes within other sectors, not just to the limited aspects of a narrowly defined specific innovation. These points will be discussed in various sections in this chapter. In short, since theoretical models cannot deal with the full complexity of the process of invention, innovation, and the utilization of new devices, some historical study is required to develop a full understanding of these processes. Also important is the role of the historical background in influencing economic and technological developments, what some refer to as path dependence (or, suggesting a less certain set of outcomes, path influenced), but where "history matters" (Rosenberg 1994, pp. 9–23). Without consideration of past events, it is difficult to understand either the present or the future.

A related set of points about the nature of innovations were made earlier by Simon Kuznets, in two articles published in the 1970s (1973, 1979). Kuznets described several important aspects of the nature of innovations and the difficulties in evaluating their effects. First, there is the great initial uncertainty concerning the complete set of the ultimate effects of any one innovation. Second, there is the great importance of complementary positive adjustments – technologically, ideologically, and organizationally (including social and legal institutions) – before the full effects (positive and negative) of an innovation can be determined. These concerns mean that it will often take a long time before all the invention's impacts can be represented as "a major transformation of their pattern of living" (1973, p. 199), as well as adequate time to adapt to the dislocations affecting productive labor and other resources used in production and the other social difficulties caused by the

[1]Neither Nate nor Stan can find a published source for this claim. Arrow himself is not sure if, and where, it appears in print. The quote is from Arrow (2012, p. 43). It might be noted that this difference between theoretical models and historical complexity applies generally to all theoretical models.

introduction of new innovations (1973, pp. 202–208). Most innovations do present negative effects and cause reductions in welfare. Mokyr (2014), citing Tenner (1996), points to examples of innovations providing positive benefits but with offsetting costs such as DDT, sugar beets, lead for paint and gasoline, and asbestos, while Kuznets (1973, pp. 205–208) points to their impact on the environment and the increase in pollution. Some of these difficulties such as possible deterioration of the natural environment can, once recognized, and with appropriate political and technological developments, be overcome. In some cases, these can be accomplished by appropriate use of price incentives, but in some cases, it may require government-introduced regulatory policy. This, however, can be a lengthy and expensive process and may offset only some part of the difficulties. Kuznets did believe in the long-run net beneficial outcome of the cumulative process of innovation, demonstrated in his brief comparison of what the world of 1960 would have looked like if innovation had actually ceased one century earlier, particularly in regard to consumer goods (1973, pp. 189–190; 1979, pp. 66–69).

The Role of "Learning-by-Using"

This chapter is intended to draw more attention to certain aspects of the historical study of technological change and to the contribution of economic history to its theoretical analysis. It will first give attention to the background to certain innovations and the initial expectations of what benefits they could provide. Then it will discuss a number of reasons for what is often regarded as the relatively slow impact on measured total factor productivity and then describe why innovations often have significantly greater impacts on the economy than just in those sectors in which the innovation occurred. Given his major contributions to the study of these issues, we will draw heavily upon the published works of Nathan Rosenberg, but we shall extend several of his points and arguments.

The difficulties in "predicting and preparing for" specific innovations and preparing for all the effects of any innovation have been well illuminated by Nathan Rosenberg (2010, pp. 153–173), in an essay entitled "Uncertainty and Technological Change," dealing with the differences between the initial expectations of inventors and the ultimate role played by their innovations. The initial expectations often reflected the very particular problem that the invention was trying to solve, and even the innovators were unable to anticipate the subsequent improvements and developments that would take place. Thus, the early development of the steam engine was concerned with providing a means to pump water out of flooded mines (Rosenberg 2010, pp. 164–165). The "first railroads were expected to serve only as feeders into the existing canal system or were to be constructed in places where the terrain had rendered canals inherently impractical" (Rosenberg 2010, pp. 162–164; MacGill 1917, p. 291). Alexander Graham Bell saw the telephone as being mainly "improvements in telegraphy," not as its replacement (Rosenberg 2010, p. 156). Marconi saw the major use of his wireless innovation as being an aid mainly to ships, for either ship-to-ship or ship-to-shore communication (Rosenberg 2010, p. 156). More

recently, some believed that the main function of the transistor was expected to be the development of better hearing aids for the deaf (Rosenberg 2010, p. 157). Obviously more examples of such varieties of incorrect expectations can be given, but it is clear that innovations made in response to particular needs, or for a specific purpose, will often turn out, when improved and more fully developed, to have much different and broader uses, with their contribution to economic change being much larger and often in an unexpected direction than earlier anticipated.

There are several related reasons for the underestimation of the full effect of an innovation. First, we often date the introduction of an innovation quite early in the development process, where it can best be described as "primitive" (Rosenberg 1994, p. 69). With use – what we can describe as "learning by using" (Rosenberg 1982, pp. 120–140) – and with further experimentation, the specific piece of hardware (the innovation proper) will be improved upon its initial state, making it more productive, and also may be seen to have further, often unexpected, uses, which add to the benefits from the initial innovation, benefits not anticipated when the innovation was introduced.

"Learning by using" is to be distinguished from the more familiar concept of "learning by doing" since the latter refers more directly to the gains in productivity in the production process due to repetitions in the process of production. "Learning by using" refers to the emergence of new problems which arise from the process of production of the new innovation which must be solved to permit its utilization, problems which cannot be known until production is begun and are generally unexpected. The importance of "learning by using" is that most innovations, when introduced, are at a rather early stage and therefore require some improvement. Often the ability and need to make improvements can only be known and accomplished after the new technology is introduced. At an early stage, neither theoretical nor empirical approaches to the analysis of the new technology could anticipate many of the problems that will arise in the production process, and it is only by observation of the actual process that the problems are revealed and the basic information needed to make improvements known. Thus there may be a considerable time before the benefits are obtained.

"Learning by using" can account for a large part of overall productivity change. The impact of "learning by using," as well as the lag between the introduction of an innovation and its impact on measured total future production, however, lacks the dramatic appearance that comes from the study of the application of new scientific knowledge or the initial introduction of the new physical machinery. Further, these adjustments may, unlike the basic invention, not be patentable, leaving a less observable record. Yet the improvements made in the process of production are often crucial to making innovations productive and efficient. Given the inability of any model to describe all the possible operating eventualities, more information awaits the actual use of the innovation. "Learning by using" may be regarded as providing a joint product with the good produced, with elements of cost shared between the production of the good and the future benefits derived by the new knowledge, or else as a "free good" resulting from its production, an externality

resulting from the start of production, with all the costs attributed to the production of the good.

Important aspects of the learning process have been studied in several key articles. Jamasb (2007) and Stein (1997) present models that incorporate learning in the innovation process. Stein notes also the spillover of external benefits and costs to other firms, while Breschi et al. (2000) point to the possible differences in the nature and rate of innovation between new and old firms. Jovanovic and Lach (1989) point to the benefits that accrue to later entrants able to take advantage of what has been learned by earlier producers. Rantisi (2002) looks at learning as a function of the clustering of similar firms which provides for sharing of knowledge and practices.

Also important, as described in detail by Kuznets, is that to obtain the full set of benefits and to offset the costs of an innovation may take time, as there are often a variety of complementary adjustments, material and institutional, that must be made (Kuznets 1973, pp. 185–201; Kuznets 1979, pp. 56–99). Two particularly dramatic examples relate to the development of energy sources for the economy. The initial limited effect of the development of electricity upon the measured productivity of the economy was due, in large measure, to the need for technological and institutional changes to permit the widespread use of this innovation. To obtain more benefits in the manufacturing sector, it was necessary to redesign and reshape the factory floor, as well as to take advantage of the locational flexibility that had not previously been permitted to factories. The expanded use of electricity permitted new technologies in other sectors, such as metallurgy and steel production, benefits not immediately apparent when the basic innovation was introduced. Electricity has, of course, had a dramatic impact upon nonindustrial aspects of the economy, including transportation and the lighting of streets and houses, and has been the power source for many consumer goods (Mowery and Rosenberg 1998, pp. 105–109; Hughes 1983). To permit the widespread use of electricity by businesses and consumers required wiring, above and below ground, and this meant the increased ability of the state and/or the private sector to impinge on the property rights of individuals and businesses. While governments had long used the power of eminent domain, electrification required a considerably more extended use of this legal principle, dealing with many more individuals over larger areas, to be successful.

In the early twentieth century, petroleum was to become the important source of energy in the economy. Petroleum was not then a new product; the earliest major US discovery of oil had occurred in Pennsylvania in 1859 (Rosenberg 1982, pp. 185–186). Indeed, so uncertain was oil's future at this time that even as shrewd a businessman as Andrew Carnegie, contemplating the future prospects for oil, tried to corner the market since he expected that the United States would soon run out of oil (Sabin 1999). Fortunately for himself, Carnegie had a diversified investment portfolio. It took several decades before new discoveries of oil increased its supply, and significant increases in the demand for oil for use in various products, before the full impact, was achieved. This required, for example, the innovation and many successful refinements to the automobile with its internal combustion engine as well

as improvements in the airplane, both depending on oil for fuel (Mowery and Rosenberg 1998, pp. 47–70; Mowery and Rosenberg, in Rosenberg 1982, pp. 163–177). Vincenti (1990), is a detailed examination of the role of "learning by using" in airplane invention. To get the full benefits from these transportation developments, extensive expenditures by federal, state, and local governments, as well as by firms in the private sector, were necessary. The public sector assumed responsibility for building highways, roads, and bridges to permit private travel and the business movement of goods. For air travel, governments provided airports and traffic controls as well as safety regulations. For the automobile, the private sector provided the production and sale of automobiles and trucks, in both of which there were relatively rapid technical improvement in production, as well as a network of private stations to service autos and trucks as well as to make gasoline and oil available to needy customers. Various credit arrangements, such as installment credit as earlier pioneered by the Singer Sewing Machine Company, were also introduced to permit individuals and firms to afford the costs of purchasing cars and trucks.

These examples of what is required for all of the effects of an innovation to occur can be repeated for many other cases where developments after the initial introduction of an innovation were important, whether within the same sector as the innovation or elsewhere in the economy and whether they were innovations of hardware or of institutions. This latter point has been raised by Kuznets (1979, pp. 56–66) who states (p. 65), "It is the interplay of technological advance and organizational, economic, and social adjustments that the crucial feature of the innovation, the *application* of new technological element, lies." As Rosenberg (2010, p. 163) notes about the long time before electric power had a large impact on factory production that "such technological innovations commonly require significant organizational changes as well."

General Purpose Technology

A particular type of innovation that has a widespread set of uses and effects in several sectors of an economy has come to be called a general purpose technology (see Rosenberg and Trajtenberg, in Rosenberg 2010, pp. 97–135; Bresnahan and Trajtenberg 1995). These have been described as "a certain type of dramatic innovations" that "has the potential for pervasive use in a wide range of sectors that drastically change their modes of operation" (Helpman 1998, p. 3; see also Lipsey, et al. 2005). While often these were not expected to have such a wide range of uses when initially innovated, the general purpose technology invariably developed many new applications after its first adoption, which was intended for a specific purpose. The key examples discussed are the steam engine in the eighteenth and nineteenth centuries, the electric motor in the late nineteenth and early twentieth century, and the semiconductors, the laser, and the computer in the late twentieth century. To be fully effective as a general purpose technology, there must be a large range of complementary innovations as well as related changes in technology and organization in several different sectors of the economy.

The evolving nature of general purpose technologies is one explanation for the uncertainty of the full impact of new technologies, since the full set of the uses of an innovation often go far beyond its original intent. While the original incentive may be for an improvement aimed at one specific use, as new improvements take place, there are a wider range of different, unexpected uses in other sectors. One implication of this is that subsequent advances may take place in sectors other than the focus of the original innovation, posing issues of coordination among the different sectors. This problem of decentralized decision-making may result in a lower rate of overall technical advancement than if changes were more centralized. The time needed for the development of complementary technologies and other adjustments to take advantage of network externalities to make full use of the general purpose technology means that a long time may be required before marked changes in the measured rate of technological progress can be observed.

Faulty Predictions

Even more faulty have been the predictions, often by eminent scientists, that the stage of development had been reached that no further innovations could occur or at least none that could have substantial impacts in generating high employment or rapid economic growth. Such distinguished nineteenth-century economists as John Stuart Mill and Alfred Marshall presented some similar claims, as did the twentieth-century economist Alvin Hansen, in the Great Depression of the1930s (Mill 1895, II, pp. 334–340; Marshall 1920, pp. 67–68, 242–244; Hansen 1939). For more optimistic expectations by Mill, see Hollander (1985, I, p. 223; II, pp. 881–888). Unlike many others, Mill regarded the stationary state as a desirable outcome. Indeed, most periods of economic decline have provided proponents of such a decline of innovation, as John Taylor has pointed out (Taylor 2014). Most recently, such a claim has been made by economists such as Benjamin Friedman and Robert Gordon, despite the body of past evidence to the contrary (Gordon 2012, 2014; Friedman 2013; see, however, Mokyr 2014).

Competition Between Old and New Technologies

Given the nature of the economy, it is to be expected that new methods and innovations will emerge in competition with older technologies. The persistence of earlier technologies can often be the cause of delays in benefits for the new innovation slowing the rate of introduction of the new methods. Some of this may be due to improvements made to older technology, which keep them competitive with the new for at least a longer period of time. The long-term existence of a capital stock based on the older technology, which no longer needs to cover fixed costs, means that relatively lower prices may lead to some continued use of the older technology. Some of the lag may be due to the investors of the old technology who may, via the use of market forces or government action, work to reduce or exclude

the new. Similarly, laborers who prefer the economic conditions under the old technology may use the market or the government to prevent or delay the introduction of new methods, such as containerization (Levinson 2006). The late nineteenth-century political commentator Henry Sumner Maine (1897), in his argument against democracy, claimed that if workers had been able to vote on the introduction of innovations, the Industrial Revolution could not have taken place. Other delayed impacts may reflect government-chosen policies, at times in response to citizen's wishes. Tariffs (or their absence) have long played a major role in the timing of the introduction of a new technology. The nature of the patent system and its changes, over time, will affect the incentive to innovate as well as their diffusion (Khan 2005).

In the early days of the introduction of railroads in New York State, in competition with the Erie Canal, there were several attempts to reduce the railroad's competitive edge, such as limiting railroad operations to times when the canals were closed, requiring railroads to pay a toll equivalent to that of canals for freight carried, and a requirement that the railroad freight charge be the same as canal charges (MacGill 1917, pp. 291–294, 316–322, 344, 353–356, 368, 389, 398–400, 489, 495, 533–557; Engerman and Sokoloff 2006, pp. 110–112). Other states and nations introduced policies to limit expansion of railroads at the expense of canals. Pennsylvania introduced a tax on the Pennsylvania Railroad in 1846 to "guarantee the states against losses that might be sustained as a result of competition between the new railroads and the public works" (Hartz 1948, pp. 267–271; Dunlavy 1994). Ohio, similarly, had passed legislation to require railroads "to reimburse the state for half the canal tolls lost in all freight that the road carried between cities located on the Ohio Canal," as well as other limiting regulations (Scheiber 1969, pp. 270–317). None of these state legislations lasted very long, but they do indicate the type of problems confronted by innovations in competing with entrenched interests who were able to use governmental power.

Another consideration affecting the timing and magnitude of the introduction of an innovation and its full accomplishments is the cyclical nature of the economy, reflecting expectations of the future path of profitability as well as the availability of capital for investments required (Rosenberg and Frischtak, in Rosenberg 1994, pp. 62–84). The influence of cyclical changes can explain the clustering of innovations, as well as the lag between innovation and introduction into production, a point stressed by Schumpeter (Rosenberg 1982, pp. 5–7).

The Axiom of Indispensability

In determining the benefit-cost ratio of all the expenditures on research and development leading to innovations, it is important to remember that we should not look only at successful innovations and ignore the costs of failed attempts to develop new techniques, often in direct competition with those methods that have been successful. Thus, estimating the return to the antebellum canal network should not stop with the measured benefits from the Erie Canal, but needs to also deal with the losses of the

six other cities that, at roughly the same time, unsuccessfully competed against the Erie Canal (Engerman and Sokoloff 2006, pp. 97–98, 112. See also Rosenberg 2010, pp. 275–279; 1982, pp. 55–62).

An important consideration relating to estimates of the benefits of an innovation is what Robert Fogel called the axiom of indispensability (Fogel 1964, p. 10; cf. Rosenberg 1982, pp. 27–29). Fogel claimed that in the absence of the innovations that made the railroad successful, resources could possibly have been devoted to seeking other means of overland transport, such as the automobile, which might then have been introduced earlier than it was and which, as did the railroad, could improve its efficiency over time. Thus Fogel denies that the railroad was necessarily indispensable for US economic growth. Given some limitation upon the magnitude of resources that society will devote to innovating and improvements, the expenditures on a particular set of innovations and improvements will reduce expenditures on alternatives which, even if not ultimately as effective as the successful innovation was, may have been nearly so successful as was the adopted successful innovation. That such a possibility is not fanciful can be seen in current debates as to whether the pattern of change in the internal combustion engine came at the expense of devoting resources to developing such possible alternatives as the electric car, leaving us far behind in adapting to the current climate crises.

A further issue raised by Fogel's axiom of indispensability is the possibility that alternative innovations could have been made to replace any one specific innovation or several related innovations. This points to a broader set of questions concerning the possibility of alternative innovation in different parts of the world. This has been most frequently discussed in the context of arguing about the differences between East and West and the causes of the economic rise of the West. Most studied have been the nature and also the impact of innovation in China compared to that in Europe. There are several questions. One is the contention made by Needham (1969; see Winchester 2008) about the greater early successes of China in innovations than in Europe, an early lead that over centuries disappeared as economic and other expansions in Europe came to exceed those of China. Second, why, in many cases, did early modern Europe do more to make these innovations practical and useful than did China – whether due to cultural factors or taste differences, the range of usable knowledge, differences in relative factor prices and resource scarcities, or some limits of technological skills (for this discussion, see Allen 2011; Landes 1996; Jones 1981; Rosenberg and Birdsall 1986; and Mokyr 2002) among those discussing this point? Third, suggested most directly by the Axiom of Indispensability, is it possible or probable that East and West pursued different technological and institutional means to achieve the same general aim? Given differences in historical background and resources, were there differences in technological development that emerged before large-scale contact between these societies? In today's world with rapid communication and much day-to-day contact among scientists and inventors, the possibility of major divergences might seem doubtful. Nevertheless, the examination of the existence of such differentials at earlier stages of science and technical development should prove to be of importance and of interest.

Linear vs Chain-Linked Models

Despite his important role for economists and economic historians in pointing to the importance of technological change in accounting for economic growth, Schumpeter's story is in some ways still incomplete (Rosenberg 2000). He distinguishes between the major innovation and the subsequent improvers, whom he describes as "mere imitators." Thus he downplays the importance of those improvements made after the introduction of the innovation. These "imitators" can be heavily involved in enhancing the productivity of an innovation (Rosenberg 2000, pp. 55–78). The "imitator" may not get the glory that goes to the innovator, but it is often the imitators who reap the largest financial rewards. The "first mover" innovator may not be the greatest financial beneficiary of change, a phenomenon true not only for innovators but also for the economic growth of nations, as pointed out in an article by Ames and Rosenberg (1963; also Engerman and Sokoloff 2012).

Schumpeter further argues that the importance of major innovations plays a great role in contributing to the maintenance of the capitalist system (Rosenberg 1994, pp. 47–61). Capitalism creates new structures, new commodities, new technologies, new sources of supply, new markets, and new forms of organization – which drive out existing structures by the process he calls "creative destruction" (Schumpeter 1942, pp. 81–86), the mechanism by which new innovations drive out older systems. To Schumpeter, it is by the major innovations and in big jumps in the innovations, rather than by minor changes, that, he argues, capitalism is able to keep expanding (Rosenberg 1982, pp. 3–33).

One customary view of the innovation process, which has been called the linear model, suggests a rather "smooth, well-behaved linear process" from new developments in science to invention to innovation to production to marketing. This model allows for no feedbacks and no interactions among the various steps and is compatible with a Schumpeterian emphasis on innovation as an exogenous process and with technological change being regarded as discontinuous. Distinctions are made between the current scientific frontier and the past accumulation of scientific knowledge. The reality of the innovation process clear, however, to those who study the historical process of innovation has been better described as a chain-linked model, "complex, variegated, and hard to measure." This can include feedbacks and temporal interactions among accumulated science, innovation, production, and marketing (Kline 1985; Kline and Rosenberg in Rosenberg 2010, pp. 173–202). Developments at each stage influence, and are influenced by, what happens at the other stages, as for example, the contribution of technological improvements to the progress of science. This view is compatible with recent studies of innovation that regard it as being incremental and continuous, with attention given to the importance of small improvements based primarily on experience and "learning by using," with the prototypical case being the aircraft industry (Vincenti 1990; Rosenberg 2010, pp. 153–172, Rosenberg 1982, pp. 120–140; Mowery and Rosenberg, in Rosenberg 1982, pp. 161–177).

The introduction of innovations and improvement do not necessarily begin with new scientific information, but often are based on a preexisting state of knowledge. It

is often that developments in technology, as with the microscope, permit new scientific discoveries. These may result from "learning by using," with the benefits that result from solving problems that arise in the production process.

For these and other reasons, such a chain-linked model is more realistic than the linear model and serves to highlight the difficulties and complexities of the innovation process as it actually takes place. Innovations may not be based only on the newest science but can draw on the accumulations of past scientific development. This linked-chain model has been seen to be quite useful for describing productivity changes in the airplane, as well as the increased importance of electricity in the economy (Vincenti 1990; Kline 1985). In both cases, there were many unanticipated difficulties, necessitating modifications in product design as well as in operating and maintenance procedures. And, as seen in the case of nineteenth-century America and twentieth-century Japan, the technologically advancing nations can benefit from imitating the developments in the scientifically more advanced nations and need not themselves develop new innovations.

Conclusion

This chapter is intended to draw together some aspects of technical change and innovation that have been understated in the recent literature. The studies of the historical process by which innovations are made, introduced, and contribute to economic growth demonstrate the complexity of the process which is masked in some theoretical discussions.

The complexity of the process by which innovations occur and are introduced and diffused throughout the economy has become recognized. Large-scale technological steps have long been the primary focus in the examination of technological change. But more recently, the importance of what seem to be relatively minor adjustments have led to some shift in emphasis in historical and economic studies. This has led to a greater understanding of the great uncertainty in forecasting future technological changes, of the often long-delayed measured achievements of what are regarded as new major technologies, and of the need to bring in the study of institutions into the analysis of technological change. Consideration of these factors will not only increase our historical knowledge but also serve to enrich our theorizing about these questions.

Acknowledgments We wish to thank Philip Hoffman, Zorina Khan, Joel Mokyr, and the editors of this volume for very helpful comments on earlier drafts.

References

Allen RC (2011) Global economic history: a very short introduction. Oxford University Press, Oxford

Ames ED, Rosenberg N (1963) Changing technological leadership and industrial growth. Econ J 73:13–31

Arrow KJ (2012) The economics of inventive activity over fifty years. In: Lerner J, Stern S (eds) The rate and direction of inventive activity revisited. University of Chicago Press, Chicago, pp 43–48
Breschi S, Malerba F, Orsenigo L (2000) Technological regimes and Schumpeterian patterns of innovation. Econ J 110:388–410
Bresnahan TF, Trajtenberg M (1995) General purpose technologies 'engines of growth'? J Econom 65:83–108
Dunlavy CA (1994) Politics and industrialization: early railroads in the United States and Prussia. Princeton University Press, Princeton
Engerman SL, Sokoloff KL (2006) Digging the dirt at public expense: governance in the building of the Erie canal and other public works. In: Glaeser EL, Goldin C (eds) Corruption and reform: lessons from America's economic history. University of Chicago Press, Chicago, pp 95–122
Engerman SL, Sokoloff KL (2012) Economic development in the Americas since 1500: endowments and institutions. Cambridge University Press, Cambridge
Fogel RW (1964) Railroads and American economic growth: essays in econometric history. Johns Hopkins Press, Baltimore
Friedman BM (2013) Brave new capitalists paradise: the jobs? New York Review of Books, 60 (November 7), 74–76
Gordon RJ (2012). Is US economic growth over? Faltering innovation confronts the six headwinds. National bureau of economic research, working paper 18315
Gordon RJ (2014) The demise of U.S. economic growth: restatement, rebuttal, and reflections. National bureau of economic research, working paper, 19895
Hansen AH (1939) Economic progress and declining population growth. Am Econ Rev 29:1–15
Hartz L (1948) Economic policy and democratic thought: Pennsylvania, 1776–1860. Harvard University Press, Cambridge, MA
Helpman E (1998) General purpose technologies and economic growth. MIT Press, Cambridge, MA
Hollander S (1985) The economics of John Stuart Mill, 2 vols. Toronto University Press, Toronto
Hughes TP (1983) Networks of power: electrification in western society, 1880–1930. Johns Hopkins University Press, Baltimore
Jamasb T (2007) Technical change theory and learning curves: patterns of progress in electricity generation technologies. Energy J 28:51–71
Jones EL (1981) The European miracle: environments, economics, and geopolitics in the history of Europe of Europe and Asia. Cambridge University Press, Cambridge
Jovanovic B, Lach S (1989) Entry, exit, and diffusion with learning by doing. Am Econ Rev 79:690–699
Khan BZ (2005) The democratization of invention: patents and copyrights in American economic development, 1790–1920. Cambridge University Press, Cambridge
Kline SJ (1985) Research, invention, innovation, and production: models and reality. Stanford University: Department of Mechanical Engineering, Stanford
Kuznets S (1973) Innovations and adjustments in economic growth. In: Population, capital, and growth: selected essays. Norton, New York, pp 185–211
Kuznets S (1979) Technological innovations and economic growth. In: Growth, population, and income distribution: selected essays. Norton, New York, pp 56–99
Landes DS (1996) The wealth and poverty of nations: why some are so rich and some so poor. Norton, New York
Levinson M (2006) The box: how the shipping container made the world smaller and the world economy bigger. Princeton University Press, Princeton
Lipsey RG, Carlaw KI, Becker CT (2005) Economic transformations: general purpose technologies and long-term economic growth. Oxford University Press, Oxford
MacGill C (1917) History of transportation in the United States before 1860. Carnegie Institution, Washington, DC
Maine HS (1897) Popular government: four essays, 5th edn. J. Murray, London

Marshall A (1920) Industry and trade: a study of industrial technique and business organization. Macmillan, London
Mill JS (1895) Principles of political economy: with some of their application to social philosophy, 2 vols. D. Appleton, New York
Mokyr J (2002) The gifts of Athena: historical origins of the knowledge economy. Princeton University Press, Princeton
Mokyr J (2014) The next age of invention. City J 24:12–21
Mowery DC, Rosenberg N (1998) Paths of innovation: technological change in 20th century America. Cambridge University Press, Cambridge
Needham J (1969) The grand titration: science and society in east and west. George Allen & Unwin, London
Rantisi N (2002) The competitive foundations of localized learning and innovation: the case of women's garment production in New York City. Econ Geogr 78:441–462
Rosenberg N (1982) Inside the black box: technology and economics. Cambridge University Press, Cambridge
Rosenberg N (1994) Exploring the black box: technology, economics, and history. Cambridge University Press, Cambridge
Rosenberg N (2000) Schumpeter and the endogeneity of technology: some American perspectives. Routledge, London
Rosenberg N (2010) Studies on science and the innovation process: selected works. World Scientific, Singapore
Rosenberg N, Birdsall LE Jr (1986) How the west grew rich: the economic transformation of the industrial world. Basic Books, New York
Sabin P (1999) A dive into nature's great grab-bag: nature, gender and capitalism in the early Pennsylvania oil industry. Pa Hist 66:472–505
Scheiber HN (1969) Ohio canal era: a case study of government and the economy, 1820–1861. Ohio University Press, Athens
Schumpeter JA (1942) Capitalism, socialism, and democracy. Harper and Brothers, New York
Stein J (1997) Waves of creative destruction: firm-specific learning by doing and the dynamics of innovation. Rev Econ Stud 64:265–288
Taylor JB (2014) Will the real secular stagnation thesis please stand up. Wall Street Journal, (January 5):A17
Tenner E (1996) Why things bite back: technology and the revenge of unintended consequences. Knopf, New York
Vincenti WG (1990) What engineers know and how they know it: analytical studies from aeronautical history. Johns Hopkins University Press, Baltimore
Winchester S (2008) The man who loved china: the fantastic story of the eccentric scientist who unlocked the mysteries of the middle kingdom. Harper, New York

The Cliometric Study of Innovations

Jochen Streb

Contents

Abstract

Per definition, cliometric studies of innovations use statistical methods to analyze large quantities of data. That is why historical patent statistics have become the standard measure for innovation. I first discuss the advantages and shortcomings of patent data and then show that the distribution of patents across countries, regions, or inventors is characterized by two salient features: its skewness and its persistence over time. To explain these features, the influence of various supply-side, demand-side, and institutional factors will be discussed. I will stress the importance of path dependency. This chapter ends with a closer look at technological transfer that came along with patent assignments and foreign patenting.

Keywords

Patent · Patent statistics · Human capital · Skewed distribution · Technological transfer · Path dependency · Innovation · Region · Patent law · Access to market

J. Streb (✉)
Abteilung Volkswirtschaftslehre, Lehrstuhl für Wirtschaftsgeschichte, Universität Mannheim, Mannheim, Germany
e-mail: streb@uni-mannheim.de

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_18

Introduction

Economic historians agree on the stylized fact that innovations are the main driver of long-run economic growth. For example, Greg Clark (2007, pp. 197–202) estimates, on the basis of a growth accounting exercise, that about three quarters of long-term growth of output per worker in the industrialized world has to be directly attributed to the permanent increase in productivity which, in his opinion, mainly resulted from the myriad of smaller and larger innovations that were developed to improve the efficiency of production processes. An important corollary of this empirical observation is that the unequal geographical distribution of innovations might be the key factor for explaining why some nations became rich and others stayed poor. That is why cliometric studies of innovations usually concentrate on two main tasks. First, they aim for measuring the distribution of innovations across space and time. Second (and based on this measurement), they try to identify those factors that have influenced the innovation of nations, regions, or firms. To perform this task with the method that differentiates cliometric studies of innovations from other research projects in innovation history – advanced statistical analysis –mass data are needed, the collection of which is at the same time one of the major methodological challenges. The epistemic interest of this research program is clearly related to the field of development economics: underdeveloped countries of today might learn from historical experience how to foster their own innovative capabilities and therefore their future economic performance. In the following, I will discuss the problems and results of measuring innovations under the headings "quantifying innovations" and "skewed distribution." The cliometric approaches to elucidate the development and diffusion of innovations are presented under the subtitles "explaining innovations" and "technological transfer."

Quantifying Innovations

In the early twentieth century, Schumpeter (1934, p. 66) provided his famous and still very instructive definition of innovation by distinguishing five different cases: the introduction of a new good or a new quality of a good, the introduction of a new method of production, the opening of a new market, the conquest of a new source of supply of raw materials or half-manufactured goods, and the carrying out of the new organization of any industry. The practical research problem of economic historians who aim at basing their empirical research on Schumpeter's definition is how to collect complete data about these rather different types of innovations in a way that allows consistent comparisons across space and time. Compilations of historical innovations that are usually provided by scholars of the history of technology are by no means comprehensive and frequently show a considerable selection bias because historians tend to prefer both basic innovations to incremental innovations and product innovations to process and organizational innovations. That is why economic historians usually rely on patent statistics as the standard measure to quantify past innovations. This preference is obviously based on the implicit assumption that,

in comparison to the compilations of innovations by historians, patent statistics offer a more complete and less biased overview of the universe of innovations. In general, two types of patent statistics have to be distinguished. Patents applied for are a measure for innovations that were appraised to be new and potentially profitable by the applying inventor. In patent systems, where the patent office is vested with the task to reject patent applications because of lack of novelty, patents granted can be interpreted as a measure for the subset of innovations which were additionally judged to be new by the impartial technical experts of this administration. Both groups of patents can differ considerably. In pre-First World War Germany, for example, only about 40% of patent applications successfully passed the technical examination by the patent office (Burhop and Wolf 2013, p. 76).

Patent statistics have obvious shortcomings too. Griliches (1990, p. 1669) highlights the three most important of them: "Not all inventions are patentable, not all inventions are patents and the inventions that are patented differ greatly in 'quality', in the magnitude of innovative output associated with them." The first part of this statement points out that patent statistics can only contain information about product and process innovations but fully neglect, as most of the compilations of innovations, the last three types of innovations on Schumpeter's list that are in general not patentable. To close this gap of knowledge, survey-based studies in modern innovation economics sometimes explicitly ask for information about organizational innovations in marketing, procurement, or internal organization of a company. In economic history, however, comparable mass data are usually not available. The same is true for input indicators, such as R&D expenditures by private firms or public research organizations, which are also often used in nonhistorical studies of innovations, which concentrate on the development in the last decades.

The second part of Griliches' statement refers to the fact that the propensity to patent varies considerably across industries. Whereas some industries try to appropriate the return of their innovations with the help of patenting activities, others prefer keeping them secret instead. The formula for Coca-Cola, for example, has never been patented because its public disclosure in a patent application would have allowed competitors to imitate this product after the end of the patent protection. Given these differences in industries' patenting activities, it could be misleading to interpret a particular industry's comparatively low number of patents automatically as a sign for its alleged below-average level of innovation. To assess the magnitude of this measurement problem in cliometric studies of innovations, Moser (2012) uses an alternative source to identify them. She looks at the number of British and American exhibits presented at world's fairs between 1851 and 1915. The historical catalogues used to guide the visitors through the exhibition of a particular world's fair comprise information about the exhibitor's name, location, and a description of the innovation. The latter allows Moser to assign every exhibit to exactly one of ten different industries. Because the catalogues also provided information about whether or not the exhibit was patented, she can also calculate the patenting rates of the exhibits. At the Crystal Palace exhibition in London in 1851, for example, about 89% of British exhibits and 85% of the American ones were without patents. In the light of this observation it is hard to maintain the general claim that historical patent

statistics offer a sufficiently precise overview of innovative activities. In addition, Moser identifies considerable differences in industries' propensity to patent. In 1851, industry-specific patenting rates of British exhibits ranged from 30% in manufacturing machinery and 25% in engines to a mere five percent in mining and metallurgy. Moser concludes that patenting rates were especially low in those industries where innovations were difficult to imitate. In the middle of the nineteenth century, this argument also applied to chemicals, because modern methods of chemical analysis that allowed chemical products to be "reengineered" had not yet been developed. Even though patenting rates gradually increased over the course of the second half of the nineteenth century, Moser's analysis clearly shows that patent statistics are by no means a perfect measure for historical innovations. On the other hand, patent statistics are often the only source of mass data available for cliometric studies of innovations. When using this second-best measure, researchers are therefore well advised to control for industry effects in their regression analysis.

The third part of Griliches' statement addresses the problem that patent counts allocate the same weight to every patent, no matter whether it had a high or a low economic value for the patentee or society. This is an additional reason why inferring the level of innovation from the raw number of patents can lead to considerable measurement error. For this particular problem, however, scholars found various ways to deal with it. Ideally, one would like to assign each patent an individual weight that quantifies its technological or economic significance. Townsend (1980), for example, rated historical patents related to coal mining according to their importance, on a scale from 1 to 4. This procedure might be recommendable for specific industry studies, but does not work for large patent populations where the careful evaluation of every single patent would be very time-consuming and would require engineering competence in a wide range of technological fields. In order to address this problem, economic historians use three other methods to identify patents with a high economic value. Figure 1 illustrates these methods. We already know that the set of patents filed in a particular country is only a more or less large subset

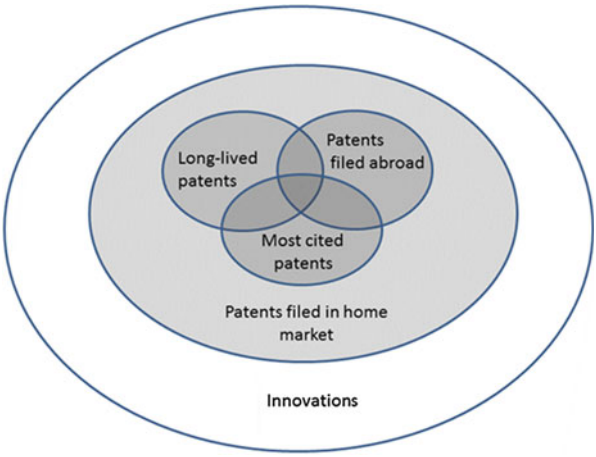


Fig. 1 Identifying valuable patents

of all innovations that have been developed there in a given time period. Among all patents filed in the home market are in turn three non-disjoint subsets that, for different reasons, all might represent valuable patents. These are the subsets of foreign patents, long-lived patents, and most-cited patents.

An inventor can apply for a patent not only in his home market, but also in foreign countries. Getting a foreign patent, however, imposes additional costs in the form of expenses for patent lawyers and translators, fees for filing and renewing, and the longer-term costs of international disclosure of the underlying technology. Future returns on a foreign patent can arise from two major sources. A patentee can use the temporary patent protection to increase his profits either by exporting the innovative good or by licensing foreign producers to manufacture and sell it in their respective home markets. After weighing the costs and benefits of foreign patenting, most inventors decide to file a patent only in their home country. Only the most promising innovations will also be patented abroad. That is why foreign patents might represent an especially valuable part of a country's patent stock. Today, the so-called triadic patents that are simultaneously filed at the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO), and the Japanese Patent Office (JPO) are used to identify a country's best innovations.

Economic historians usually concentrate on foreign patenting in the United States for two reasons. First, early on the United States established a large and developed market in which only excellent foreign innovations could take hold. Second, the USPTO provides comparatively detailed and long-term historical patent statistics. The most comprehensive cliometric analysis is provided by Cantwell (1989) who analyzes the patenting activities in the United States of 17 industrialized countries and 27 sectors for the years 1890–1892, 1910–1912, and 1963–1983. A shortcoming of this kind of identification strategy is that the volume and structure of foreign patents are probably not independent of the characteristics of the foreign country where they are filed. In general, firms will seek patent protection only in those foreign countries where two preconditions hold: first, the potential market for their innovation is large, and second, the probability of imitation is high. What is more, some countries might even discriminate against foreign inventors by delaying or even declining the granting of their patent applications (Kotabe 1992). As a result, the portfolio of a country's foreign, and therefore valuable, patents might look very different depending on whether it has been derived from foreign patenting activities in, for example, Germany, Japan, Spain, or the United States.

In historical patent systems like those of Germany or the United Kingdom, where patent holders had to renew their patents regularly by paying a renewal fee, valuable patents can alternatively be identified by their individual life span (Schankerman and Pakes 1986; Sullivan 1994). Legislators had introduced patent renewal fees in the hope that many patent holders who were not able to profitably exploit their patents would give them up early and thereby make the new knowledge that was documented in the patent file publicly usable long before the maximum possible patent duration would have elapsed. If this mechanism worked as intended, a long life span of a historical patent can be seen as a reliable indicator of its comparatively high private economic value. In the German Empire, for example, a patent holder

had to decide annually whether he wanted to prolong his patent by another year. The renewal fee amounted to 50 Marks at the beginning of the second year and then grew steadily up to 700 Marks at the beginning of the fifteenth and final possible year of patent protection. The resulting cancellation rate was high. About 70% of all German patents that were granted between 1891 and 1907 had already been cancelled after just 5 years. About 10% of all patents were still in force after 10 years and only about 5% reached the maximum age of 15 years. Streb et al. (2006) interpreted those German patents that survived at least 10 years as the valuable patents within the German Empire.

However, the method of identifying valuable patents by their individual life span has three shortcomings. First, it can only be employed if the respective patent law stipulated the obligation to renew patents annually or, as in the British case, after 3 (later on: 4) and 7 years of patent protection, respectively. This was not the case in the often-researched US patent system, where patentees only had to pay a registration fee. Second, in industries with a high rate of technological progress, even patents representing important basic innovations might have been cancelled after just a few years as the technological frontier moved on. Third, in a world with imperfect financial markets, private inventors and smaller firms with limited financial capacity might have been forced by comparatively high renewal fees to give up their patents even though they still represented a high economic value (Macleod et al. 2003). Both types of short-lived but valuable patents will be systematically ignored by the life-span approach.

In academics, the value of a scientific article is often measured by the numbers of citations it received in following publications. A similar measure can be used to identify valuable patents. The idea is that the more often a particular patent is cited in subsequent patent specifications, the higher inventors evaluate its technological and economic significance (Jaffe and Trajtenberg 2002). Unfortunately, before the First World War, it was not common practice to refer to a preceding patent for defining prior state of the art. Even though most citations appear within one decade of patent issue, Nicholas (2011b) found that some British patents of the interwar period were still cited in US patents in the decades after the Second World War. Nuvolari and Tartari (2011) identified another way to make use of the concept of most-cited patents in cliometric studies. Their basic research design is to exploit Bennet Woodcroft's "Reference Index of Patents of Invention" published in 1862. This volume provides a list of references to technical and engineering literature, legal proceedings, and commentaries in which a patent is mentioned for each English patent granted between 1617 and 1841. Nuvolari and Tartari assume that the absolute number of references assigned to a particular patent shows its visibility in the contemporary technical and legal discussions and is therefore a reasonable indicator for its underlying value.

Depending on both data availability and the particular research agenda, a researcher is free to choose the most appropriate among the aforementioned methods for identification of valuable patents. However, to sharpen the definition it might be worthwhile to employ two or more methods simultaneously and concentrate on

those valuable patents which lie in intersections of the three subsets of foreign patents, long-lived patents, and most-cited patents depicted in Fig. 1.

Summing up, due to the scarcity of alternative sources for mass data, the vast majority of cliometric studies of innovations are studies of patenting activities. The main problem with this approach is that patent statistics neglect all innovations that were never patented, either because inventors preferred secrecy to patenting as a means to appropriate the return of their innovations or because the patent law did not provide for patenting particular innovations. Organizational innovations are an example of the latter problem. On the other hand, the use of patent statistics has the important advantage that researchers can choose between different sophisticated methods of identifying the valuable innovations within the set of all patents granted.

Skewed Distribution

A striking (and often neglected) feature of patent statistics is that the distribution of patents across countries, regions or inventors is highly skewed. Figure 2, for example, displays the number of long-lived German patents that were held by firms and private inventors located in the 20 most innovative foreign countries before the First World War. This represents the intersection between each country's long-lived patents and its patents filed in Germany, indicating a subset of particularly valuable patents.

Before the First World War, the United States dominated foreign patenting activities in Germany, with 29% of all long-lived foreign patents. Overall, the respective shares of the three (five) most innovative countries came to 63 (82) percent. This ranking of technological leadership has been persistent over time. On a world scale, the United States, Great Britain, France, and Germany (which, by definition, cannot show up in Fig. 2) have dominated foreign patenting activities for more than 120 years (Cantwell 1989; Hafner 2008). The only country that was able to join this exclusive club of technological leaders was Japan in the second half of the twentieth century. Cantwell suggests that we should explain the inability of most backward countries to achieve a similar level of innovation by the fact that in most industries new knowledge is generated as an incremental, cumulative and path-dependent process. As long-term paths of research and development provide no major shortcuts for latecomers, the technological leaders are in general far ahead of their followers when it comes to the development of major innovations.

Assuming that transaction costs (search and information costs, bargaining costs, monitoring, and enforcement costs) generally increase with distance, so-called gravity models predict that geographical (and cultural) proximity fosters bilateral foreign trade flows. Burhop and Wolf (2013) show that the same was true for international trade in German patents during the pre-First World War period. All other things being equal, the frequency of patent transfers decreased with growing distance between the buyer and the seller of a particular German patent. In addition, similar evidence can be found for the more general case of foreign patenting

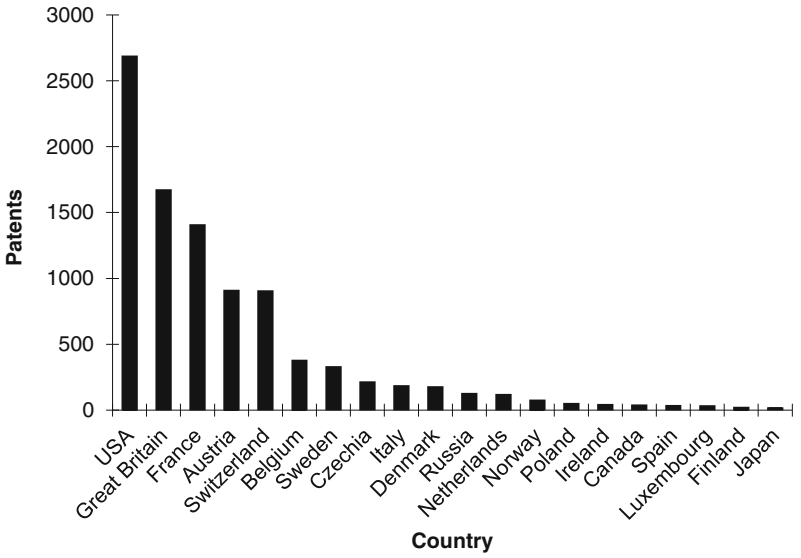


Fig. 2 Long-lived German patents of the 20 most innovative foreign countries before the First World War (Source: Degner and Streb 2013, p. 24)

activities. In particular, the comparatively high number of long-lived German patents that countries such as Austria or the modern-day Czech Republic possessed (see Fig. 2) might have resulted from direct proximity to this large neighboring economy. In contrast, these two countries played no major role in the American patent market, where Canada held a relative high number of patents.

The very uneven distribution of innovation across countries is mirrored within the innovative countries themselves; an observation that is reminiscent of the self-similarity of fractal geometry. In an influential paper that triggered many cliometric studies of innovations, Sokoloff (1988) points out that in the early nineteenth century, the level of patents per capita in southern New England and New York surpassed those of the rest of the United States by a factor of 20. Between 1890 and 1930 most Japanese independent inventors lived in the areas around Tokyo and Osaka (Nicholas 2011b). Streb et al. (2006) reveal that the long-lived German patents granted to domestic patentees before the First World War were also not uniformly distributed across the different German regions but were, as shown in Fig. 3, geographically clustered in the districts along the Rhine as much as in Greater Berlin and Saxony. A particularly high level of innovation, it seems, is a characteristic of regions rather than countries. For that reason, scholars have concentrated recently on the analysis of regional innovation systems (Malmberg and Maskell 2002).

Firm-level data indicate that above-average innovation of regions, in turn, is often based on achievements of just a few very innovative firms. Degner (2009), for example, presents the astonishing result that from 1877 to 1900 two thirds, and

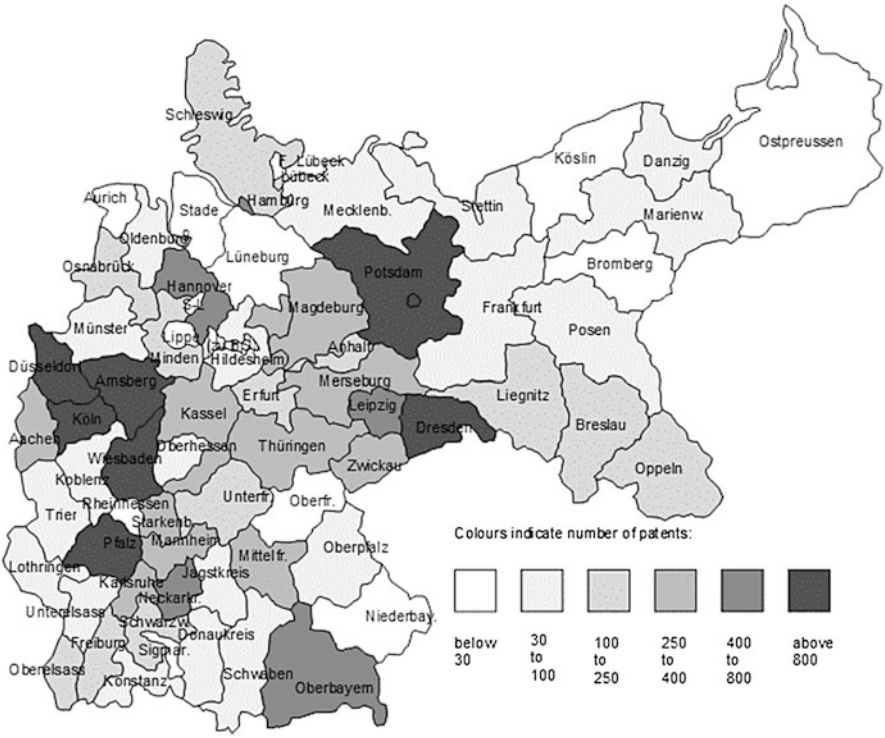


Fig. 3 The geographical distribution of high-value patents in Germany, 1878–1914 (Source: Streb et al. 2006, p. 364)

from 1901 to 1932 between 40% and 55%, of all long-lived German patents granted to domestic firms were held by only the 30 most innovative firms. That this distribution of innovation across firms was extremely skewed is emphasized by the fact that more than 266,000 firms with more than five workers existed in Germany in 1930. Many of the firms on Degner's list, such as Siemens or BASF, are also among the most innovative German firms of the early twenty-first century.

To conclude, many empirical observations lead to the conclusion that innovation, measured by the number of (valuable) patents, is a rare and persistent characteristic both at the macroeconomic and the microeconomic level. Surprisingly, most cliometric studies of innovations do not address these features explicitly.

Explaining Innovations

Traditionally, scholars have argued about whether an observed increase in innovations was primarily evoked by supply-side or demand-side factors. Mokyr (1990), for example, takes the view that demand-side factors might influence the direction of innovative activities but cannot explain the absolute level of technological creativity

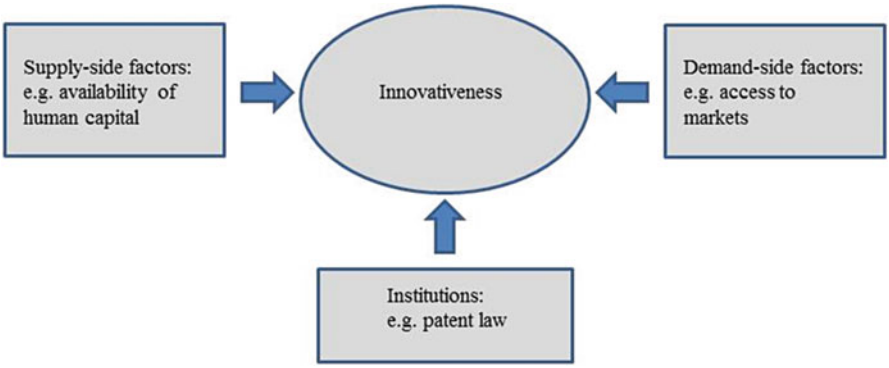


Fig. 4 Determinants of innovation

in a society. In his opinion, the latter was historically determined by various supply-side factors such as geography or the availability of basic technological knowledge. He also believes that demography and its influence on labor costs and popular preferences like the degree of risk aversion or the openness to new (technological) information were important. More recently, researchers also explored how the detailed anatomy of patent legislation influences the volume and structure of innovations. Figure 4 depicts the relationship of these three approaches.

Supporters of the view that it is the supply side of the economy that drives innovation stress the importance of human capital. In general, human capital comprises the stock of all qualifications and skills that increase an individual's productivity in economic activities. It can be acquired by formal education and learning by doing and therefore accumulates over the lifetime of a worker or researcher. Like physical capital, however, human capital can also be devaluated. Such a scenario is likely to occur in the aftermath of technological shocks. Handloom cotton weavers, for example, were highly paid specialists at the end of the eighteenth century, but were quickly replaced by unskilled adults and even children after Edmund Cartwright invented the power loom in 1785. Unfortunately, exact measures for human capital do not exist. Researchers therefore often rely on imperfect proxies like literacy rates, years of schooling, formal degrees, or even the Whipple index, which measures the extent of age heaping in a society (Baten and Crayen 2010).

At least since the Second Industrial Revolution, human capital has become an indispensable input in industrial innovation processes. In the late nineteenth century, chemical and electrical engineering companies invented the new organizational concept of the R&D department. Thus, for the first time in history, scientists and engineers collaborated to search systematically for new goods that could be profitably sold by their employer. In other industries such as mechanical engineering, drawing offices, and experimental departments, which also had to be equipped with well-trained employees, became an increasingly familiar sight. Human capital was now needed even for the purely imitating activities of firms. Reverse engineering, for example, meant in practice that workers had to have the skills to disassemble

complex machinery, to record each component with the help of engineering drawings, and to produce replica parts and fully functional copies. That is why Benhabib and Spiegel (1994) hold the view that human capital is essential for enlarging a country's level of technology by making possible either the imitation of foreign superior technology or the development of its own innovations. In their empirical approach, they measure a country's capability to innovate by its human capital stock, which they estimate by enrollment rates in primary, secondary and higher education. Specifically, a country's potential to imitate is approximated by the gap between the productivity level of the technological leader and its own inferior productivity level. The extent to which this potential can actually be used for catching up depends again on the available human capital stock. Analyzing the reasons for cross-country variation in growth rates of GDP per capita of 79 countries between 1965 and 1985, they confirm that in order to grow economically, emerging countries could rely on adopting foreign technology while industrialized countries had to develop better technology. These different growth strategies might also demand different strategies of human capital formation. Acemoglu et al. (2006) suggest that backward countries that want to catch up by imitating foreign technology should invest primarily in secondary education, whereas countries at the technological frontier should concentrate on increasing the quality and quantity of tertiary education.

Comparing the development of the synthetic dye industry in Great Britain, Germany, and the United States before the First World War, Murmann (2003) identifies the relative abundance of well-trained domestic chemists as one of the key factors that explain why German firms came to dominate the industry, as measured by both innovations and share in worldwide sales. From this observation arises the question whether the availability of an appropriate stock of human capital also influences innovation on a more disaggregated level. To answer this question, Baten et al. (2007) analyze the patenting activities of 2,407 firms located in the 52 districts of the state of Baden, Germany, around 1900. They measure regional human capital formation as the number of students in technical and commercial schools of secondary and tertiary level per 1,000 inhabitants. If the efficiency of a firm's R&D primarily depended on locally available human capital, firms that were located in districts with many students should have displayed more innovations than firms in districts with a below-average number of well-educated people. The econometric results suggest that Baden's small- and medium-sized firms relied on hiring graduates from technical and commercial schools in their geographical neighborhood. By contrast, Baden's large innovative firms were apparently able to cross geographical boundaries and acquire new researchers and engineers from distant German and foreign regions.

Principal-agent theory assumes that a worker's productivity depends not only on his human capital but also on the personal effort he is willing to make on the job. If the employer cannot observe the exact effort level because of asymmetric information and is therefore not able to reward diligence or punish sloth, a worker is not likely to do more than what is necessary to keep his job. This hypothesis might also be true for employees in industrial R&D departments, especially if they receive their pay in the form of a fixed salary. If a researcher does not participate in the company's

additional profits generated by his own innovations, he has no incentives to dedicate himself to the development of new goods and processes with all his heart and mind. Theoretically, an employer can set such incentives by paying a variable salary that increases with a researcher's output. Burhop and Lübbers (2010) explore whether this kind of incentive scheme worked in the R&D departments of German chemical and electrical engineering industries around 1900. They analyzed the contents of individual researchers' working contracts and found that among the three firms Bayer, BASF, and Siemens, only Bayer offered ex-ante contracted bonus payments that depended on the profits resulting from the employed researcher's inventions. In contrast, BASF and Siemens implemented discretionary reward schemes with no clear link between the level of bonus and a researcher's individual achievements. Regression analysis reveals that a high share of bonus payments in total compensation significantly increased the number of long-lived patents granted to a firm. Moreover, individual experience also mattered: total patent output rose with the average tenure of researchers.

If human capital has been the decisive bottleneck of innovating activities in history, its unequal distribution across countries and regions might help to elucidate the even more skewed geographical distribution of patents. Sokoloff and Khan (1990) disagree with this supply-side argument. They assume that during early American industrialization the skills and knowledge that were needed for successful patenting activities were widely spread among the general population. In their view, it was the unequal access to mass markets for innovative goods that explains why some regions became innovative and others did not. This demand-side argument is based on the assumption that the expected profitability of a patent increases with the size of the market in which the patented innovation might be sold. As land transport was prohibitively expensive before the introduction of railways, firms that were either located near highly populated metropolitan areas or able to transport their innovative goods at low costs on navigable waterways to distant markets had arguably much higher incentives to take out patents than did firms in more remote areas. To support this hypothesis, Sokoloff (1988) demonstrates that previously non-innovative northeastern American regions in the neighborhood of canals increased their patenting activities considerably after the completion of these waterways. Analyzing the biographical information on 160 "great American inventors," Khan and Sokoloff (1993) show that men of great technological creativity who did not already live in the traditional centers of innovation in New England and New York tended to move there. Interestingly enough, New England and New York kept their above-average level of innovation even after other American regions had gained similar market access due to the large extension of the railway network. This observation implies the likelihood of path dependency, which we will address below in more detail.

Demand factors not only influence innovation, but also firms' original choice of location. That is why it is necessary to distinguish clearly between a firm's choice of location and its decision to patent. Sokoloff is well aware of this problem and therefore controls for the division of the labor force between agriculture and manufacturing. It turns out that the estimated positive relationship between a

firm's proximity to navigable waterways and the intensity to patent is robust to the inclusion of this variable, which is supposed to measure the level of industrial activity in a region. Hence, in Sokoloff's sample, demand factors seem to influence the geographical distribution of patents independently of the original choice of location. The German case, however, suggests that the aggregated level of industrial activity might not be the adequate variable to distinguish between demand effects on firm location and on the decision to patent, respectively. German industries widely differed in their propensity to patent. The patent classes "electrical engineering," "chemicals including dyes," and "scientific instruments" together comprised more than one quarter of all long-lived patents granted between 1877 and 1918 (Streb et al. 2006). In addition, many valuable patents in the field of mechanical engineering were spread over several patent classes, such as "machine parts" or "steam engines" and less obvious ones like "weaving" or "agriculture" (which included textile machines and agricultural machines, respectively). The uneven propensity of industries to patent matters because of their simultaneous uneven geographical distribution across Germany. Obviously, the broad west-east strip of German regions with an above-average number of high-value patents, depicted in Fig. 3, was also the favored location of those industries in which most of the high-value patents originated. Long before the German patent law of 1877 actually came into force, the original choice of location for these industries might have been influenced by a variety of factors, such as the expected market volume or the availability of raw materials and intermediate products. Large (and later very innovative) chemical firms like BASF or Bayer, for example, preferred to settle on the banks of the Rhine, which was not only an important navigable waterway but was also used as a water source and a way to dispose of effluents. The great majority of chemical firms located themselves along waterways independently of their later decision to patent. Consequently, waterway areas had an above-average density of chemical firms, and because of this industry's high patenting activities, also had a higher number of patents than regions with a similar industrial activity level that were dominated by industries that patented less than the average. The same argument holds for mechanical and electrical engineering. Firms engaged in the field of mechanical engineering were especially concentrated in the geographical neighborhood of iron and steel producers, namely, in the Greater Ruhr area, and near textile firms, namely, in Saxony. Berlin was the center of German electrical engineering. To test the robustness of the relationship between a firm's proximity to metropolitan areas or mass transportation infrastructure and the propensity to patent proposed by Sokoloff, it would therefore be advisable to control not only for the general level of industrial activity in a region but also for the respective activity levels of different industries located in it.

Another point is worth mentioning. Sokoloff and his coauthors concentrate on the period of the First Industrial Revolution in early nineteenth-century America when the comparatively low level of human capital needed to invent a new steam engine or textile machine was widely dispersed among merchants and artisans. That is why, in the early nineteenth century and before, it might have been the access to mass markets for innovative goods that made a potential inventor into an actual one. During the Second Industrial Revolution of the late nineteenth century, however,

when basic innovations occurred in chemicals and electrical engineering, broadly dispersed general technical knowledge and skills might no longer have been sufficient for achieving a major technological breakthrough. This assumption is also supported by the fact that, in this period, the share of independent inventors among all patentees declined steadily while the respective share of researchers in industrial R&D departments increased (Nicholas 2011b, p. 1003). By then, the unequal geographical distribution of patenting has been rather determined by the unequal supply of higher education. It is therefore conceivable that the increasing importance of science and technology for innovation processes over the course of the nineteenth century shifted the main emphasis from demand-side factors to supply-side factors when it comes to explaining innovations.

Yet another argument in favor of the view that innovation is mainly driven by demand-side factors is the observation that upstream manufacturers' search for innovations is often driven by the concrete needs of their downstream customers. Streb et al. (2007) observe a statistically significant bidirectional Granger causality between German net cloth exports and patents in the technological classes "dyes" and "dyeing," which suggests that during the German Empire, the knowledge exchange between chemical and textile firms created an upward cycle of endogenous growth. Specifically, after the invention of many synthetic dyes in the last third of the nineteenth century, German chemical companies soon realized that textile manufacturers were not able to process synthetic dyes with their traditional equipment. That is why the former also engaged in the development of new chemical and mechanical procedures suitable for processing synthetic dyes. In a next step, this new knowledge was communicated to the downstream textile industry. The main channel of this knowledge transfer was the newly invented customer consulting service of the German dye manufacturers, which regularly informed textile firms about both new dyes and new dyeing methods. The German textile firms subsequently increased their international competitiveness to a considerable extent by exporting cloth colored with the innovative dyes. The increasing demand for synthetic dyes by the prospering textile firms in turn encouraged further R&D projects by the innovative chemical firms that led to new patents and again, via customer consulting, to additional economic benefits of the German textile industry. This upward cycle, however, was not infinite. It came to an end when the synthetic dyes technology had matured.

Various cliometric studies of innovations (Burhop and Lübbers 2010; Cantwell 1989; Khan and Sokoloff 1993) imply that the outstanding innovation of certain regions, companies, and independent inventors might have been built up in a path-dependent process. Degner (2012) elaborates on this hypothesis. The starting point of his theoretical considerations is the emergence of a new technology, such as the aforementioned chemical synthesis of dyes in the middle of the nineteenth century. Inspired by the economic opportunities that come along with a new technological field, in a first round of R&D, many newly founded companies with similar innovation capabilities will try to arrive at innovations. Given the high uncertainty of the innovation processes, however, only a few of these companies will succeed. Those firms will possess two advantages in the following second round of R&D.

They can now build on the scientific and economic knowledge their employees have acquired during the first round of R&D. In addition, the sales of the innovations developed in the first round of R&D might have led to the establishment of large financial reserves that will allow the innovative firms to expand their R&D capacities and therefore carry out several innovation processes simultaneously in the second round of R&D. Both advantages taken together considerably increase the probability that the winners of the first round of R&D will also make innovations in the second round – which, in turn, will foster their level of innovation in the third round of R&D even more. In contrast, firms that failed in the first rounds of R&D will soon no longer have a chance to catch up to the growing advantage of the early innovators. In the longer run, a path-dependent process will split initially very similar companies into few very innovative and many non-innovative companies.

To test his theoretical model, Degner analyzes the patenting activities of more than 1,000 German firms between 1877 and 1932. His striking result is that a firm's stock of valuable patents is a robust predictor of future patenting activities, whereas neither firm size, access to capital market, market structure, nor regional human capital endowment have a robust, significant influence on the number of valuable patents. Future research will show whether these empirical observations can be generalized. If this is the case, both the skewness of distribution of innovations across firms and regions (where innovative firms and individuals are clustered) and the persistence of innovation could be explained by the process of path dependency outlined by Degner.

Until now we have interpreted patent statistics as an admittedly imperfect but still objective measure for innovations. This view neglects the possibility that the introduction or change of a particular patent law itself might influence both the level and the direction of innovation activities. From a theoretical perspective, the introduction of formal intellectual property rights promises to foster innovation. The argument is that, in a world without patent protection, many inventors would have to fear economic losses because competitors would imitate innovations quickly and sell them at prices that only cover their own production costs but not the original inventor's R&D costs. Expecting this ruinous competition in advance, many potential inventors might decide to forego R&D projects that would otherwise lead to socially useful innovations. To fight this underinvestment in R&D, governments introduced patent protection, which allow successful inventors to recover their R&D costs by selling their innovations as a temporary monopolist.

This simple textbook explanation of the beneficial effects of formal intellectual property rights might be misleading in a more complex historical setting in which emerging countries struggle to catch up to the technological leaders. Murmann (2003), for example, argues that German chemical companies owed their meteoric rise to world market dominance in the late nineteenth century to a large extent to the absence of a German patent law before 1877, which made it possible to imitate British and French synthetic dye innovations and sell them in the unprotected German market. During this period of ruthless imitation, German imitators learned to master the new technology, build up R&D departments, and develop their own innovations. Perhaps unsurprisingly, after learning by imitation was completed,

German chemical firms began to lobby for the introduction of a domestic patent law because they now judged their newly acquired capability to innovate more profitable than their traditional imitating strategy. Richter and Streb (2011) confirm Murmann's narrative for the case of German machine tool makers who, in the second half of the nineteenth century, used various channels such as reverse engineering, visiting international exhibitions and foreign firms, scrutinizing international patent applications, and hiring foreign craftsmen and engineers to imitate superior American technology. In the early twentieth century, many of these former product pirates became internationally renowned innovators of machine tools.

In the nineteenth century the Spanish government found an elegant solution to have the best of both worlds: a full-fledged national patent system while maintaining the possibility to imitate superior foreign technology for free. So-called patents of introduction could be granted to Spaniards who were the first to introduce a foreign innovation into the Spanish market. For this technological transfer, the authorization by the original foreign inventor was not needed (Sáiz and Pretel 2013).

To conclude, good imitators can become good innovators when unsecure intellectual property rights give them the time they need to adjust to international competition in innovation. There is, however, one important caveat: Degner (2012) has shown that it is in general very difficult to catch up to the accumulated stock of experience innovative firms in industrialized countries have already built up in many historical R&D projects. Nevertheless, developing countries may realize that it does not pay to be an early complier with international rules of law with respect to intellectual property rights. Doing so may not only amplify the dominance of the traditional technological leaders but can also slow down the speed of technological and economic progress in their domestic industries.

Moser (2005) questions the alleged innovation-stimulating effects of patent protection on a more general level. Based on her research on exhibits presented at world's fairs between 1851 and 1915, she shows that countries without domestic patent laws did not display lower levels of innovation than countries with a long-standing tradition of patent protection. Switzerland, for example, which switched towards a fully functioning patent system only in 1907, regularly presented a comparatively high number of high-quality innovations at the world's fairs as measured by the jury prizes they received for exceptional novelty and usefulness. Moser has to admit, however, that patent laws might have influenced the direction of innovation activities. According to her research, countries without domestic patent protection concentrated their R&D activities on industries for which secrecy was a comparatively efficient means to appropriate the return to innovation. Many innovations in food processing, for example, such as milk chocolate, baby foods, and ready-made soups, were developed by Swiss or Dutch inventors in periods when neither country had a patent law. From this perspective, the current leading position of Dutch and Swiss companies in international markets for consumption goods might be a legacy of a long-gone era without patent protection.

Nicholas (2011a) adopts another approach to test for the influence of patent protection on innovation. His starting point is the observation that international patent systems differed considerably with respect to the fees a patentee had to pay to keep a patent in force. At the end of the nineteenth century, the total costs of

maintaining a patent for 15 years came to £265 in Germany, £84 in Belgium (for a 20 year term), £60 in France, and £54 in Italy. The important outlier was the United States where a mere £7 secured a 17-year patent protection, a fact which is believed to have promoted the "democratization" of innovation activities in this country (Sokoloff and Khan 1990). Nicholas does not exploit this cross-country variation, but concentrates on the English case where patent fees were reduced from £175 to £154 in 1883 for a 14 years' term. To be precise, the English fee reduction did not affect the two renewal fees payable by the end of the fourth year (£50) and by the end of the seventh year (£100). Only the initial fee, which was due at the beginning of the patent protection, declined from £25 to £4.

Following the arguments by Macleod et al. (2003) a first payment of £25 might have been prohibitively expensive for many potential English inventors. From this group's perspective, the fee reduction of 1883 represented the first affordable patent protection for their inventions. One would therefore expect an increase in English patenting activities after 1883, something which actually happened. Nicholas, however, wanted to know whether the 1883 reform also fostered innovation as measured by valuable patents, which he identified with the help of both the number of citations they received and their individual life span. Using a difference-in-difference regression to analyze changes in valuable English patents relative to the control group of valuable patents granted to English patentees in the United States, he concluded that the decrease in patent fees did not increase innovation. This finding has an important methodological implication. If the level of patent fees only influenced the number of total patents, but not the number of the valuable ones among them, an international comparison of the latter would be possible even when the variation of national patent fees was considerable.

National patent laws also differed with respect to other features. For example, some countries introduced technical examination or compulsory licensing clauses, while others did not. Moreover, some patent administrations discriminated against foreign inventors, while others did not (Khan 2013; Moser 2013). Because of these various differences, scholars should exercise a degree of caution when comparing patenting activities across different countries.

Technological Transfer

A major merit of patent laws is that they create a reliable legal framework for the diffusion of technology both within countries and between countries. The first channel through which the diffusion of technology can take place is the public disclosure of new knowledge. A patentee is required to provide a detailed technical description of his innovation in the patent specification that is then made available to the general public. Even though others are not allowed to exploit this information for the economic purpose specified in the patent during its period of validity, they can immediately use it as a starting point for related R&D projects. To prove that this diffusion mechanism already worked in the nineteenth century, Moser (2011) shows that innovation activities in the US chemical industry became less geographically

concentrated after this sector's propensity to patent had increased in the late nineteenth century.

To analyze the volume, direction, and impact of international technological transfer empirically, researchers traditionally rely on international data for bilateral trade flows or FDI. However, patent specifications can also serve foreigners as a source of new knowledge, especially when these patents were published in their native language. That is why Eaton and Kortum (1999) measure the direction of technological transfer by patenting activities in foreign markets. They conclude that since the end of the Second World War the world's long-term productivity growth has been mainly driven by the foreign patenting activities of a few leading research economies. The United States has been the dominant source of new knowledge, followed by Japan and Germany. Khan (2013) changes the perspective from countries of origin to the recipient countries. Interestingly enough, the average share of foreign patents in all patents granted varied considerably across countries between 1840 and 1920, for instance, from 78% in Canada, 59% in Spain, 34% in Germany, 22% in the United Kingdom, to only 7% in the United States. Based on this observation, she develops the hypothesis that lower rates of patenting by foreign inventors indicate a higher level of innovation of their domestic competitors. This might be true if inventors of any other country first and foremost engaged in those foreign markets where they did not fear the high technological creativity of the domestic population. Note, however, that Khan's considerations contradict the traditionally held assumption that foreign patenting concentrates in countries where the probability of imitation is high due to a comparatively rich endowment of technical competencies and skills.

According to Khan's statistics, the German patent market was a preferred destination of foreign inventors relative to other third markets. Using the concept of revealed technological advantage, Degner and Streb (2013) analyze the international patterns of technological specialization on the basis of foreign patenting activities of 21 countries from the European core, the European periphery, and overseas between 1877 and 1932 in Germany. It turns out that the countries of the European core revealed technological strength in the old technological fields of the First Industrial Revolution and in the new technological fields of the Second Industrial Revolution. Great Britain, for example, excelled in textiles, machine tools, electrical engineering, chemicals, and mass-consumption technology. In contrast, the Eastern and Southern European countries of the European periphery demonstrated technological strength only in the well-known technological fields of the First Industrial Revolution, such as Spain or Poland in the textile, coal, and steel industries. This difference suggests that a country's technological advantages were significantly influenced by its current stage of economic development. While the economically advanced countries of the European core had already explored the prospects of the more science-based technologies of the Second Industrial Revolution, the less advanced countries were still engaged primarily in the traditional technological fields of the First Industrial Revolution. This finding supports Cantwell's (1989) hypothesis that backward countries were not able to catch up to the superior level of innovation of the leading research economies.

A closer look at the performance of individual countries reveals further insights. The availability of domestic natural resources obviously influenced a country's technological specialization. Most of the countries with their own natural deposits of coal, iron, or other nonferrous metals, especially Belgium, Luxembourg, the modern-day Czech Republic, Poland, Norway, and Spain, displayed strong advantages in the technological field of the coal and steel industry, which included mining technologies for nonferrous metals. France, the Netherlands, and Denmark used their advanced agriculture to concentrate on innovations that fostered the mass consumption of foodstuffs and drinks. It is not surprising that Italy and France displayed great technological strength in the field of motor cars. Canada, however, which is not renowned for manufacturing automobiles, also revealed some technological advantage in this field before the First World War. Therefore, historical patterns of technological specialization might also produce information about abandoned national paths of technological development that would be otherwise forgotten.

The second channel through which patent protection facilitates the diffusion of technology is by decreasing the transaction costs of information exchange. As long as intellectual property rights were insecure, inventors who wanted to sell new ideas always had to fear being cheated out of their financial compensation. After the introduction of patent protection, it became easier and less risky to transfer knowledge via patent assignments, and particularly creative inventors specialized in invention activities. Lamoreaux and Sokoloff (1999, 2001) claim that the particular features of the American patent system, namely, the very low registration fee and the requirement that only the "first and true" inventor was entitled to apply for patent protection, were the key behind the pronounced division of labor in American innovation activities. Specialized inventors concentrated on the creation of technical inventions and then sold their new knowledge via patent assignment to established firms that took over the task of manufacturing and selling the innovation. Patent agents or lawyers often acted as an intermediary between inventors and companies. The relative importance of this type of technological transfer is demonstrated by the fact that around 1900, about one third of American patents was fully or partly assigned after issue. The historical German patent market was less liquid than the American one which can be explained at least partly by insufficient inventor protection (Burhop 2010). The German patent law ruled that the first applicant, not the initial inventor, was entitled to a patent grant. As a result, many innovations that were created in industrial R&D departments were directly granted to the company and not to the employed researcher. This is another example of how the details of a national patent law can significantly influence the outcome of innovation processes and therefore patent statistics.

Future Research

Until now, most cliometric studies of innovations have concentrated on patenting activities in leading research economies such as the United States, the United Kingdom, Germany, or Japan. To learn more about imitating and innovating in less advanced countries, future cliometric research projects should take a closer

look at patenting activities in the European periphery and overseas. The greatest challenge will be to harmonize the different national patent statistics and merge them into one unified data base that would allow for testing for the various determinants of patenting activities on the basis of a broad international panel. Given the obvious shortcomings of patent statistics, researchers should also keep on searching for alternative historical mass data which include information on innovations that have never been patented.

Another desideratum is to get more information about the microeconomics of historical R&D management. Here, the research idea of scrutinizing historical working contracts of employed researchers (Burhop and Lübbers 2010) might be a good starting point for further empirical analysis. Surprisingly enough, cliometric studies in innovations have widely neglected to research the impact of innovations on economic performance. It would be very interesting, however, to learn more about how the skewed (and persistent) distribution of innovations across countries, regions, and inventors affected the respective distributions of economic outcome indicators such as GDP per capita, productivity, or profit.

References

Acemoglu D, Aghion P, Zilibotti F (2006) Distance to frontier, selection, and economic growth. J Eur Econ Assoc 4:37–74

Baten J, Crayen D (2010) Global trends in numeracy 1820–1949 and its implications for long-term growth. Explor Econ Hist 47:82–99

Baten J, Spadavecchia A, Streb J et al (2007) What made southwest German firms innovative around 1900? Assessing the importance of intra- and inter-industry externalities. Oxf Econ Pap 59:i105–i126

Benhabib J, Spiegel MM (1994) The role of human capital in economic development: evidence from aggregate cross-country data. J Monet Econ 34:143–173

Burhop C (2010) The transfer of patents in imperial Germany. J Econ Hist 70:921–939

Burhop C, Lübbers T (2010) Incentives and innovation? R&D management in Germany's chemical and electrical engineering industries around 1900. Explor Econ Hist 47:100–111

Burhop C, Wolf N (2013) The German market for patents during the 'second industrialization', 1884–1913: a gravity approach. Bus Hist Rev 87:69–93

Cantwell J (1989) Technological innovation and multinational corporations. Basil Blackwell, Oxford

Clark G (2007) A farewell to alms: a brief economic history of the world. Princeton University Press, Princeton/Oxford

Degner H (2009) Schumpeterian firms before and after World War I: the innovative few and the non-innovative many. Z Unternehm 54:50–72

Degner H (2012) Sind große Unternehmen innovativ oder werden innovative Unternehmen groß? Eine Erklärung des unterschiedlichen Innovationspotenzials von Unternehmen und Regionen. Jan Thorbecke, Ostfildern

Degner H, Streb J (2013) Foreign patenting in Germany, 1877–1932. In: Donzé P-Y, Nishimura S (eds) Organizing global technology flows. Institutions, actors, and processes. Taylor & Francis, New York/Oxford, pp 17–38

Eaton J, Kortum S (1999) International technology diffusion: theory and measurement. Int Econ Rev 40:537–570

Griliches Z (1990) Patent statistics as economic indicators: a survey. J Econ Lit 33:1661–1707
Hafner K (2008) The pattern of international patenting and technology diffusion. Appl Econ 40:2819–2837
Jaffe A, Trajtenberg M (2002) Patents, citations and innovation: a window on the knowledge economy. MIT Press, Cambridge, MA
Khan BZ (2013) Selling ideas: an international perspective on patenting and markets for technological innovations, 1790–1930. Bus Hist Rev 87:39–68
Khan BZ, Sokoloff KL (1993) 'Schemes of practical utility': entrepreneurship and innovation among 'great inventors' in the United States, 1790–1865. J Econ Hist 53:289–307
Kotabe M (1992) A comparative study of the U.S. and Japanese patent systems. J Int Bus Stud 23:147–168
Lamoreaux NR, Sokoloff KL (1999) Inventors, firms, and the market for technology: US manufacturing in the late nineteenth and early twentieth centuries. In: Lamoreaux NR, Raff DMG, Temin P (eds) Learning by doing in firms, organizations, and nations. University of Chicago Press, Chicago, pp 19–60
Lamoreaux NR, Sokoloff KL (2001) Market trade in patents and the rise of a class of specialized inventors in the nineteenth-century United States. Am Econ Rev 91:39–44
Macleod C, Tann J, Andrew J et al (2003) Evaluating inventive activity: the cost of nineteenth-century UK patents and the fallibility of renewal data. Econ Hist Rev 56:537–562
Malmberg A, Maskell P (2002) The elusive concept of localization economics. Towards a knowledge-based theory of spatial clustering. Environ Plann A 34:429–449
Mokyr J (1990) The lever of riches: technological creativity and economic progress. Oxford University Press, Oxford
Moser P (2005) How do patent laws influence innovation? Evidence from 19th-century world fairs. Am Econ Rev 95:1214–1236
Moser P (2011) Do patents weaken the localization of innovations? Evidence from world's fairs, 1851–1915. J Econ Hist 71:363–382
Moser P (2012) Innovation without patents: evidence from world's fairs. J Law Econ 55:43–74
Moser P (2013) Patents and innovation: evidence from economic history. J Econ Perspect 27:23–44
Murmann JP (2003) Knowledge and competitive advantage. The coevolution of firms, technology, and national institution. Cambridge University Press, Cambridge
Nicholas T (2011a) Cheaper patents. Res Policy 40:325–339
Nicholas T (2011b) Independent invention during the rise of the corporate economy in Britain and Japan. Econ Hist Rev 64:995–1023
Nuvolari A, Tartari V (2011) Bennet Woodcroft and the value of English patents, 1617–1841. Explor Econ Hist 48:97–115
Richter R, Streb J (2011) Catching-up and falling behind: knowledge spillover from American to German machine tool makers. J Econ Hist 71:1006–1031
Sáiz P, Pretel D (2013) Why did multinationals patent in Spain? Several historical inquiries. In: Donzé P-Y, Nishimura S (eds) Organizing global technology flows. Institutions, actors, and processes. Taylor & Francis, New York/Oxford, pp 39–59
Schankerman M, Pakes A (1986) Estimates of the value of patent rights in European countries during the post-1950 period. Econ J 96:1052–1076
Schumpeter JA (1934) The theory of economic development. Harvard University Press, Cambridge, MA
Sokoloff KL (1988) Inventive activity in early industrial America: evidence from patent records, 1790–1846. J Econ Hist 48:813–850
Sokoloff KL, Khan BZ (1990) The democratization of Invention during early industrialization: evidence from the United States, 1790–1846. J Econ Hist 50:363–378
Streb J, Baten J, Yin S (2006) Technological and geographical knowledge spillover in the German empire, 1877–1918. Econ Hist Rev 59:347–373

Streb J, Wallusch J, Yin S (2007) Knowledge spill-over from new to old industries: the case of German synthetic dyes and textiles 1878–1913. Explor Econ Hist 44:203–223
Sullivan RJ (1994) Estimates of the value of patent rights in Great Britain and Ireland, 1852–1976. Economica 61:37–58
Townsend J (1980) Innovation in coal-mining: the case of the Anderton Shearer Loader. In: Pavitt K (ed) Technical innovation and British economic performance. Macmillan, London, pp 142–158
Woodcroft B (1862) Reference Index of English Patents of Invention, 1617–1852. G. E. Eyre & W. Spottiswoode, London

Arts and Culture

Karol Jan Borowiecki and Diana Seave Greenwald

Contents

Abstract

The economic history of arts and culture includes both "high culture" – like the fine arts, theater, and classical music – and popular culture, such as pop music, movies, and newspapers. This chapter focuses primarily on the high arts but also provides a cursory description of the literature addressing more popular cultural production. The four sections of this chapter correspond to four key areas of inquiry in the economic history of arts and culture: what are relevant data about the arts and how to capture them, how market forces encourage the consumption and supply of culture, how artistic production is linked to geography and clustering, and what drives creative output. This chapter surveys scholars' engagements with these questions across a wide range of art forms and time periods. It

K. J. Borowiecki (✉)
Department of Business and Economics, University of Southern Denmark, Odense, Denmark
e-mail: kjb@sam.sdu.dk

D. S. Greenwald
National Gallery of Art, Washington, DC, USA
e-mail: d-greenwald@nga.gov

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_31

concludes with a discussion of why the study of the economic history of arts and culture represents unique opportunities for interdisciplinary collaboration and is particularly relevant to present-day service economies.

Keywords
Art · Creativity · Innovation · Art markets · Culture

Introduction

A chapter on arts and culture is a necessary component of a handbook of cliometrics, if only to honor the Greek god for whom the discipline is named. According to Greek mythology, Clio was the muse of history, seated alongside her fellow muses responsible for the inspiration of poetry, music, dance, and other arts. Clio herself is sometimes credited with being the "proclaimer" of all these arts – announcing, preserving, and praising them for history (Graves 2012: 577–781). Like the muses, the economic history of arts and culture is multidisciplinary. Economists, economic historians, sociologists, and art historians have all made significant contributions to the subfield. Therefore, while this chapter will focus primarily on contributions by economic historians, it will necessarily summarize and engage with work originating in other fields.

Although considered a subcategory of cliometrics, the economic history of arts and culture is vast. It includes both "high culture" – like the fine arts, opera, classical music, and ballet – and more popular culture, such as pop music, movies, novels, newspapers, and video games. In just one chapter, it is impossible to summarize all of the work done on these many subjects. Therefore, this chapter focuses primarily on the high arts, the expertise of the co-authors. While we focus on the high arts, it is important to note that these are *cultural* products; they are a consequence of the wealth and tastes of individuals existing in a particular cultural setting. We do our best to also include an at least cursory description of the literature addressing other more popular cultural production, what Ruth Towse describes as "media economics. . .the area of the broadcasting, audiovisual and publishing industries" (Towse 2010: 6). Nonetheless, this chapter should not be viewed as a definitive survey of *all* cliometric work that can be classified as dealing with arts and culture. It is, instead, an introduction to the field, including suggestions for future reading for those interested in pursuing the economic history of arts and culture.

This chapter is divided into four sections, each of which responds to the four key questions in the economic history of arts and culture. The first section, "Problems with Data and Culture," asks what are relevant data about arts and culture and how do we capture them. The research question central to the section "Art Markets and Their Logic," is: if artistic output is a response to market forces, how have these forces changed over time to encourage and form consumption and supply? "Cultural Output and Geography," the third section, examines where artistic consumption and output are located. The market for the arts is *not* a globalized commodities market where interchangeable goods can be traded for a worldwide price. Instead, it is

centered in specific places and often involves a live performance. The final section, "Capturing and Fueling Creativity," summarizes the literature that explores whether it is possible to understand what drives creative output and therefore determine how to encourage it. Understanding what drives creativity is of significant policy interest as economies become service economies. In a service economy, economic growth is driven by highly educated workers who are often employed in positions and industries that demand creativity and innovation. Understanding the economic history of artistic output can help policy makers understand what will encourage creative output in the future.

In the conclusion to this chapter, we will describe the future of the subfield and how we believe it will continue to address these lines of inquiry and which new questions may be on the horizon. We encourage readers considering delving into the economic history of the arts in greater depth to explore the reference list included.

Problems with Data and Culture

To address research topics in the economic history of the arts – such as how and why individuals create high-quality artwork and understanding how art markets function – scholars need good data about historical cultural output. Like in other areas of economic history, it is challenging to find these data. However, locating high-quality historical data takes on a couple of important new dimensions when the study of arts and culture is involved. There are several obstacles that scholars face when determining what is relevant data about arts and culture and how to capture these data.

The first obstacle faced is the ability to translate the aesthetic quality and value of a work of art, music, literature, or performance into a data point. First, scholars need to adjust their usual data-driven methods to measure the unique and often subjective quality of cultural output. For example, should scholars use artists as their level of observation – therefore aggregating professional performance across a career – or should they focus on individual works, which can capture the good and bad works by famous artists and the occasional influential work by an otherwise unsuccessful artist? There is no correct answer to this question, but it is an example of the ways in which economic historians need to adjust to the particularities of studying artistic output.

A second difficulty is related to survivor bias, which is particularly pronounced in the study of the history of arts and culture. Unlike a census or other systematic data collection, the focus of historical artistic data is *not* on average individuals, but rather exceptional ones. Data typically only exist for famous artists, those who "made-it" and are observable in the present, often decades or centuries after their deaths. Third, the market for arts and culture is particularly opaque. Even today, an important proportion of works of art are sold by private dealers for prices that are never made public; the bottom-line of movie or theater producers, if private, are similarly privileged. Therefore, when cliometricians go in search of historical artistic data, they face two problems: locating data from years past and locating data generated by markets that are relatively small and where pricing is frequently private. Because of

the importance of data quality in this subfield, many contributions to the existing literature are significant in part because they provided new arts data that allowed for new insights.

One of the problems that economic historians of art struggle with the most is how to capture the *quality* of artistic output. However, before tackling quality questions, it is surprisingly difficult to even grasp the true *quantity* of art produced and consumed over time. For the visual arts, several sources are available. Auction data can, of course, capture large samples of art produced across several centuries; art from ancient times to the present can be sold at a contemporary auction. Several online databases – such as artnet and Blouin Art Index – provide this data for the past few decades; hard copy works like Gerald Reitlinger's three-volume opus *The Economics of Taste* (1961–1970) lists works sold at auction from 1760 to 1960 (although this data has been critiqued since publication, notably Guerzoni 1995). Other sources of the quantity of output over time include exhibition data like those used in Greenwald (2018a), museums increasingly accessible lists of their permanent collections, and catalogue raisonnés – complete lists of a given artist's production – assembled by art historians. Finally, a number of scholars have located sales data for historical markets. Many recent papers, in addition to making analytical contributions about art markets, have contributed significant new troves of data about historical artistic production. Recent examples of work that contributes exceptional art market data include Federico Etro and Laura Pagani's study of art markets in seventeenth-century Italy (Etro and Pagani 2012); Kim Oosterlinck, Christian Huemer, and Geraldine David's work about the multinational nineteenth-century art dealer Goupil & Cie (David et al. 2014); and Richard Goldthwaite's recent "economic biography" of an early Italian composer (Carter and Goldthwaite 2013).

Of course, scholars are not only interested in charting and understanding greater quantities of artistic output but also in tracking and understanding what drives higher *quality* creative work. Following standard economic theory, many scholars use prices to approximate the value of goods and services, and this includes artistic output. Within visual arts, this can be done rather conveniently with past results from art auctions and under the assumption that market prices reflect the true quality of a painting. First in a landmark article (Galenson and Weinberg 2001) and then in the book *Old Masters and Young Geniuses* (2007), economist David Galenson used auction prices as a proxy for the quality of an artwork. Focusing on nineteenth-century French artists, his arguments depended on auction prices decades after the creation of an artwork as a proxy for intrinsic quality. Other scholars working on the visual arts, notably Christiane Hellmanzik and Douglas Hodgson, have also used auction prices as indicative of quality (e.g., Hellmanzik 2010, or Hodgson 2011). There is, however, some bias in these approaches in so far as prices are affected by certain unobservable trends and fashions – whether contemporary or historical – and it can be difficult to account for these. Take, for example, the significant price increases of modern art since the turn of the twentieth century, especially in comparison with, say, Baroque art over the same time period. This is unlikely a reflection of higher perceived quality and artistic merit of modern art, but rather a shift in fashions (the taste for religious art is on the decline) and market trends (the

expectation of higher returns on modernism attracts further investment). While using auction data is useful for the study of artworks, it has its limitations outside visual arts and cannot be applied easily to music or literature. A composer's hand-written music manuscripts can be auctioned for a non-negligible amount, but here the link between quality and price is less clear. The manuscripts are historical artifacts, not the works of music themselves.

Another way of assessing the quality of an artist's oeuvre is to use contemporaneous expert opinion – namely, opinions formed at the time a cultural work was produced – to judge the quality and importance of a work of art. Helpfully, there are expert opinions published not only about visual arts but also about the performing arts, music, movies, and books. Economists have tried to gauge the ability of neutral experts – like critics and the awards committees – to pick "winners" in arts and culture, both financially and in terms of artistic reputation. In two 2003 articles, Victor Ginsburgh (Ginsburgh and van Ours 2003; Ginsburgh 2003) examines the role of expert opinions in a classical music competition, American movie awards, and international book prizes. He concludes that positive expert opinions or awards given shortly after the production of a work are usually correlated with economic success. However, these pronouncements do not correlate with the long-term survival or acclaim of the work of art or artist. In the visual arts, Kathryn Graddy specifically tested the correlation between quality rankings created by renowned seventeenth-century art critic Roger de Piles for artists active in his own period and the prices attained by those artists on the auction market over the subsequent three centuries. She finds that his highest-ranked artists attain both better financial returns on the market and art historical acclaim (Graddy 2013). It is therefore unclear how reliable contemporaneous expert opinions can be.

Therefore, many scholars choose to engage with *retrospective* expert opinion. In short, they use "best of" lists, comprehensive dictionaries of artists, writers, and musicians (e.g., Grove Art Online, Grove Music Online, and the Encyclopedia Britannica), and other retrospective compendia to gauge the importance of a given cultural producer and his or her output.[1] Using the length of entry about a producer in these compendia is a strategy to measure the significance and quality of a producer's contributions to his or her field. O'Hagan and Kelly (2005)[2] discuss this approach and the relevant literature in depth. They also developed their own selection algorithm using a column-inch method, in which they look at the space devoted to each artist in an art dictionary and gauge the quality of an artist by how much space is allocated to him or her. The underlying assumption that the length of entry correlates with an artist's importance finds support in empirical evidence. However, there are also drawbacks to this approach. First, artists who lived longer have naturally longer

[1]For example, in "The Dictionary of Composers and Their Music" (1979 & 1993), the two musicologists Gilder and Port provide information about who wrote what, and when. The dictionary is a recognized survey of the most influential classical compositions and served often as a source for composer's output (e.g., Borowiecki 2013).

[2]http://www.tandfonline.com/doi/abs/10.3200/HMTS.38.3.118-125

biographies, albeit longevity may also correlate with their fame; these individuals simply had more time to produce potentially famous work. Second, artists in early periods, especially prior to the eighteenth century, typically have shorter biographies because documentation and historical sources from this earlier period are less common and less well-preserved. Third, and perhaps most difficult to deal with, is the existence of biases to the home or target market country of the reference work; this issue persists even in the most prestigious publications. Therefore, building on selections of artists from just one reference work will create a bias that is inevitably directed toward the country of origin of the source and, as a result, one will over-represent artists of a particular nationality or gender and under-represent others – women artists, artists of color, and artists from the global south are chronically underrepresented.

To overcome these issues, one would ideally use a larger number of international dictionaries and pursue various selection algorithms, depending on the objectives of research. An example of this kind of extensive work is *Human Accomplishment* by Charles Murray (2003). In this book, the chapter "Excellence and Its Identification" describes how Murray selected the most prominent people in various fields – including visual arts, music, and literature – and outlines his comprehensive selection methods. His selections are based on numerous international dictionaries and reference works to mitigate the risk of country- or marketing-biases. Murray's lists of famous achievers have been often used to identify samples in other research projects (e.g., in visual arts, Hellmanzik 2010, or in music, Borowiecki 2013). Though not included in Murray's sources, other data sources that can be used to gauge the importance of an artist are the collections of museums and performance programs at operas, symphonies, and theaters, which are all determined through a combination of audience demand and the subjective choice of curators or artistic directors (Alfred Loewenberg's "Annals of Opera" [1978] is a reference work on notable performances, used by, among others, Giorcelli and Moser 2016).

Though applicable to many fields and a practical method for gauging an artist's quality and prominence, using data from encyclopedic lists of the collections of cultural institutions is vulnerable to significant sample bias. To even qualify for inclusion in these publications and institutions, an artist, composer, or author had to have made a meaningful contribution to the relevant arts canon. The sample of creatives presented in these publications is non-random (Gilder and Port 1993; Borowiecki 2013). Hence, one may ask: what about the lives and works of ordinary artists? There is no dictionary of mediocre artists or low-quality artworks. However, recent research has proposed ways to work around this obstacle. One approach is to obtain data from censuses. For example, Borowiecki (2018) pursues this method and collects data on thousands of average artists using the United States census as recorded in the Integrated Public Use Microdata Series database. The comprehensive decennial population census provides information from 1850 on occupational status, which can be used to identify various high art occupations, including artists (visual arts), authors (literary arts), musicians (music), and actors (performing arts). Working at the level of individual works, Greenwald (2018a, b) explores data about historic exhibitions that captures *all* the artists who participated in contemporary

art exhibitions at a series of venues around France and the United States during the nineteenth century. There are also scholars working with online searches for certain artists or artworks, or downloads of certain types of music from streaming platforms, in order to move closer to un-curated demand by art audiences (Borowiecki and O'Hagan 2012; Waldfogel 2014; Aguiar 2017).

As is common in other research areas in economic history, there is no one perfect data source for information about arts and culture. Survivor bias related to works of art and artists – in the visual arts, music, literature, and other areas – is particularly pronounced. Notable, artwork and artists dominate the available data. It is convenient to assert that famous artists have made very significant contributions to the arts and our cultural heritage, and therefore constitute a particularly appealing sample to look at. Famous artists are also the right sample if one is interested in understanding what drives outstanding creative output and genius. Though economic historians of arts and culture have conducted exhaustive archival research and come up with novel methods for capturing creative output, persistent sample bias is a reality that must be acknowledged and addressed in this subfield, perhaps more than in other cliometric disciplines.

Art Markets and Their Logic

Unsurprisingly, many works of economic history dealing with arts and culture address research questions related to *markets* for art and culture. Economic historians generally accept that the output of any good is a response to market forces; therefore, just like for other goods, scholars examine how these forces have changed over time to encourage the consumption and supply of creative output.

Art markets – particularly markets for visual arts – are peculiar in a number of ways. First, "value" in these markets is subjective. Judgment of the relative quality of cultural product depends on the combination of expert opinion and audience reception. Furthermore, judgments of a book, film, or other work of art's value can change significantly over time. Consider the example of Vincent van Gogh, whose work received little recognition during his lifetime. Despite producing approximately two thousand paintings, he sold very few of them while he was alive and died penniless at the age of 37. Today, his paintings are among the most valuable in the world. Largely because of these subjective valuations, there are significant asymmetries in information between suppliers and consumers in cultural markets. This creates ideal conditions for the presence of intermediaries – art dealers, record labels, movie studios, publishers, etc. – to mediate interactions between the cultural producers and their audiences. Later in this section, we will describe the growing literature dedicated to these intermediaries, their development, and their behaviors. This section often focuses on the market for visual arts, because it is a large literature and engages with research questions related to markets for other art forms; however, we also make an effort to discuss studies of markets for other cultural goods.

When studying art markets, scholars' primary research questions generally fall into one of three categories. The first category is the description of a particular

historically significant moment in the history of art, including moments of significant transformation in market structure. The second category is the performance of art as an asset over time, which depends in large part on the structure of the art market. The third category consists of efforts to understand how markets for art and culture have been formed by the broader socioeconomic context in which they exist, and how, in turn, these markets have had long-term effects on communities where they are present. This section describes studies in each of these areas.

There are several major contributions to the economic history of arts and culture that provide specific data and historical detail about significant moments in the history of art markets. A foundational example of this type of work is the literature about art markets in Golden Age Holland. Economist John Michael Montias's examinations of seventeenth-century Dutch painting (Montias 1982, 1989) provided extensive price data about paintings produced during this period and social and economic detail about the communities where art was produced and consumed. His scholarship was the first in a series of economic history works in this area conducted by both art historians and economists, included Filip Vermeylen, Neil De Marchi, and Hans Van Miegroet (e.g., Vermeylen 2003; De Marchi and Van Miegroet 1994). Other scholars have tackled the Italian Renaissance and Baroque periods, like Richard Goldthwaite's work on Renaissance Florence (Goldthwaite 1993) and Federico Etro and Laura Pagani's study of art markets in seventeenth-century Italy (Etro and Pagani 2012; Etro 2018), which presents exceptional data from contracts between artists and patrons. A recent notable example of this kind of work dealing with another period and geography is Kim Oosterlinck, Geraldine David, and Jeroen Euwe's research about the art market in in German-occupied Europe during World War II. Using novel data, they show that demand for art was booming during the war and that this demand is attributable to a wartime need for portable and saleable assets (Oosterlinck 2017; David and Oosterlinck 2015; Euwe 2007). Across this body of work, scholars seek to understand the particularities of supply (who, why, and how people are selling art) and demand (who purchasers were and why they decided to acquire artworks) at clearly defined moments in time. There is research with similar goals in other areas of the arts. For film markets, for example, Pokorny and Sedgwick (2010) chart the development of the US film markets from 1929 to 1999, focusing on a comparison of the 1930s and 1990s. In particular, they show that similar levels of profit variability existed in both decades; however, while in the 1930s the main source of profits was from low- to mid-budget films, in the 1990s the main source of profits was high-budget films.

Other research seeks to understand *transitions* from one art market structure to another. Frederic Scherer, for example, describes the evolution of European music markets from the 18th to the 19th centuries. In particular, he examines how composers transitioned from being employed by their patrons to being entrepreneurs seeking support and income from a variety of sources (Scherer 2004). In the 1960s, William Baumol and William Bowen used a blend of historical and contemporary data to examine how the costs of producing and supplying the performing arts had increased over the twentieth-century. They attributed this increase to large fixed costs – including growing costs of labor – for the performance of an opera, ballet, or

piece of theater and termed the problem a "cost disease." In turn, they argued that this cost pressure influences demand either by making the ticket price for the performing arts out of reach of many potential audience members or demanding that the state or private foundations provide significant subsidies for productions. This phenomenon damages the economic viability of the performing arts (Baumol and Bowen 1965), although this theory has been debated and questioned since its presentation (Cowen 2000).

Baumol and Bowen were interested in an on-going transformation in the market for the performing arts. Other scholars have produced purely retrospective work that focus on past changes. One of the most studied transitions in the structure of art markets took place in France during the nineteenth century. In *Canvases and Careers: Institutional Change in the French Painting World*, sociologist Harrison C. White and art historian Cynthia A. White examine the transition in late nineteenth-century France from a system of centralized state support for the visual arts to what they call the "dealer-critic" system, a decentralized market-based system for the exhibition and sale of contemporary art (White and White 1993). Many works have followed the Whites' lead in exploring and explaining why the market for arts and culture has changed over time. David Galenson and several co-authors have published a number of papers in direct response to *Canvases and Careers.* A paper published with Robert Jensen argues that the fracturing of a centralized state-sponsored exhibition system in nineteenth-century France can be partially attributed to artists' entrepreneurial impulses to exhibit and sell in their own group shows (Galeson and Jensen 2002). In another article, Galenson and Bruce Weinberg describe how the new decentralized market-based system encouraged and rewarded innovation among artists and helped foster the advent of modern art (Galenson and Weinberg 2001).

We describe this literature about nineteenth-century art markets in detail not only because it is much-cited in the field but also because it introduces an important topic in the economic history of the arts: the changing importance of certifiers and intermediaries in art markets over time. Under the decentralized "dealer-critic" system, the number of individuals participating in the market grew. This created a growing role for market intermediaries like art dealers and auction houses, including Christies, Sothebys, and Drouout. Some of the most studied intermediaries are contemporary art dealers. Works including Bystryn (1978) and Fitzgerald (1995) demonstrate that dealers have become essential for the economic success – and perhaps the art historical staying power – of modern artists. The most recent major contribution to this area, Olav Velthuis' *Talking Prices* (2005), presents compelling ethnographic research suggesting how pricing in the primary art market is fluid and depends on the ability of particular dealers to consistently adjust to the immediate conditions of supply and demand.

The role of market intermediaries is the focus of studies of other markets for cultural goods, including books, journalism, and music. In book history, publishers, printers, and booksellers are a particular focus (St Clair 2004; Dittmar 2011; Finkelstein and McCleery 2012). Scholars working on the economic history of the music industry have examined technologically driven challenges posed to traditional

intermediaries like record companies and the musicians they represent. As the contemporary music market changes radically (the initial dominance and then decline of records and CDs, the arrival of legal and illegal downloads, and now the rise of streaming), what counts as the "history" of the changing music market has been accelerated. Debates over intellectual property protection and its effect on the quality, quantity, and format of music produced have been central to the literature about intermediaries in music. Many articles (e.g., Oberholzer-Gee and Strumpf 2007 and Waldfogel 2012) have found that illegal downloading on services like Napster had little to no effect on the observed decline in record sales in the beginning of the twenty-first century and did not lead to deteriorating quality of music produced during the same period. As music piracy becomes less common – particularly in Europe and North America – scholars have instead turned towards assessing the effects of legal digital music streaming (Aguiar and Martens 2016). This literature about the music industry is directly related to another recurring theme in the economic history of the arts: the importance of copyright and protection of intellectual property. We will address this topic in detail in section "Capturing and Fueling Creativity."

As the art market and number of intermediaries grew from the nineteenth-century forward, buyers began to invest in art for financial as well as aesthetic reasons. Art became a valuable and easily saleable asset (Velthuis 2011). There is a large literature dedicated to judging the performance of art as an asset over the long run. In general, these papers find that art performs poorly as an investment. Useful surveys of these findings are conducted by Frey and Eichenberger (1995), Goetzmann et al. (2011) and Renneboog and Spaenjers (2013). The particularities of markets for art – persistent asymmetries of information and the influence of intermediaries like auction houses and art dealers – have a profound effect on art's performance as an asset. David, Oosterlinck, and Szafarz (2013) and a number of papers by Orley Ashenfelter, Kathryn Graddy, and Christophe Spaenjers examine inefficiencies in auctions for art and other luxury goods and how these inefficiencies negatively impact the performance of art as an asset (Graddy and Ashenfelter 2011a, b; Spaenjers et al. 2015).

The third and final group of studies of art markets examines the interaction between art markets and the broader social and economic context in which they exist. The most cited contribution in this part of the literature is from sociology. In *Distinction: A Social Critique of the Judgment of Taste*, famed sociologist Pierre Bourdieu asserts that "there is an economy of cultural goods [with] a specific logic" (Bourdieu 1984: xxv).[3] This logic is determined by the education and upbringing – the *habitus* – of the participants in the cultural economy. Bourdieu writes: "One can say that the capacity to see (*voir*) is a function of the knowledge (*savoir*). . .A work of art has meaning and interest only for someone who possesses the cultural competence, that is the code into which it is encoded. . .A beholder who lacks the specific

[3]Pierre Bourdieu, *Distinction: A Social Critique of the Judgement of Taste*, trans. Richard Nice (Oxford: Routledge, 1984), xxv.

code feels lost in a chaos of sounds and rhythms, colors and lines, without rhyme or reason."[4] Because transacting in this cultural economy demands this encoding, Bourdieu concludes that class may be not only expressed but also defined and transmitted to future generations through the consumption of cultural goods. Finally, he argues that formation of class distinction on the basis of cultural tastes is particularly effective. While upwardly mobile people may be able to accumulate financial capital in one lifetime, cultural understanding is the product of "total, early, imperceptible learning, performed within the family from the earliest days of life and extended by a scholastic learning which presupposes and completes it. . ..[B] ourgeois families hand [this] down to their offspring as if it were an heirloom" (Bourdieu 1984: 59). Since its publication, *Distinction* has established as given the assertion that cultural activities, tastes, and institutions are potent sites of class formation. The effects of this assertion have been far-ranging across fields that deal with the arts, including in cultural economics and the economic history of the arts.

Bourdieu used contemporary survey data and quantitative methods to support his assertions. While cultural economists working on contemporary issues have also used survey data to study connections between artistic consumption and socioeconomic background (e.g., Ateca-Amestoy 2008), economic historians have needed to look for other sources that allow them to study this interaction in the past. Therefore, this area of the economic history literature often focuses on how demand for art is related to and enshrined in bricks-and-mortar cultural institutions. This scholarship asks: how is the particular demand for or supply of art encouraged by and preserved in museums, operas, and symphonies? Furthermore, how do these institutions continue to form art markets after their founding?

In the visual arts, there have been several studies of the relationship between demand for art and the founding of art museums. Though not an economic historian, sociologist Paul Dimaggio has made major contributions to the economic histories of museums and other arts institutions. Building on Bourdieu's work, Dimaggio examines nineteenth-century American patrons' socially motivated founding of new museums, operas, and symphonies. He argues that these institutions changed how cultural output was delivered, and put a solid fence around what was considered "high culture" and what was "popular culture" (Dimaggio 1982). These categories of art and their distribution remain separated from one another today, and an artist's inclusion in a major museum collection (or the performance of a composer's work at a major opera house or symphony) remains an important certifier of an artwork's status as "high culture" and of its quality. Bruno S. Frey and several co-authors have made important contributions that document this phenomenon both in historical settings and today (e.g. Frey 2013).

Currently, the most active area of inquiry into the effects of bricks-and-mortar arts institutions is related to how the founding of an institution can form the supply of, and demands for, art in a given place for centuries *after* the institution was founded.

[4]Ibid.

Readers will note that almost all of the literature described here deals with the arts in a particular time and place. Even financial history papers that aim to measure the returns on art as a globally traded asset must concede to the importance of local and idiosyncratic interventions of intermediaries like those described in Velthuis (2005). Artistic output is not equivalent to a commodity like wheat or coal. Instead, trading in cultural goods is often done face-to-face, depends on personal relationships and takes place in dedicated physical institutions. These geographical factors are explored in greater detail in the next section, "Geography and Art."

Geography and Art

In the economic history of art, scholars often return to the topic of geographic clustering – both of consumers and suppliers of art. They have observed that across historical eras that the most active audiences and the best-known artists tend to be centered in particular cities. This section describes the work demonstrating this phenomenon and explores the effects of clustering on artistic output. It begins with studies of demand for arts and culture, and then transitions to research focusing primarily on supply.

Two recent contributions have discussed the persistence of demand for culture in cities with a long legacy of supporting the arts. In a paper titled "Phantom of the Opera" (Falck et al. 2011), Olivier Falck, Michael Fritsch and Stephan Heblich use the location of Baroque opera houses in Germany as a quasi-natural experiment to test whether or not high human-capital employees are drawn to cities with cultural amenities. They find that the presence of a Baroque opera house increases the current presence of educated workers. This suggests that there is persistence of "cultured" activities – and the reputation of being cultured – in certain cities. There have been several comments and challenges to this paper, but a recent replication study (Bauer et al. 2015) did find the same effect for Baroque opera houses; although there were also positive effects on the share of high-capital workers for historical brothels and breweries. The "Phantom of the Opera" paper did not set out to gauge demand for the arts; instead, it sought to understand why highly skilled workers choose to live in certain places, and how this clustering affects the growth path of a certain city and region. In contrast, Borowiecki (2015b) does set out to understand the persistence of support for the arts and demand for cultural goods in certain cities in Italy. He finds that cities and provinces that had high-levels of cultural activity in the past – measured by the number of active composers in the area – continue to have more concerts and operas performed in the present. Furthermore, residents still spend relatively more money on high-culture activities as compared to other entertainment, like sporting events. Demand for and support of the arts, once established in a given city, appears to persist.

Measuring clusters of demand for the arts is a new and small literature when compared with scholarship that aims to understand the geographic distribution of artists. One of the earliest quantitative studies on this topic dates back to the turn of the twentieth century, when Gustav Michaud explored the spatial spread of artists

and intellectuals in the United States and described his findings in the article "The Brain of the Nation," published by the Century Magazine in 1905 (Michaud 1905). This area of research has received considerable attention, especially in the last 20 years as many countries transition to a service-based economy. Studies of agglomeration economies have established as given that economic activity is concentrated geographically. However, geographic clustering of successful artists is remarkably more concentrated, and it has often been observed that throughout history the global population of prominent artists is located in a handful of cities. For example, Mitchell (2016) documents the extent of geographic clustering of a sample of 370 significant poets and writers in the UK and Ireland born 1700–1925. She identifies the Greater London Area as an unrivalled cluster, with 79 (or 21%) of writers born within this region, and during its peak more than 50% of all authors working there. Dublin, Edinburgh, Oxford, and Cambridge emerge as the only other cities that see minor clustering of authors at any point in the sample. Other major contributions that document this intense clustering include Hall (2006), Murray (2003), and Schich et al. (2014).

At Trinity College Dublin, John O'Hagan, now Professor Emeritus, leads a research group specifically dedicated to the study of creative clustering. This group has produced a series of projects compiling detailed databases that cover the lifetime migration histories of hundreds, or even thousands, of famous artists, ranging from visual arts, to music, to literary arts. Research by this group and by Maximillian Schich show that the clustering of artists is not driven primarily by large numbers of births of artists in given cities, but rather by patterns of migration. Schich et al. (2014) use a database of 150,000 notable people – including artists – to show that the median distance between birth and death place for these people has remained largely constant from the fourteenth to the twenty-first century.

O'Hagan and Hellmanzik (2008) specifically examine famous visual artists' migration and clustering patterns for four periods (based on their date of birth): Renaissance Italy, Europe in the first half of the nineteenth century, and the Western world in general for the periods 1850–1899 and 1900–1949. These data support Schich's conclusions that famous artists clustered at a remarkably high level in all periods. Florence and Rome dominated in Renaissance Italy, with significant clustering because of artists' birthplaces and domestic migration. Paris and London witnessed a marked clustering of artists born in the first half of the nineteenth century. These two cities were the main work locations for 55 of the 72 artists studied. The French capital continued to dominate among artists born in the second half of the nineteenth century, while artists born in the first half of the twentieth century clustered in New York City, with all prominent American artists clustering there. The geographic distribution of music composers is studied by O'Hagan and Borowiecki (2010), who examine annual migration and clustering patterns of 522 important composers of the last 800 years. In line with visual artists, it is shown that Paris has been a major center for composers. The concentration of composers in this one city perhaps reflects the general prominence of Paris as a cultural city. Some composers were born in the French capital, but most migrated to it. Table 1 outlines the broad pattern of migration of prominent composers over the

Table 1 Type of movement of famous music composers by century

	Movement						
	None		Internal		External		All
Century of Birth	Total	Relative	Total	Relative	Total	Relative	Total
12th	2	0.50	2	0.50	0	0.00	4
13th	0	0.00	2	0.50	2	0.50	4
14th	2	0.18	8	0.73	1	0.09	11
15th	0	0.00	31	0.61	20	0.39	51
16th	14	0.13	66	0.63	24	0.23	104
17th	14	0.17	52	0.62	18	0.21	84
18th	16	0.17	41	0.44	36	0.39	93
19th	27	0.18	88	0.59	34	0.23	149
20th	2	0.09	16	0.73	4	0.18	22
All	77	0.15	306	0.59	139	0.27	522

Source: O'Hagan and Borowiecki (2010)

centuries. O'Hagan and Borowiecki find that 86% of all prominent composers spent the longest period of their working lives away from their place of birth, 59% migrated to a city within the country of their birth, while the remaining 27% migrated to work in another country.

These studies show a significant rate of migration going back several centuries for artists, musicians, and other creative people. This suggests that creatives existed in an integrated global world many centuries before average people regularly migrated towards urban clusters. Importantly, as the markets for art and culture were not (and are still not) globally integrated, the suppliers of art have had to travel to cultural centers where art markets, institutions, consumers, and other producers are located (Florida 2014).

These migration patterns appear to hold for more than just notable artists. In an effort to address the issue of sample bias in the subfield, Borowiecki (2018) examines differences in the clustering intensity and location choice between famous and average artists in the United States from 1850 to the present. This is done using both census records to cover average artists and data from art dictionaries listing famous artists. He shows that the geographic spread of the census artists (the "average" artists) is greater, which implies a lower clustering intensity – albeit still very noticeable – in comparison with the famous creatives. This reconfirms the previous statements that extraordinary achievers concentrate more than average individuals. It is interesting to observe that for both populations of artists the same dominant clusters emerge: New York City, followed by Boston, Chicago, Los Angeles, and San Francisco. Borowiecki (2018) also shows interesting differences across artistic domains. However, some discipline-specific clusters emerge. New Orleans has a very high concentration of births of musicians, while Seattle constitutes an important work location for literary artists.

Often, businesses that help sell the creative output of artists also cluster, usually near the artists themselves. Several studies of the locations of art dealers over time

show that like artists themselves, they cluster intensively in the same place – even in the same neighborhood in a culturally active city. Examples of this geographic work on art dealers includes art historians Pamela Fletcher and Anne Helmreich's study of the nineteenth-century London art market (Fletcher and Helmreich 2012) and the mapping of the location of Parisian art dealers from 1815 to 1955 created by the Artl@s research group at the École Normale Supériuere (Saint-Raymond et al. 2015). Significant clusters of intermediaries have also emerged in theater and cinema. Broadway in New York City and the West End in London are dense clusters of theater activity. Organizations at the heart of these clusters – and organizations in the same city or metropolitan area but outside the principal cluster – have produced significant innovations in theater. (Castañer and Campos 2002 provides a comprehensive overview of this literature.) Hollywood is a globally dominant cluster in the film industry. Several articles have examined how this cluster emerged and how it continues to propagate its dominance. Research has particularly focused on the large budgets of early Hollywood films and the ability of films made in southern California in the early to mid-twentieth century to penetrate foreign markets (Miskell 2016; Bakker 2008, Sedgwick 2000). Just as art dealers cluster near artists, Hollywood and Broadway are two clusters that unite both the creators of the art themselves – actors, directors, writers, etc. – and the individuals and organizations who provide the funding and distribution necessary to create a movie or theatrical production (Caves 2000; Scott 2004).

Observing geographic clustering is an important contribution to economic history. However, not surprisingly, economists and economic historians are not content with just making this observation. They want to understand why clustering fosters creativity. For policy making in today's service-driven economy, scholars seek to understand how governments and institutions can fuel innovation and creative activity, either by fostering clusters or with other policy interventions. The next section addresses these questions.

Capturing and Fueling Creativity

Is it possible to understand what drives creative output and how to encourage it? While scholarship surveyed in the first three sections of this article certainly touches on this question, this section presents papers that make permutations of this research question their primary focus. Therefore, this section deals more with the work of scholars who are interested in using historical data and case studies to examine how employment and economic production in the "creative industries" – a fluid category including industries ranging from visual arts and fashion to advertising, journalism and software design – have developed over time (Towse 2010).

Importantly, this work aims to understand how creativity and creative industries develop. Creative and arts sectors are seen as a "key ingredient for job creation, innovation and trade" (UNCTAD 2010) and are believed to constitute opportunities for depressed cities and developing countries to participate in high-growth areas of

the world economy. Typically, scholars working from this policy perspective describe their research area as "cultural economics." In this literature, creativity is sometimes modeled as a result of rational decision-making (Frey 2013) or as a function of some objective and quantifiable factors, such as general education or experience (Bryant and Throsby 2006). However, the most recent contributions have looked at how specific extraordinary individuals make their discoveries and produce creative output due to peer effects and emotional drivers, how creativity can be induced by an increased demand for innovation, and how intellectual property protection can foster creativity.

Famous artists have been shown to exhibit remarkable clustering patterns. With these observations in mind, Hellmanzik (2010) provides an important study on the existence of location premiums by exploiting a sample of prominent modern artists born between 1850 and 1945. She combines auction data with records about whether and when an artist worked in Paris or New York, the two main cluster locations of that period. The findings suggest that paintings created in Paris or New York have been valued higher by 11% and 43%, respectively. Furthermore, artists working in one of these two cluster locations are shown to reach a peak in the age–price profile of their work significantly earlier than artists working elsewhere. Similar results are presented by Mitchell (2016) for writers. Mitchell finds that an author becomes more productive by around 11% each year when residing in London.

When teasing out the causal effects of clustering, scholars must ask whether geographic clusters attract the most creative artists, or whether artists who cluster are more productive because of positive externalities associated with cluster locations. In other words, is self-selection driving the empirical evidence on better performance in geographic clusters, or does a clustering benefit exist? This question – and escaping the endogeneity problems one faces in answering it – is of considerable policy importance not only for the arts but also for other sectors (see Rosenthal and Strange 2004). Borowiecki (2013) uses data for a global sample of 116 prominent music composers born between 1750 and 1899 to them to answer this question. A historical approach enables him to exploit the variation in the geographic distance between a composer's birthplace and a geographic cluster as an exogenous source of clustering, and thus to credibly assert that the association between clustering and productivity is a causal relationship rather than simply a correlation. Borowiecki finds that geographic clustering increases creativity: composers were writing around one additional influential work every 3 years they spent in a cluster. Drilling down further into the dynamics of notable artistic clusters provides some clues as to how clusters encourage higher-quality artistic production. Borowiecki (2015a) develops a simple theoretical framework explaining the trade-off between agglomeration economies (peer effects) and diseconomies (peer crowding), which suggests that the productivity gain due to the presence of peers is characterized by an inverted U-shaped relationship and eventually decreases if the peer group size becomes very large. These theoretical predictions are supported by data for music composers: a composer was about 10% more productive when an additional prominent composer was located in the same city. However, the effect is nonlinear and may begin to decrease for very large numbers of peers.

What drives the benefits from clustering with peers? The literature has provided three distinct answers about the internal dynamics of a cluster: knowledge exchange between peers in the same field, interaction between individuals from diverse creative fields, and competition between peers. Geographic proximity facilitates spillover effects between individuals in the same field (Marshall 1890). Therefore, in cities with a particularly high concentration of artists, synergies and spillovers may positively impact the individual's ability to innovate. Porter (1996) suggests that local competition in specialized, geographically concentrated economic activities may constitute a significant stimulus for growth. The competitive working environment experienced by artists when they are concentrated forces innovation. One can find anecdotal evidence for this argument. In 1778, Mozart was in Paris, the most crowded creative market he ever lived in. His productivity peaked in this year, and he wrote 19 influential compositions, which is three times higher than his annual work-life average of around 6.6 compositions.[5]

The literature has suggested further factors external to clusters that can both foster the creation of a creative cluster and innovation more generally. These factors include specific market demand for innovation, differing levels of intellectual property protection, and political competition for creative talent. In their influential study about modern art, Galenson and Weinberg (2001) suggest a significant shift in demand for innovation in art. The authors explore how an artists' quality of artworks, approximated with auction price data, changed over his or her lifetime and show that the peak age occurred much earlier for later cohorts, who were exposed to increased demand for innovation. It was therefore not by chance that Paul Cézanne and Pablo Picasso created their most significant works within 1 year of each other, although they were born more than 40 years apart.

Apart from market conditions in a given time and place, many scholars have discovered that copyright protection over time and across countries and regions has created different creative environments. Ruth Towse has contributed extensively to this literature and has examined the effects of copyright protection on a variety of creative industries and conflicts between artists and intermediaries over ownership of intellectual property (Towse 2001). Her work deals primarily with contemporary policy and economic conditions. More historical in her focus, Petra Moser has researched the effects of copyright and patent protection across a variety of industries, geographies, and eras (e.g., Moser 2011, 2013). Recently, she has worked with artistic data – specifically book publishing in nineteenth-century Britain (Li et al. 2017) and the effect of Napoleonic rule on Italian operas (Giorcelli and Moser 2016) – to examine the effects of copyright on innovation. Giorcelli and Moser exploit variation in the adoption of copyrights due to the timing of Napoleon's military victories in Italy. The authors are therefore able to estimate the causal effects of copyrights on creativity and find that basic levels of copyright protection increased not only the quantity of creative output but also its quality. These creativity gains are

[5]The category of "influential compositions" is recorded by the the musicologists (experts' selection) in Gilder and Port (1993).

explained by the fact that copyrights reward the greater composing effort necessary to produce high-quality work.

Beyond intellectual property rules, other political and policy factors can have an important effect on artistic output. Vaubel (2005) develops the hypothesis that competition among neighboring states may favor cultural innovation. This hypothesis is then backed up with the empirical observations that European instrumental music had its breakthrough during the Baroque era and that the most famous composers came from the two countries characterized by the highest degree of political fragmentation: Italy and Germany. Vaubel then measures the average duration of employment as a proxy for competition on the demand side and shows that famous Italian and German composers of the Baroque period changed their employers significantly more often than their French and British counterparts did. These insights suggest not only that political fragmentation has promoted musical composition and encouraged quality but also stimulated the mobility of composers.

While much of the research about creativity has focused on the effects of clustering and specific policies, there is a growing economic history literature that explores how biological factors drive artistic creativity. Thinkers since Aristotle have theorized that creativity and emotional state are linked. Other disciplines – notably applied psychology literature – have probed this link, along with the apparent correlation between famous creative people and psychiatric disorders like depression and bipolar disorder. Economic historians of the arts have recently engaged with this literature by looking at the letters of famous artists and seeing how their emotional state correlates with their creative output. Borowiecki (2017), for example, uses textual analysis to calculate the extent of positive and negative emotions expressed in a large number of letters written by Wolfgang Amadeus Mozart, Ludwig van Beethoven, and Franz Liszt. This allows him to create well-being indices capturing emotional states throughout each man's lifetime. He then shows that negative emotions have a causal impact that leads to increased creativity, as measured by the number of high-quality compositions written.

While it is clear that clustering drives increased creative output, this greater output may have come at a cost for artists' physical and emotional well-being. Fierce competition between peers anecdotally led to depression and nervous breakdowns, as was the case for Maurice Ravel, who was diagnosed with neurasthenia in 1912 immediately after the failure of his ballet "Daphnis et Chloé." His condition was presumably aggrevated by an exceptional performance of his fellow Frenchman and competitor Claude Debussy's "Prelude to the Afternoon of a Faun," also performed in Paris just 10 days earlier. With anecdotes like this one in mind, Borowiecki and Kavetsos (2015) argue that the concentration of talent is likely to have adverse effects in terms of health and well-being. They attribute these outcomes to the continuous mental strain individuals go through in order to achieve their aspirations, which become more intensified in settings where one's peers thrive. Borowiecki and Kavetsos approximate for peer competition in various ways and suggest that it reduces composers' longevity. For example, all else equal, a 1% increase in the number of composers located in the same area and time reduces composer longevity by about 7.2 weeks. Essentially, clustering leads to greater

quantity and quality of creative output, but this can come at a cost for one's wellbeing. These findings are thought-provoking, especially if one considers that the first fundamental theorem of welfare economics also argues that competition is indispensable in producing Pareto-optimal outcomes.

Conclusion

Economic historians have rarely turned their attention to the arts. Often, when they did decide to study the arts, it was a mid-career decision driven not primarily by a belief in the centrality of arts and culture as topic in economic history but because they had a personal love for art, music, or another cultural good. These later career engagements created *excellent* scholarship – and there are exceptions to the characterization of scholars in this subfield starting midcareer. (In fact, the co-authors of the underlying chapter are just such an exception.) However, the fact remains that studying the economic history of the arts has long been a secondary topic in our field.

Happily, this is now changing. This change has been largely driven by the shifting composition of developed economies. Developed economies are now dominated by services – including the ever-growing category of "arts, entertainment and recreation".[6] The continued growth of the arts sector and other service industries demands creativity and innovation; this innovation often emerges from clusters of creative people, whether it is in Silicon Valley, a major metropolis like Paris, or in smaller clusters around universities and other cultural institutions. Today's economic dynamism therefore shares many common features with artistic output over the course of history. Understanding what fosters and drives creative achievement means understanding what fosters and drives economic success today. For this reason, the economic history of art and culture has become more important in the broader field and is poised to become increasingly important.

Beyond applications to contemporary economics and policy questions, cliometric approaches to the study of the arts are having a growing impact on other humanities fields that have typically resisted quantitative research. The rise of the digital humanities – first in the study of literature and now in art history, musicology, and other fields – has made humanities scholars more receptive to the use of quantitative methods (Jockers 2013; Moretti 2005; Fletcher and Helmreich 2012). However, literary scholars, art historians, and other humanities researchers often do not have the prerequisite statistical and computational training to compile and analyze quantitative evidence. This creates a golden opportunity for cliometricians to contribute to these other fields and use quantitative methods to address long-standing research questions in these areas. Consider the following example: cliometricians have documented the geography of creative clusters and explained how these have evolved over time and across various regions of the world. It could certainly be

[6]These numbers for the United States are available from Federal Reserve Economics Data (FRED). URL: https://fred.stlouisfed.org/series/CES7071000001

argued that art and music historians have been aware that Paris was an important center for the arts. However, there would quite likely be less agreement on how important it has been, the timing of its prime, and how it compared with other major cities. In other words, the extent of the dominance of Paris has not yet been quantified, nor has it been compared across creative domains. In this way, cliometrics provides a representative, objective, and robust measurement of the importance of cities, which can later be used within the humanities in various contexts.

Unfortunately, cliometricians have not yet seized this opportunity for collaboration. Practitioners in the digital humanities have engaged mostly with computer scientists and statisticians to implement tools such as 3-D mapping, algorithmic literary analysis, and linguistic patterns in textual corpora. Economists and economic historians should take this opportunity to have an impact on another field. Early efforts include the "Genius for Sale!" conferences, organized in 2014 at the University of Oxford and 2016 at Brandeis University.[7] These conferences invited scholars working both in the humanities and social sciences to present on topics related to the arts in their broader economic contexts. We need to create more collaborations like these that may, in turn, result in important interdisciplinary publications.

Cross-References

- ▶ History of Cliometrics
- ▶ Innovation in Historical Perspective
- ▶ Institutions
- ▶ The Cliometric Study of Innovations

References

Aguiar L (2017) Let the music play? Free streaming and its effects on digital music consumption. Information Economics and Policy, 41:1–14

Aguiar L, Martens B (2016) Digital music consumption on the internet: evidence from clickstream data. Inf Econ Policy 34:27–43

Ateca-Amestoy V (2008) Determining heterogeneous behavior for theater attendance. J Cult Econ 32(2):127–151

Artnet Auctions. http://www.artnet.com

Bakker G (2008) Entertainment industrialised: the emergence of the international film industry. Cambridge University Press, Cambridge, UK, pp 1890–1940

Bauer TK, Breidenbach P, Schmidt CM (2015) "Phantom of the Opera" or "Sex and the city?" historical amenities as sources of exogenous variation. Labour Econ 37:93–98

Baumol W, Bowen W (1965) On the performing arts: the anatomy of their economic problems. Am Econ Rev 55(1/2):495–502

Blouin Art Sales Index. http://www.blouinartsalesindex.com/site/app.ai

Borowiecki KJ (2013) Geographic clustering and productivity: an instrumental variable approach for classical composers. J Urban Econ 73(1):94–110

[7]The conference was organized by one of the authors of the underlying chapter and attended by the other, and if it was not for that event, this chapter would probably not have been written.

Borowiecki KJ (2015a) Agglomeration economies in classical music. Pap Reg Sci 94(3):443–468
Borowiecki KJ (2015b) Historical origins of cultural supply in Italy. Oxf Econ Pap 67(3):781–805
Borowiecki KJ (2017) How are you, my dearest Mozart? Well-being and creativity of three famous composers based on their letters. Rev Econ Stat 99(4):591–605
Borowiecki KJ (2018) The origins of creativity: the case of the arts in the US since 1850. SDU Working Paper
Borowiecki KJ, Kavetsos G (2015) In fatal pursuit of immortal fame: peer competition and early mortality of music composers. Soc Sci Med 134:30–42
Borowiecki KJ, O'Hagan JW (2012) Historical patterns based on automatically extracted data: the case of classical composers. Hist Soc Res 37(2):298–314
Bourdieu P (1984) Distinction: a social critique of the judgement of taste. Harvard University Press, Cambridge, MA
Bryant WDA, Throsby D (2006) Creativity and the behavior of artists. In: Ginsburg VA, Throsby D (eds) Handbook of the economics of art and culture. Elsevier, Amsterdam, pp 507–529
Bystryn M (1978) Art galleries as gatekeepers: the case of the abstract expressionists. Soc Res 45(2):390–408
Carter T, Goldthwaite R (2013) Orpheus in the marketplace. Harvard University Press, Cambridge, MA
Castañer X, Campos L (2002) The determinants of artistic innovation: bringing in the role of organizations. J Cult Econ 26(1):29–52
Caves RE (2000) Creative industries: contracts between art and commerce. Harvard University Press, Cambridge, MA
Cowen T (2000) In Praise of commercial culture. Harvard University Press, Cambridge, MA
David G, Oosterlinck K (2015) War, monetary reforms and the Belgian art market, 1945–1951. Finan Hist Rev 22(2):157–177
David G, Oosterlinck K, Szafarz A (2013) Art market inefficiency. Econ Lett 121(1):23–25
David G, Huemer C, Oosterlinck K (2014) Art dealers' strategy: the case of Goupil, Boussod & Valadon from 1860 to 1914. Working Paper, Yale School of Management. 19 January, 2014.
De Marchi N, Van Miegroet HJ (1994) Art, value, and market practices in the Netherlands in the seventeenth century. Art Bull 76(3)
Dimaggio P (1982) Cultural entrepreneurship in nineteenth-century Boston: the creation of an organizational base for high culture in America. Media Cult Soc 4(4):33–50
Dittmar JE (2011) Information technology and economic change: the impact of the printing press. Q J Econ 126(3):1133–1172
Encyclopaedia Britannica (2014) http://www.britannica.com
Etro F (2018) The economics of Renaissance art. The Journal of Economic History, 78(2):500–538
Etro F, Pagani L (2012) The market for paintings in Italy during the seventeenth century. J Econ Hist 72(02):423–447. https://doi.org/10.1017/S0022050712000083
Euwe J (2007) De Nederlandse kunstmarkt: 1940–1945. Boom, Amsterdam
Falck O, Fritsch M, Heblich S (2011) The phantom of the opera: cultural amenities, human capital, and regional economic growth. Labour Econ 18(6):755–766
Federal Reserve Economics Data. https://fred.stlouisfed.org
Finkelstein D, McCleery A (2012) Introduction to book history, 2nd edn. Routledge, London
Fitzgerald M (1995) Making modernism: Picasso and the creation of the market for twentieth-century art. University of California Press, Berkeley
Fletcher P, Helmreich A (2012) Local/global: mapping nineteenth-century London's art market. Nineteenth-Century Art Worldwide 11(03). http://www.19thc-artworldwide.org/autumn12/fletcher-helmreich-mapping-the-london-art-market
Florida RL (2014) The rise of the creative class: revisited. Basic Books, New York
Frey BS (2013) Arts & economics: analysis & cultural policy. Springer Science & Business Media
Frey BS, Eichenberger R (1995) On the return of art investment return analyses. J Cult Econ 19(3):207–220
Galenson DW, Jensen R (2002) Careers and canvases the rise of the market for modern art in the nineteenth century. National Bureau of Economic Research, Cambridge, MA.
Galenson DW, Weinberg BA (2001) Creating modern art: the changing careers of painters in France from impressionism to cubism. Am Econ Rev 91(4):1063–1071

Galenson DW (2007) Old masters and young geniuses: the two life cycles of artistic creativity. Princeton University Press, Princeton, p 2007
Gilder E, Port J (1993) The dictionary of composers and their music. Wings Press, San Antonio
Ginsburgh V (2003) Awards, success and aesthetic quality in the arts. J Econ Perspect 17(2):99–111
Ginsburgh VA, Van Ours JC (2003) Expert opinion and compensation: evidence from a musical competition. Am Econ Rev 93(1):289–296
Giorcelli M, Moser P (2016) Copyright and creativity: evidence from Italian operas. Working Paper, Social Science Research Network.
Goetzmann WN, Renneboog L, Spaenjers C (2011) Art and money. Am Econ Rev 101(3):222–226
Graddy K (2013) Taste endures! The rankings of Roger de Piles († 1709) and three centuries of art prices. J Econ Hist 73(3):766–791
Graddy K, Ashenfelter O (2011a) Sale rates and price movements in art auctions. Am Econ Rev Pap Proc 101:212–216
Graddy K, Ashenfelter O (2011b) Art auctions. In: Towse R (ed) A handbook of cultural economics, 2nd edn. Edward Elgar, Cheltenham, pp 19–28
Graves R (2012) The Greek myths, Deluxe edition. Penguin Publish, New York, pp 577–581
Greenwald D (2018a) Modernization and rural imagery at the Paris Salon: an interdisciplinary approach to the economic history of art. Eco Hist Rev (forthcoming)
Greenwald D (2018b) Colleague collectors: a statistical analysis of artists' collecting networks in nineteenth-century New York. Nineteenth-Century Art Worldwide (forthcoming)
Goldthwaite RA (1993) Wealth and the demand for art in Italy, 1300–1600. Johns Hopkins University Press, Baltimore
Grove Art Online. http://www.oxfordartonline.com/groveart
Grove Music Online. http://www.oxfordmusiconline.com
Guerzoni G (1995) Reflections on historical series of art prices: Reitlinger's data revisited. J Cult Econ 19(3):251–260
Hall P (2006) Cities in civilization. Phoenix, London
Hellmanzik C (2010) Location matters: estimating cluster premiums for prominent modern artists. Eur Econ Rev 54(2):199–218
Hodgson D (2011) Age-price profiles for Canadian painters at auction. J Cult Econ 35:287–308
Jockers ML (2013) Macroanalysis: digital methods and literary history. University of Illinois Press, Champagne-Urbana, IL
Li X, MacGarvie M, Moser P. Dead poets' property-how does copyright influence price? SSRN Working Paper, 15 June 2017
Loewenberg A (1978) Annals of opera. John Calder, London
Marshall A (1890) Principles of economics, vol 1. Macmillan, London
Michaud G (1905) The brain of the nation. Cent Magaz 69:41–46
Miskell P (2016) International films and international markets: the globalisation of Hollywood entertainment, c. 1921–1951. Med Hist 22(2):174–200
Mitchell S (2016) Essays on synergies from the geographic clustering of literary artists. Doctoral Thesis, Trinity College, Dublin
Montias JM (1982) Artists and artisans in delft: a socio-economic study of the seventeenth century. Princeton University Press, Princeton
Montias JM (1989) Vermeer and his milieu: a web of social history. Princeton University Press, Princeton
Moretti F (2005) Graphs, maps, trees: models for a literary history. Verso, New York
Moser P (2011) Do patents weaken the localization of innovations? Evidence from world's fairs. J Econ Hist 71(2):363–382
Moser P (2013) Patents and innovation: evidence from economic history. J Econ Perspect 27(1):23–44
Murray C (2003) Human accomplishment: the pursuit of excellence in the arts and sciences, 800 BC to 1950. Harper Collins, New York

O'Hagan J, Kelly E (2005) Identifying the most important artists in a historical context: methods used and initial results. J Quant Interdisc Hist 38(3):118–125
O'Hagan J, Borowiecki KJ (2010) Birth location, migration, and clustering of important composers. J Quant Interdis Hist 43(2):81–90
O'Hagan J, Hellmanzik C (2008) Clustering and migration of important visual artists: broad historical evidence. J Quant Interdis Hist 41(3):121–136
Oberholzer-Gee F, Strumpf K (2007) The effect of file sharing on record sales: an empirical analysis. J Polit Econ 115(1):1–42
Oosterlinck K (2017) Art as a wartime investment: conspicuous consumption and discretion. Econ J 127(607):2665–2701
Pokorny M, Sedgwick J (2010) Profitability trends in Hollywood, 1929 to 1999: somebody must know something. Econ Hist Rev 63(1):56–84
Porter ME (1996) Competitive advantage, agglomeration economies, and regional policy. Int Reg Sci Rev 19(1–2):85–90
Reitlinger G (1961) The economics of taste; the rise and fall of picture prices, 1760–1960. Barrie and Rockliff, London
Renneboog L, Spaenjers C (2013) Buying beauty: on prices and returns in the art market. Manag Sci 59(1):36–53
Rosenthal SS, Strange WC (2004). Evidence on the nature and sources of agglomeration economies. Handbook of regional and urban economics, vol 4. pp 2119–2171
Saint-Raymond L, de Maupeou F, Cavero J (2015) Les rues des tableaux. Géographie du marché de l'art parisien (1815–1955). Artl@ s Bull 4(1):6. https://paris-art-market.huma-num.fr
Scherer FM (2004) Quarter notes and bank notes: the economics of music composition in the eighteenth and nineteenth centuries. Princeton University Press, Princeton
Schich M, Song C, Ahn Y-Y, Mirsky A, Martino M, Barabási A-L, Helbing D (2014) A network framework of cultural history. Science 345(6196):558–562
Scott AJ (2004) On Hollywood: the place, the industry. Princton University Press, Princeton
Sedgwick J (2000) Popular filmgoing in 1930s Britain: a choice of pleasures. University of Exeter Press, Exeter
Spaenjers C, Goetzmann WN, Mamonova E (2015) The economics of aesthetics and record prices for art since 1701. Explor Econ Hist 57:79–94
St Clair W (2004) The reading nation in the romantic period. Cambridge University Press, Cambridge
Towse R (2001) Creativity, incentive and reward. Edward Elgar Publishing, Cheltenham
Towse R (2010) A textbook of cultural economics. Cambridge University Press, Cambridge
United Nations Conference on Trade and Development, and United Nations Development Programme (2010) Creative economy report 2010: creative economy: a feasible development option. United Nations, Geneva
Vaubel R (2005) The role of competition in the rise of baroque and renaissance music. J Cult Econ 29(4):277–297
Velthuis O (2005) Talking prices: symbolic meanings of prices on the market for contemporary art. Princeton University Press, Princeton
Velthuis O (2011) Art dealers. In: Towse R (ed) A handbook of cultural economics, 2nd edn. Edward Elgar, Cheltenham, pp 28–32
Vermeylen F (2003) Painting for the market. Brepols, Turnhout
White HC, White CA (1993) Canvases and careers: institutional change in the French painting world. University of Chicago Press, Chicago
Waldfogel J (2012) Copyright protection, technological change, and the quality of new products: evidence from recorded music since Napster. J Law Econ 55(4):715–740
Waldfogel J (2014) Chapter 12 – Digitization, copyright, and the flow of new music products. In: Ginsburgh V, Throsby CD (eds) Handbook of the economics of art and culture, vol 2. North-Holland, Oxford, UK. http://proquest.safaribooksonline.com/?fpi=9780444537768

Railroads

Jeremy Atack

Contents

Abstract

From the outset, railroads excited the public's imagination, leading to extravagant claims regarding their contribution to economic growth and development that only grew with time. Skepticism about these claims eventually led to their debunking, most notably by Robert Fogel whose seminal work was also central to the emergence of cliometrics and inspired the first major controversy in the discipline. The flood of work that followed paints a more nuanced picture, generally assigning a larger growth and development role to the railroad than Fogel, but one far short of the hyperbole of past generations. Railroads were important to the economy and transformative in their effect but not decisive. By lowering transport costs – especially after adjusting for quality – they not only

J. Atack (✉)
Vanderbilt University, Nashville, TN, USA

NBER, Cambridge, MA, USA
e-mail: jeremy.atack@vanderbilt.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_30

diverted trade but also created new opportunities in part by generating positive externalities that promoted scale and agglomeration economies with network spillovers continent-wide.

Keywords
Railroads · United States · Historical · Growth · Development · Cliometrics

Introduction

Over the course of the nineteenth century and into the twentieth, the view that railroads were the dominant force in the economy and in society's progress had become an increasingly common theme (see, for example, Emerson in a speech from 1844 (1903, 364), United States Census Office (1864, clxix), Adams et al. (1871, 335) or Jenks (1944, 3)). Consequently, when W. W. Rostow asserted that "the introduction of the railroad has been historically the most important single initiator of take-offs" into modern self-sustained economic growth (Rostow 1956, 45), few questioned his claim. It would, however, receive increasing pushback as he elaborated that theme (Rostow 1960, 1963). One scholar, in particular, Robert W. Fogel, then a graduate student at the Johns Hopkins University, challenged Rostow's bold claims in what would become his doctoral thesis (Fogel 1962, 1964). In it, Fogel introduced a new concept, social saving – the gain to society from an innovation – to measure the contribution of the railroad to America's growth. In the process, he helped lay the foundations for what we today call "cliometrics" – the subject of these handbooks. Indeed, Fogel's work on railroads was cited as the basis for his winning the Nobel Prize in Economics (jointly with Douglass C. North) in 1993 (https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1993/press.html). As a result, this review of the putative role of the railroad in American economic development and growth also casts light upon the evolution of the field of cliometrics.

There is also a parallel body of work, dating from slightly later, regarding railroads in countries elsewhere and on every continent that I will ignore here. That literature was, by and large, motivated by the same issues as the American. Moreover, as tools – both empirical and applied (such as trade theory and geographic information systems (GIS)) – have continued to develop, railroads have remained a focus on interest (e.g., Donaldson and Hornbeck (2016)) around the world including China's latest development initiative, "One Belt, One Road."

Early Railroads

The idea of using rails to provide a smooth track for wheeled vehicles carrying heavy loads, hauled by man or animals and eventually pulled by stationary engines, dates back centuries. Agricola's famous manual on mining, for example, illustrates one

such configuration and subsequent innovation led eventually to such familiar devices flanged wheels to locate the vehicle correctly on the track (Agricola et al. 1912). Rather, the real innovation that made the *railroad* practical as a means of long-distance transportation was the development of a mobile steam engine which was small yet powerful enough to provide the necessary motive power. Such a breakthrough occurred, more or less simultaneously, on both sides of the Atlantic at the start of the nineteenth century (Bathe and Bathe 1935; Dickinson and Titley 1934). Even so and although the idea had been in the air for over decade (See Sellers 1886, p. 13 quoting from an 1812 address by Oliver Evans), it would be more than a generation before steam locomotion itself would become a reality.

The pioneering railroad with steam locomotion was built between Stockton and Darlington in northeast England in 1825 to carry coal from a mine to a coastal port. Despite its intended purpose, the first trip attracted several hundred passengers – far more than the promoters had anticipated – many of whom rode atop the coal. Following this successful "proof of concept" demonstration, the invention was innovated with the Manchester to Liverpool Railway in England and by the Baltimore and Ohio Railroad in the United States, both of which began (some) service in 1830.

Trade and Improved Transportation

Other railroads quickly followed, driven in part by local commercial interests fearful that they might be eclipsed if cheaper or better means of travel and shipping diverted business elsewhere. Indeed, such trade diversion was of explicit concern for east coast US cities like Boston, Philadelphia, and Baltimore which had already seen some of "their" traffic siphoned away by New York once the Erie Canal opened in 1825 (Condit 1980). Faced with this threat, the city fathers in Boston and Baltimore promoted rail alternatives (the Boston and Worcester Railroad and the Baltimore and Ohio Railroad) linking their ports and trading houses to points west, ideally the Great Lakes and Ohio River valley. Philadelphia facing the same decision, however, made the wrong choice and opted instead to promote a canal system that, because of topography, was a much inferior alternative to either railroads or the Erie Canal. The Pennsylvania Mainline Canal would eventually fail, and its lack of competitiveness retarded the growth of Philadelphia (although it did facilitate the development of coal mining, especially anthracite, in the state).

Trade diversion occurs because trade, like electricity or water, flows in the direction of least resistance, where that resistance is caused by trade barriers such as transport costs and other frictions (e.g., handling charges, spoilage and other losses, speed, and timeliness). Wagon transportation was extremely slow, hard on whatever was being carried (passengers or freight) and with limited capacity. It was therefore very expensive, and there was little to no technological progress in the industry to reduce costs. Water transportation, while often slow (where powered by wind, paddle, or animal as by sailboat, canoe, or canal barge), had the virtue of cheapness and relative smoothness (except where transshipment was necessary at breaks in transportation). Where powered by steam, shipping by water could be

relatively speedy and became progressively cheaper as technological change drove down costs. Rail, however, combined directness of travel, speed, gentle handling, and capacity as well as benefiting from sustained advances in technology that reduced operating costs over time, including growing positive externalities (like network effects and agglomeration economies). Trade flows thus increasingly favored rail connections over the alternatives (Swisher 2017).

Improved transportation and communication not only diverted existing trade but also created new trading opportunities, leading to increased regional (and local) specialization in agriculture and manufacturing. For example, an instrumental variable analysis of the expansion of railroads in the Midwest before the Civil War suggests that they were responsible for more than half of the urbanization that occurred in the region in that period and almost two-thirds of the increase in improved acreage (that is, land in agricultural production) (Atack et al. 2010; Atack and Margo 2011). Some of these changes occurred quickly– for example, the decline of agriculture, especially grain production, in New England and the rise of manufacturing in that region (Field 1978). Moreover, these changes had far-reaching consequences such as moving agricultural production into areas better suited or less plagued by pests (Olmstead and Rhode 2008). It also allowed manufacturers to realize scale economies and quality improvements through specialization and skill development and to take advantage of agglomeration economies as well as promoting the development of mineral resources (especially coal and iron ore) in the interior of the country.

By 1840, the US railroad mileage was comparable to that of canals but still only a small fraction of the miles of navigable river, especially those served by canoes, rafts, and other shallow water vessels. However, unlike rivers which went where nature desired, roads, railroads, and canals went where their builders intended – and usually by a more direct route. Moreover, even large teams of horses were limited in the cargo that they could haul given indifferent road surfaces, poor grading, and wheel and wagon designs of the time. Indeed, most of the nation's roads were unpaved and little changed until the twentieth century when the "Good Roads" movement (started by bicyclists but continued by early automobile pioneers) lobbied for increased government spending on roads– including by the federal government. These efforts began to pay off, particularly after World War One, with improvements to the roadbed and surface, drainage, and building of bridges over streams and small rivers. Until then, there was little or no change in road travel times or the costs of road transportation – a.k.a. productivity change. This pattern is in stark contrast to the dramatic productivity gains in water and rail transportation manifest in sharply reduced fares and freight rates, sometimes by as much as 90% that benefitted both producers and consumers and increased the demand for transportation services (Taylor 1951).

Rail Construction and Its Geography

GIS data: Railroads were "big news," and so we know a great deal about railroad construction from contemporary local newspaper reports, annual reports by railroads to their shareholders, trade journals (like the *American Railroad Journal* and

Railway Age), contemporary travel guides (like *Appletons Travel Guide* or *Rand McNally Travel Guide*, some of which were published annually), and histories, especially of individual railroads (e.g., Stover 1975, 1987).

Railroads were also marked on maps (and with ever greater geographic precision over time (Modelski 1987)). Unfortunately, these multiple sources are often inconsistent with one another and may even represent different things. Maps, for example, show rail links between points, but there was a time lag between the map being engraved and its being published. Did the engraver mark only those links that were known (and based upon what source?) when the engraving was made or did they include those railroads that they expected to be operational by the time that the map was published? Cities and towns were points on most maps to or through which railroads "connected," but most railroad companies had their own stations and freight yards that were physically separate and not (easily) connected to those of other companies. Incompatibilities between railroads, especially track gauge (see below), also produced discontinuities in practice. Rivers were also represented by lines on maps, but not all were, in fact, crossed (bridged, anyway) by railroads (also shown as lines on maps). Instead, passengers and freight (and sometimes the wagons complete with their contents and locomotives) were instead ferried across the river. Sometimes two different railroad lines connected the same two points, or a single railroad built two tracks to allow concurrent two-way traffic. Where the rail line was single-track (as most in the United States were), sidings and turnouts had to be provided along the way so that trains traveling in opposite directions might pass one another. Moreover, with regard to construction itself, this took time and might be interrupted at any moment by events both natural (e.g., inclement weather) and man-made (e.g., financial panics or corporate bankruptcy). Some rail once built might be improved (e.g., straightened or regraded); other track might be abandoned. Each of these circumstances presents a challenge for statistical reporting, some of which is apparent in Fig. 3 (compare Carter et al. 2006; Wicker 1960).

Much of the most recent work on railroads relies upon spatial location derived from geographic information systems, especially the databases and shapefiles developed by Atack (2013, 2015), reflecting a growing belief that spatial relationships and location are as importance as miles. There were also important earlier efforts to model this spatial dimension, pre-GIS (Craig et al. 1998; Fogel 1964).

While Atack's GIS data represent spatially accurate locations for rail links (many of which no longer exist), they also strip out important features such as double tracking, sidings, and turnouts because these features do not generally appear on maps or in contemporary accounts and thus cannot be dated. However, even the dating of the main lines is imprecise because it is based on the copyright date for contemporary maps which provided "evidence" of the existence of a railroad connection between points (Atack 2013, 2015, 2018). These maps are georeferenced (using ArcGIS) – a procedure designed to accommodate and adjust for inaccuracies in drafting the maps and to compensate for differences in scale and projection so that they align with known geographic coordinates.

From 1861 onward, the maps selected represent best guesses of the railroad system at approximately 5-year intervals to 1911 based upon map copyright dates. When the interval between maps was reduced, there were many more inconsistencies

between them. When the interval was increased, it was felt that important information on timing was lost. During the 1850s, the data are at approximately 2-year intervals, except in the Midwest where they are annual because good data on construction and railroad operation from contemporary sources exists (Paxson 1914). Before then, I relied upon the database of railroad construction originally assembled by an agricultural economist Milton C. Hallberg (now deceased) from Pennsylvania State University that is available online at http://oldrailhistory.com (Hallberg 2004). This database gives the year when each stretch of line was built before 1850. In each case, the precise location of the railroads (given that these involved major earthworks and structures like bridges and tunnels) was then provided by satellite imagery and US Geological Survey maps (many of which date back to the late nineteenth century and are available online) by digitizing tens of thousands of points along them (Atack 2013, 2018).

Once a GIS database has been constructed, GIS software enables many calculations to be made quickly and easily – such as proximity to a rail link (using, e.g., the ArcGIS "buffer" command) and their geographic density, calculations that Fogel (1964) also made but which took him months without GIS and computers (personal conversation, 2008). However, even with modern technology, the construction of a network like that used by Donaldson and Hornbeck (2016) still requires considerable work (particularly in establishing the rules by which railroads connected to each other and to other modes of transportation – issues on which reasonable people and factual evidence may also still disagree).

Railroad construction: Early railroads were built in both the Northeast (e.g., the Boston and Worcester) and the South (e.g., the Baltimore and Ohio). Indeed, for a brief moment in the 1830s, the South boasted the longest railroad in the world, the Charleston and Hamburg, which enabled shippers to avoid the rapids (and transshipment) on the Savannah River at Augusta, diverting river traffic away from Savannah and toward Charleston.

By the end of 1840, there were around 3000 miles of track in service or under construction (GIS-derived estimates, i.e., rail links between different points) – more than in any other country in the world (Mitchell 2007) – a preeminence that the United States retains to this day. The majority of early railroad track was in the Northeast, but over 42% of the rail links were in the South with the balance in the Midwest, where early railroads, for example, connected Toledo to southern Michigan (the Erie and Kalamazoo). Thereafter, beginning in the late 1840s, much of the push would be focused on the Midwest (and eventually, the West). The first railroad linking Chicago to the East Coast (the Michigan Southern) opened in 1852, and by 1858, the Pittsburgh, Fort Wayne, and Chicago Railroad was providing the first all-rail route to the eastern seaboard from Chicago. This, together with Great Lakes shipping, helped make Chicago the center of the world's wheat trade.

Since the Census regions differ immensely in size, perhaps a better metric for access to rail transportation is square miles of land per mile of track (Table 1) – a measure easily calculated using GIS data. These calculations show that the density of railroads in 1840 in the Northeast was 6 times that in the South and more than 30 times that in the Midwest. In the decades that followed, the density of rail links increased

Table 1 Proximity of land to railroads by Census region, 1840–1900

	Square miles of land per mile of railroad			
Census region	1840	1860	1880	1900
Northeast	105	20	9	6
South	661	124	47	18
Midwest	3434	78	21	10
West		51,559	267	50

Computed using GIS data from (Atack 2015)

everywhere as more railroads were built, but the ratio remained relatively constant between the South and Northeast in 1860. However, between 1840 and 1860, rail density in the Midwest increased sharply relative to elsewhere so that by 1860 it was one-fourth that of the Northeast. If we think of a railroad track bisecting the land, then in 1900, the average section of land (=640 acres = 1 square mile) was within 3 miles of a railroad in the Northeast (i.e., 3 miles either side of the line) and just 5 miles from a railroad in the Midwest. Even in the South, average distance to a railroad had shrunk to just 9 miles. Distances remained large only in the West, and even in that region this was only because of the vast areas of desert and marginal land that lay between about the 100th meridian and the Sierra Nevadas.

Most railroads were built by the private sector in response to market signals whether real or imagined. When they were "best" built was a matter of conjecture. Certainly, if traffic justified their existence from the get-go (like the Erie Canal, where sections of the canal were opened as built and generated immediate profits, years before the entire project was finished), then construction was "too late" from the economy's standpoint. But, were they built ahead of demand? According to an 1860 editorial in the *Picayune* (a New Orleans newspaper), "nine-tenths of our roads when first traversed by steam pass through long ranges of woodlands in which the ax has never resounded, cross prairies whose flowery sod has never been turned by the plow, and penetrate valleys as wild as when the first pioneers followed upon the trail of the savage," suggesting that railroads preceded settlement.

The evidence, however, suggests otherwise. Fishlow (1965), for example, found that most early midwestern railroads were profitable almost immediately, which is unlikely if they were built before trade justified their existence. Moreover, most of the railroads built in the Midwest during the 1850s – a major (and arguably, speculative) construction boom – served established farm regions rather than those parts of the state most recently settled. For example, 60% of the railroad mileage in Illinois by 1853 was built in the leading wheat and corn counties that made up just 25% of the state's area. Similarly, in Wisconsin, seven wheat-producing counties (plus Milwaukee), just 10% of the state's area, had 60% of the track in 1856. Those counties that got railroads in the 1850s were already growing and developing faster before the coming of the railroad. Even so, more recent GIS analysis using IV has also found that the coming of the railroad to these midwestern counties explains more than half of the urbanization – that is, population living in towns larger than 2500 person (Atack et al. 2010) – and perhaps two-thirds of the growth in land under

cultivation (Atack and Margo 2011). Indeed, there is much to suggest that railroads (and other man-made transportation systems) were endogenous to growth (Swisher 2017) and thus that a general equilibrium outcome will differ dramatically from the partial statics solution.

It is argued that the slow expansion of railroads in the South reflected the abundance of rivers in the South connecting to the coast which met area's needs for shipping low value-to-weight and nonperishable freight (e.g., baled cotton) out of the region. These rivers, however, did little to promote intra-regional trade, and what railroads were built had difficulty exchanging traffic and shipping outside the region because of differences in track gauge and the lack of bridges across major rivers – especially the Ohio. These issues would have consequences for the Confederacy in fighting the Civil War.

The mechanism was usually that a group of railroad promoters petitioned the state for a corporate charter to build a railroad between two points – sometimes the request was more specific regarding routing, probably to secure more political support in the legislature pre-adoption of state general incorporation laws. Once the privilege was secured, the promoters sought construction funds locally, nationally, and internationally primarily through the sale of fixed interest debt secured by mortgages on the property, although most New England railroads were equity-financed locally (Chandler 1954). More distant investors appreciated the "guaranteed return" and the liquidity which those investments assured by the hoopla surrounding railroads both general and specific. Local investors were often local landowners, merchants, banks, and city boosters who stood to gain either through increased business or rising land values as the benefits of that trade were capitalized into the fixed factor – an outcome critical to assessing the railroads' contribution (e.g., Donaldson and Hornbeck 2016; Fogel 1964; Swisher 2017). Actual construction would employ hundreds of manual laborers as well as skilled engineers and surveyors, and once opened, the railroad would require a locally resident staff for stations, signaling and traffic control, maintenance, and repair as well as the train crews. Railroads were big business and that business spread everywhere.

By 1860, when there were more than 25,000 miles of rail links (i.e., point-to-point connections by rail) in the United States (and 30,000 miles of track – including double tracking, sidings, etc.), the Midwest's share had grown to 38% of the nation's rail links, while the Northeast's share had shrunk to a third and the South's to 28%. This geographic distribution of railroad links in the United States is shown in Fig. 1 (this figure also shows track gauge, on which more is said below). Two features are notable with respect to the geographic distribution. First, there were large areas of the country devoid of railroads (particularly in the mountainous areas of the Appalachians). Second, there were no rail connections between the Midwest and the South, which surely handicapped interregional trade.

The outbreak of the Civil War brought an abrupt slowdown in new rail construction, although some new track continued to be built in both the Confederacy and the Union. For example, a stretch of railroad was built in Alabama and Mississippi between Selma, Alabama, and Meridian, Mississippi (but split by the Tombigbee River, which was not bridged during the war), while the Union extended the

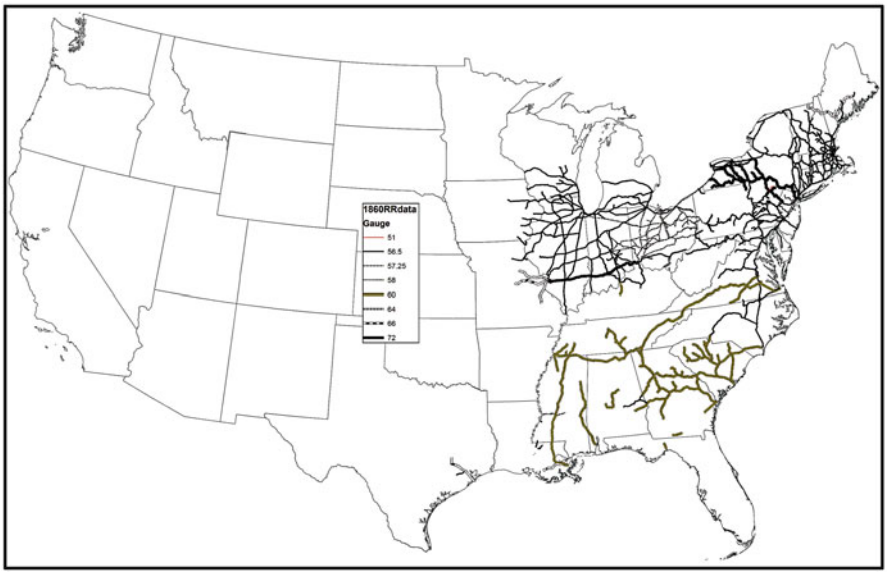


Fig. 1 American railroads in operation in 1860. (Gauge data represent the distance between rails in inches from (Taylor and Neu 1956). Standard gauge is 1435 mm (=4 feet 8 1/2 inches).) (Source: Computed using GIS data from Atack (2015))

Nashville and Northwestern Railroad westward to the Tennessee River after their occupation of Nashville. By far the most significant Civil War railroad project, however, was the authorization for the first transcontinental railroad in the Pacific Railroad Act of July 1862. Construction on this project was begun by the Central Pacific Railroad in early 1863 building eastward through the Sierra Nevada mountains. However, the Union Pacific Railroad building westward from Omaha did not break ground until July 1865, several months after the war's end.

The military on both sides were quick to recognize the importance of railroads in deploying troops to battlefields from Bull Run to Appomattox. As a result, railroads became military targets (as in the burning of Atlanta and Sherman's March), while at the same time increased intensity of use accelerated water and tear on the track. The orientation of southern railroads to southern ports rather than intra-South trade and relative to those in the Northeast and Midwest, however, made them far less useful in the conflict than the Union's railroads, although there is (as yet) no cliometric work on this (Turner 1953).

One consequence of the greater intensity of the use of railroads during the Civil War was that the Pennsylvania Railroad began to experiment with steel instead of iron rails in late 1862 or early 1863, ordering 150 tons of steel rails from British mills. This would have a substantial long-term payoff after the War as steel rails proved capable of withstanding heavier trains running at higher speed for orders of magnitude more years than the iron rails they replaced – all issues related to productivity.

Initially, the higher cost of steel discouraged replacement of iron rails with steel. In 1867, for example, iron rails cost $83.13/ton compared with $166/ton for steel from the first American Bessemer steel producer, the Cambria Iron Works. This cost difference only made steel worthwhile where wear was particularly high, as on curves and at stations. Moreover, worn-out iron rails were readily sold for scrap from which to produce new rails, but no such market for scrap steel then existed. However, as a domestic Bessemer steel industry ramped up, production costs fell (Temin 1964) so that, by 1880, with 24 domestic steel mills in production, the cost of steel had been cut to $67.50/ton while iron cost $48.25/ton. More importantly, after 1883, steel actually costs less than iron. As a result, all new track was probably laid in steel, and by the end of the 1880s, more than 80% of the nation's track was steel. This compares with only 30% at the start of the decade. Moreover, the availability of progressively cheaper steel had spillover effects for other industries (Atack and Brueckner 1982).

The replacement of wartime destruction, deferred maintenance, and pent-up demand for new lines led to a railroad construction boom after the war. This shows up clearly in all statistics regarding rail construction (Fig. 3). Unfortunately, there is no single definitive series for this. A Census of Transportation taken as a part of the federal decennial census in 1880 provides one retrospective series. Other data have been assembled from the trade magazine *Railroad Age*, and there are data on the mileage owned and the mileage operated by railroads. Similar disagreements exist about earlier data (Wicker 1960), and the addition of GIS estimates of railroad links only further muddies the water.

Between 1868 and 1873, track mileage doubled. Indeed, almost 7500 miles of track was built in 1872 – almost as much as had existed in the entire country at the end of 1849. Much of this construction took place in areas already well served – like the Midwest and the Northeast – but it also included the completion of the first transcontinental railroad in May of 1869. After 1873, construction activity varied from year to year but stayed above 1400 miles per year from 1866 through 1914, with several years when more than 5000 miles were built (Carter et al. 2006; Chandler 1965). This was especially so in the 1880s when the boom in construction, concentrated primarily in the Plains states, eclipsed even the post-Civil War construction boom. Moreover, other transcontinental railroads would follow, providing additional trunk lines in the southern and northern plains to which branch lines could connect. These facilitated the growth of mining in Colorado and Utah and of farming communities in the Great Plains states.

As a result of these late booms in construction, the density of rail track increased sharply everywhere (Fig. 2). According to *Historical Statistics*, railroad track operated (net of yards and sidings) plateaued in the 1910s and 1920s at around 300,000 miles before beginning to decline in the 1930s as the Great Depression took hold and railroads disposed of unproductive assets. Thereafter, abandonments of track have generally exceeded new construction so that by 1980 there were about 200,000 miles of track in operation, net of sidings and yards.

Track Gauge: The map of railroad links in the contiguous United States in 1900 (Fig. 2) does not separately identify track gauge because, by then, virtually all the nation's railroads had adopted a single gauge –"standard gauge" of 4 feet 8.5 inches –

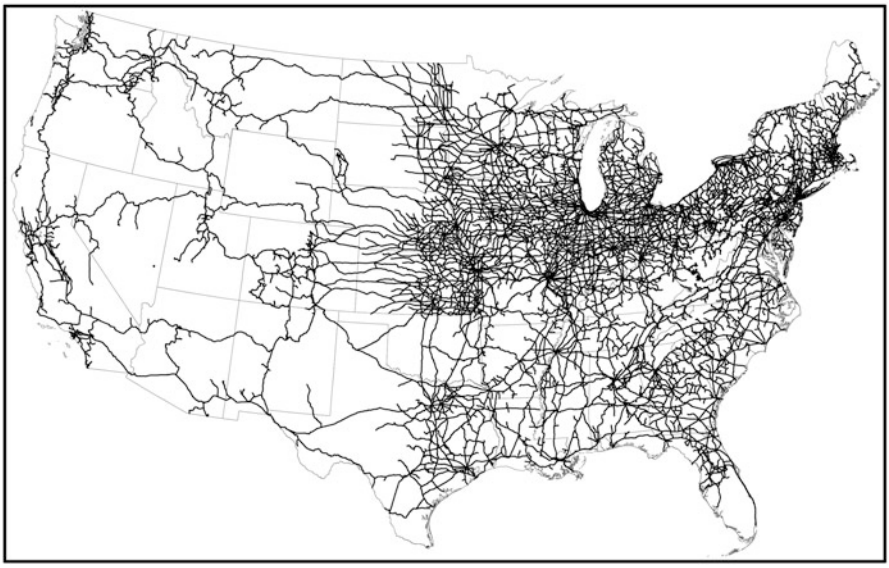


Fig. 2 Railroad links in 1900. (Source: Computed using GIS data from Atack (2015))

and locomotives and rolling stock could (in theory, anyway) be used anywhere. This harmonization of track gauge began in earnest following the Civil War and was finally realized in 1886 when the last of the southern railroads converted (Gross 2016; Puffert 2000).

The earlier multiplicity of track gauges reflected institutional, technical, and economic forces, and, despite harmonization, there would remain isolated pockets, particularly in the higher mountains, where narrow gauge railways still held sway because tight curves could not be negotiated by larger locomotives or wagons (i.e., technical considerations dominated). Absent terrain constraints, technical considerations argued for a broader rather than a narrower gauge to take advantage of the scale economies offered by the larger enclosed volume of wagons and carriages. However, broad gauge wagons not only have difficulty negotiating sharp curves (despite the invention of flexible trucks), but their greater laden weight challenged locomotives with only limited power on steeper grades. Standard gauge was a compromise well suited to most terrain although broader gauges, like the 5′ used in much of the South and the 6′ gauge adopted by the Erie in New York, were preferred where the terrain was relatively flat and the track was straight. Rail wagons operating on 5′ track were also allegedly better suited to transporting the standard cotton bale than those built for standard gauge. Of course, other considerations sometimes intruded like the decision by many of Maine's railroads to use a 5′ 6″ gauge because that was the gauge of the Grand Trunk Railroad in Canada, and Maine hoped to siphon trade from that system to Portland and other ports along their coast. These incompatibilities between gauges "locked in" users (for other examples, see Arthur 1989, but also see Liebowitz and Margolis 1994).

The initial adoption of standard gauge in America traces to the Baltimore and Ohio railroad, America's first steam railroad. As the pioneer railroad, there was no domestic source of locomotives and rolling stock from which to order equipment, so these came from England and were built to the prevailing English gauge. However, since the B&O was starting with a tabula rasa, this was of no consequence, and the B&O had no reason to specify differently – especially because their route to the Ohio River was anything but straight and level.

These explanations, however, do not satisfactorily explain the adoption of a 4′ 10″ gauge by most of the railroads in Ohio, which, not surprisingly, became known as the "Ohio gauge" and was stipulated in many of the Ohio railroad charters. Perhaps the best rationalization of this anomaly is that the first locomotive delivered to the Mad River and Lake Erie Railroad (the first railroad chartered in Ohio and the second to start operations) had been built by Rogers, Ketchum, and Grosvenor in Paterson, New Jersey, for a New Jersey railroad to a 4′ 10″ gauge. But that order was cancelled before delivery, and the locomotive was resold to the Mad River Railroad. The difference between standard gauge and Ohio gauge, however, was of no practical consequence as the earliest Ohio railroads did not connect to any others. Moreover, when they eventually did interconnect, locomotives and rolling stock could use either gauge because of the width of the wheel tread and the bullnose of the rails which kept trains centered on the track. All of this makes track gauge, especially the decades-long persistence of the "Ohio gauge," an excellent historical example of path dependence (Arthur 1994; Puffert 2000, 2009).

In 1860, eight different track gauges (besides narrow gauges of 4′ or less) were in use by American railroads (Fig. 1), although three of these 4′ 8.5″, 4′ 9.25″, and 4′ 10″ were compatible with one another. Indeed, the 4′ 9.25″ gauge was used by just one line, a spur off the Mad River Railroad connecting Corey and Findlay, Ohio. It appears to have been a compromise between standard and Ohio gauge.

Nationwide, standard gauge dominated (Table 2), accounting for at least 60% of the track mileage (net of sidings, double tracking, etc.), and those compatible with standard gauge accounted for 72% of the nation's rail links. However, the dominant gauge in the South was a 5′ gauge, accounting for 18% of the rail links in 1860, and was the last to change (Gross 2016), although many of the railroads built in Virginia and North Carolina adopted standard gauge from the start or had switched to

Table 2 Railroad links by track gauge, 1860 (shading denotes compatible gauges)

Gauge (inches)	Mileage	Share of total links
51	34	0%
56.5	15335	60%
57.25	16	0%
58	3133	12%
60	4486	18%
64	183	1%
66	852	3%
72	1486	6%

Source: GIS calculation from (Atack 2015) following classification by (Taylor and Neu 1956)

standard gauge by 1860. This profusion of track gauges prevented the American rail system from being a network and dissipated the positive network externalities (Puffert 1991, 2000, 2009; Taylor and Neu 1956). A question, however, remains as to how large these externalities could have been, especially earlier on (Arthur 1994; Liebowitz and Margolis 1994, 1995).

These artificial breaks from breaks in gauge were compounded by delays in bridging many rivers (the Mississippi was bridged relatively early at Rock Island in 1856, but no railroad crossed the Ohio until 1871 – partly because of opposition from steamboating interests and constitutional concerns relating to navigability) and the reluctance of railroads connecting to the same cities to build "Union" stations and have common freight yards. For example, Colton's 1856 map of the Boston area held by the Library of Congress shows six distinct railroad depots in Boston proper terminating different railroads from the south, west, and north. Moving freight or passengers between these sometimes involved traversing the entire city and its narrow, congested streets. The first "Union" station was opened in Indianapolis in 1853 and was jointly owned and operated by railroads serving the city. Most, however, date from much later; Chicago's first effort, for example, wasn't until 1881, while Washington DC's Union Station dates from early in the twentieth century.

Such breaks in transportation were the bane of travelers but through traffic a boon to local businesses and workers (laborers, teamsters, etc.) providing transportation between the breaks (but congesting the streets) and lodging and food for weary travelers, many of whom likely missed connections (Bleakley and Lin 2012). Indeed, such local interests underlay the "Erie gauge war" in the early 1850s when the city of Erie, Pennsylvania, fought to maintain a break of gauge in that city between the broad gauge of the Erie Railroad out of New York and standard and Ohio gauge railroads (Kent 1948), providing a concrete example of Frederic Bastiat's economic sophism of the "boon" to be realized from a "negative railroad" (Bastiat and Stirling (trans) 1922, especially Chap. 17).

Eventually the positive network externalities prevailed everywhere, and gauges were harmonized (Puffert 1991, 2009). But this came at a cost. For example, the adoption of standard gauge in the South led to the diversion of southern traffic from a mix of rail and ocean shipping to all-rail shipping and led to the decline of shipping through ports such as Savannah and Charleston (Gross 2016). The changeover, however, was a triumph of planning. Like the earlier conversion of the southern portions of the Illinois Central, the change in gauge was accomplished in a single 24-hour period, by stationing crews along the track ready to move the rails and fix axles and wheels as soon as traffic was shut down – a result of and triumph for meticulous planning and coordination.

Railroad Finance and Construction

One part of Fogel's attack on Rostow's argument was to diminish the importance of railroads as a critical factor in American economic growth, especially during what Rostow termed "the take-off into self-sustained growth" (Rostow 1956, 1960).

During that phase, dated by Rostow to between 1843 and 1860, railroad construction supposedly played a decisive role in the United States (as it did in many other countries) through its demand for iron, wood and coal, transportation equipment, and other inputs. The timing clearly coincides with a pronounced uptick in railroad construction (see Fig. 3).

However, Fogel determined that railroad demand probably played little role in developing these other industries – at least during Rostow's time frame. For example, much of the iron used by railroads (whether for rails or in fabricating parts for locomotives and rolling stock) was imported from England since domestic mills had difficulty producing the requisite quality. Moreover, most locomotives at that time burned wood, not coal, and most railroad ties were split rather than milled. There was thus an impact on lumbering (an agricultural pursuit) but not sawmilling (an industrial activity) (Fogel 1964). True, domestic iron was used for railroad spikes but more was used in shoeing horses, while demand for railroad rolling stock almost certainly reduced the demand for wagon, canal barges, and steamboats – that also used iron.

This is not to deny that railroads had an impact. Railroad demand was absolutely critical to the emergence and the growth of a domestic Bessemer steel industry. It also accounted for a substantial fraction of coal consumption once locomotives

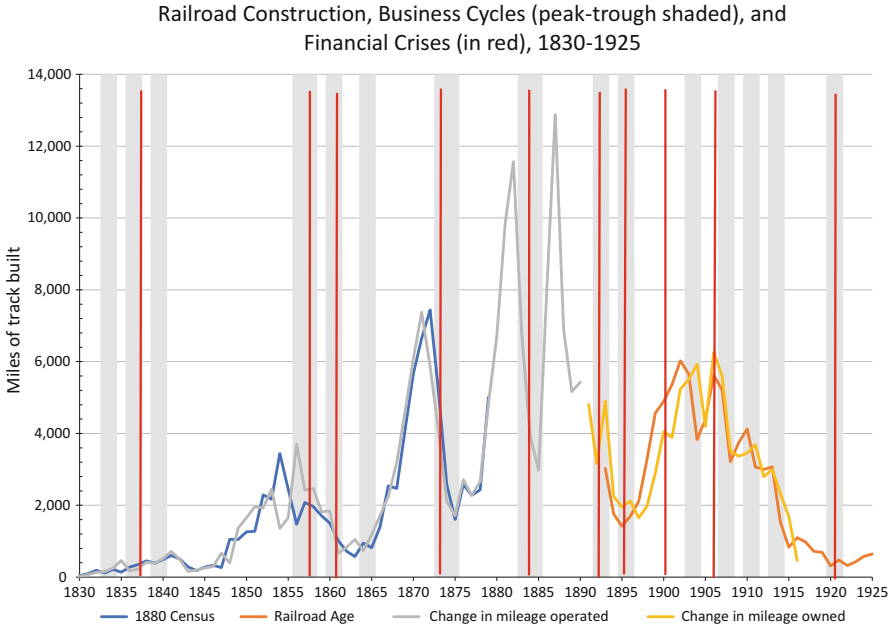


Fig. 3 Railroad construction and economic cycles, 1830–1925. (Source: Railroad construction from (Carter et al. 2006, series Df882–885); business cycle dating from (Davis 2006); financial "crises" are an amalgam of events affecting US banks and financial markets)

switched from burning wood. Initially, most engines were fueled with wood which was available in abundance in the eastern United States. But, as more land entered cultivation and the railroads pushed westward into treeless prairies and deserts, wood became less readily available and more expensive, so coal increasingly became the fuel of choice in the 1850s. Some railroads, such as the B&O, used coal from the start, despite the difficulties of burning anthracite. The growing availability of bituminous (softer and more easily burnt) coal and declining coal prices made the switch easier. By 1880, more than 90% of railroad fuel was coal (White 1968, pp. 83–90), accounting for a significant fraction of coal production. In 1915, for example, US railroads consumed 28% (122 million tons) of bituminous coal production and 7% of anthracite production (Fernald 1918, p. 179).

The pattern of rail construction in Fig. 3 shows a distinct cyclical pattern and correlates closely with the dating of the business cycle (the gray-shaded columns denoting peak to trough in the cycle by calendar year) and also with financial panics and crises (the red vertical lines). Although the dating of business cycles is imprecise (in part because these come from a mix of lagging, coincident, and leading indicators) and none of these data enjoy a broad consensus, they are suggestive. According to Harley (1982), railroad construction cycles played a major role in the Kuznets business cycle. This is not surprising. Each mile of single-track railroad, for example, represented at least 350 tons of iron or steel and employed hundreds during construction, moving earth and ballast to grade the rights-of-way, lay the track, and build refueling stops for water and coal (or wood). Just the Bessemer steel and locomotives account for more than 6% of the Davis index of industrial production (2006). Consequently, railroads figure prominently in the writings of business cycle scholars from Thorp et al. (1926), Schumpeter (1939), and Fels (1959) onward, playing a major role in most cycles of economic activity both long and short, regardless of whether or not their role was decisive (Fogel 1964).

Harley (1982) also argues that the cyclical pattern of post-War railroad construction reflected the periodic breakdown of informal oligopoly agreements among the railroads regarding exclusive construction rights in particular areas. These agreements came about because construction at a particular time in a specific location generally precluded construction later (although there were cases of what today would be called "greenmail" – beginning construction on a new line in the hope and expectation that one will be paid to drop the project – e.g., the New York, West Shore, and Buffalo Railroad, although the ploy did not always work). Later construction, when the economy had grown more, would clearly be more profitable and thus more desirable from the railroad's standpoint than construction today. This provided a strong incentive to delay construction, but only if a railroad could be sure that those higher profits would be reserved exclusively to it.

Thus, railroads had an incentive to collude, but as gains increased with settlement, so too did the incentives to break ranks. Harley's case study of Kansas concluded that once settlement reached about 35% of its potential, these incentives could become irresistible although the optimal cooperative strategy was to delay construction until settlement was 75% complete (Harley 1982). By comparing actual

mileages of new rail construction in the Great Plains with those predicted by these strategies, he concluded that railroads were likely locked in competition during the depression of the 1870s (1873–1878) and after 1883 but cooperated at other times.

Railroads represented major financial investments, costing $25–30,000 per mile early on. This expenditure was comparable to that for canals but substantially more per mile than turnpikes. It was also about 10 times the average investment in a farm and 20 times or more the capital invested in the typical manufacturing firm at the time. For railroads built in mountainous areas (like Appalachia and especially the Rockies and Sierras) that required more cut and fill or in swampy areas (like much of the coastal South) where the ground was too unstable to support ballast, ties and rails cost even more to build per mile. By 1860, perhaps $1 billion had been sunk into railroads – an aggregate investment equivalent to that in all US manufacturing at the time – and would more than quadruple before the end of the century.

The cost of building each railroad was sufficiently large that most had to depend upon pooled investment, secured primarily from capital markets rather than banks, although some early railroad charters also had banking privileges (e.g., the West Feliciana Railroad and Banking Company). The raising of external capital was facilitated by grants of incorporation and limited liability to railroad promoters by the individual states. These assets were publicly traded and represented a major component on many markets. In 1885, for example, railroads accounted for more than 80% of the stocks traded on the New York Stock Exchange and almost two-thirds of those traded on the Philadelphia market (O'Sullivan 2007). Much of the financing for construction (and sometimes more than was actually spent on construction) was raised through the sale of debt, often secured by mortgages on land, rather than equity (except in New England where railroads were shorter and capital needs therefore less (Chandler 1954)). This debt traded on both domestic and especially foreign markets (where it was often denominated in pounds sterling) and quickly eclipsed in volume all other public investments – including US government debt – during most of the nineteenth century. The money raised not only financed construction but also provided a pool of money for fraud, bribery, and corruption (see, e.g., United States. Congress. House 1873).

Much of the construction finance raised by railroads went into grading the right-of-way, drainage, and bridging – investment that was completely immobile and had few other uses (except for roads – which many would become in the twentieth century as railroads abandoned their rights-of-way). It was literally a "sunk cost." Much of the rest went into track and ties, ballast, rolling stock, and the like – a fixed, if not sunk, cost. As a result, railroads had low marginal (and variable) costs relative to total costs. This cost structure is made for volatile shipping rates with changes in demand since roads could minimize losses by just covering variable costs. While low rates during economic downturns benefitted shippers, the shortfall in earnings comprised the ability of railroads to meet interest payments on their debt, making them highly vulnerable to bankruptcy in adverse economic circumstances (e.g., 1857, 1873, and 1894) and (potentially) hostile takeovers and mergers (Adams et al. 1871). Investors soon found this to their cost when the New York and Erie Railroad filed for bankruptcy in 1859 – the first trunk railroad in America to fail – although many

smaller railroads had preceded it and many more, both large and small, would follow in the subsequent decades, for example, in 1873 and more especially in 1893. Indeed, in the 1870s, perhaps a fifth of the railroad mileage in the country belonged to railroads in default. Defaults peaked again in the 1890s when more than 40,000 miles of track was in the hands of receivers and the Pennsylvania Railroad eventually absorbed over 800 other railroads, many of them during the 1890s.

More difficult years would follow in the twentieth century as the nascent trucking industry began to compete with railroad in the shipping of freight, especially beginning in the 1930s (Schiffman 2003). Construction of the interstate highway system accelerated the decline of rail traffic (or at least its share of traffic) as did the switch in home heating and electricity generation away from coal. In the late 1960s and early 1970s, six major rail systems in the Northeast filed for bankruptcy, including the Penn Central Railroad, and freight and passenger rail traffic threatened to collapse, leading to Congressional intervention. This led to the creation of Conrail in 1976, initially a government-owned enterprise, as the principal rail freight shipper in the Northeast and Midwest. Conrail would eventually be sold back to the private sector in 1987. Meantime, Amtrak was created in 1971 with government subsidies to continue to provide rail passenger service although it owns none of the track over which its trains operate, accounting in part for its poor service record.

Government Intervention and Inducements

From the start, governments interfered with railroads – the Baltimore and Ohio Railroad, for example, was the brainchild of Baltimore's city fathers and facilitated by the speedy grants of charters by the states of Maryland and Virginia. Such intervention, however, was neither new nor uncommon (Hughes 1991). The case for railroads was that as durable and sizable investments, they needed the privilege of incorporation to provide for continuity over their long (technical) life and to help raise funds for construction. In return, states often demanded that the railroad be built to a certain standard and track gauge and connect various locations within a specific time frame, sometimes exacting "help" with other matters as well. As part of this mutuality, railroads also got help against extortion by land owners along the right-of-way through recourse to eminent domain and the right to bridge rivers and divert streams. Sometimes, other inducements were given, as in cases where the private returns were viewed as marginal or submarginal. Thus, for example, the federal government in 1850 offered grants of federal land to the Illinois Central Railroad to encourage construction of that railroad from Chicago southward through sparsely settled parts of the state. This policy built upon prior experience encouraging road (e.g., in Ohio) and canal construction (e.g., the Illinois and Michigan Canal). When completed in 1856, the Illinois Central was the longest railroad in the world. It would subsequently extend its reach into Iowa and into the South following the Civil War (although it did not initially have a bridge connection across the Ohio River and the gauge on the other side of the river was the southern 5′ gauge until 1881– that is, several years ahead of the switch by the rest of the South's railroads (Gross 2016)).

As an incentive, grants of land contingent upon construction were seen as a universal success and would be repeated beginning with the Pacific Railway Act of 1862, authorizing the first transcontinental railroad. In aggregate, these land grants to the railroads were huge. Between 1850 and 1871, the federal government gave away 131 million acres – over 203,000 square miles – an area larger than California and New York combined. These land grants potentially provided the railroad with a right-of-way and an asset whose value the railroad's construction would dramatically enhance. Moreover, that land could serve as a security for loans, especially mortgages (whose market was already well developed). However, there are questions as to whether these inducements were really necessary.

A review of the contemporaneous ex ante traffic projections for the first transcontinental railroad (Duran 2013) suggests that such subsidies were not required (although they may have reordered priorities in a capital-constrained world). These estimates are also consistent with retrospective projections of realized returns with and without the land grants (Fleisig 1975; Fogel 1960). Indeed, ex post analysis for the Central and Union Pacific railroads and the Great Northern has concluded that private rates of return exceeded the opportunity cost of capital even in the absence of land grants. For the Santa Fe and the Northern Pacific, the value of land grants may have tipped the balance ex post in favor of construction, while construction of Texas and Pacific Railroad was probably a mistake ex post (Mercer 1974). Even where returns were marginal – as in the early years of the Union Pacific – later returns almost certainly compensated (Fogel 1960). Regardless of the adequacy of private returns, though, there is universal agreement that the social returns were large (i.e., there were externalities that the railroads could not capture and whose value likely far exceeded the opportunity cost of the land ceded to the railroads (Fogel 1960; Mercer 1969)).

Those land grants would, however, also create problems of premature (or abortive) settlement based on putative misrepresentations (certainly, overly optimistic forecasts) (Libecap and Hansen 2002; Raban 1996) to would-be settlers, both those already in the United States and throughout Europe. Moreover, by retaining ownership of much of this land, the railroads could cross-subsidize activities (e.g., in the shipment of coal mined on railroad land) and block competition (e.g., from pipelines) that led to further public intervention.

Governments also interfered with the operation of railroads, in particular, the terms and conditions of service, albeit initially in a roundabout way. As competition between railroads intensified as more lines were interconnected and shippers had choices regarding carriers and routings, the railroads colluded to restrict its effect (Riegel 1931). For example, in 1869, the three dominant railroads shipping between Iowa and Chicago (the Northwestern, the Burlington, and the Rock Island) formed the Iowa Pool by gentlemen's agreement to share 55% of freight revenues and 45% of passenger revenues equally in the hopes of restricting price competition between them. Other regional groupings of carriers would later enter similar agreements, such as the Southern Railway and Steamship Association organized by Albert Fink beginning in 1873.

Such arrangements were not challenged immediately or directly, but governments found the authority to regulate freight rates as a result of an expansive Supreme Court ruling in the case of Munn v Illinois (1876) (Higgs 1987; Hughes 1991). That case originated in a challenge to an 1871 Illinois law, pushed by a recently created fraternal organization known as the National Grange of the Patrons of Husbandry, requiring the licensing of grain elevators and setting maximum rates for the storage and handling of grain, most of which arrived by rail. By then, Chicago was the center of the world's grain trade, supported by trading organizations like the Chicago Board of Trade and with a large infrastructure of grain elevators, many with common ownership, that separated railroads from ships that carried American grain worldwide. Illinois brought suit against the elevator of Ira Munn and George Scott for operating their grain elevator without a license and for charging more than the maximum permitted rate under the Illinois Warehouse Act. After conviction in state court, Munn appealed to the US Supreme Court on the grounds of an illegal taking without due process under the 14th Amendment. In its decision supporting the state of Illinois, the US Supreme Court ruled that common law permitted the regulation of business "affected with the public interest" (which would include railroads as "common carriers"), although Justice Field filed a vigorous dissent against this expansive ruling. The result of the decision, however, was a flood of regulation, much of it covering railroads, from other states, such that, by 1886, 25 of them had passed laws similar to Illinois's.

Another of Illinois's "Granger" laws, this one outlawing freight rates that charged more for a short haul than a long haul "under substantially similar circumstances" was challenged in 1886 by the Wabash, St. Louis, and Pacific Railroad. In their decision on this case, however, the Supreme Court ruled in favor of the railroad since the shipment in question involved the regulation of interstate commerce by the state. Within a year of that ruling, though, Congress passed essentially the same legislation regulating freight rates at the federal level and establishing the Interstate Commerce Commission (ICC) to oversee implementation. This ushered in an era of increasing federal interference in railroad operation such as the Elkins Act of 1903 outlawing rebates and the Hepburn Act of 1906 which gave the ICC the authority to set just and reasonable rates, prescribe accounting practices, and prohibit railroads from carrying any commodity which they or a subsidiary produced. The effect of these laws was profound and long-lasting (MacAvoy 1965; Martin 1971; Ulen 1980). Indeed, the federal government itself would even briefly assume total control over the nation's railroads between 1917 and 1920 as part of the World War One emergency. The rise of trucking, and especially tractor-trailers in combination with interstate highways, private cars, and the airlines, siphoned off much of the railroad's traffic after World War Two. This led to a reorganization and reorientation of the railroading industry by the creation of Amtrak in 1971 and Conrail in 1976, while the passage of the Railroad Revitalization and Regulatory Reform Act of 1976 and the Staggers Rail Act of 1980 dramatically reduced the authority of the ICC over the nation's railroads until the agency was abolished at the end of 1995.

Innovation and Productivity Change in American Railroads

According to estimates by Fishlow, between 1839 and 1909, railroad services grew at an average annual rate of 11.6% – about three times faster than income and commodity output at the time. No other single major sector of the economy grew as rapidly (Fishlow 1966). Part of this growth, of course, reflected the dramatic expansion in the miles of track and the spread of railroads from a few isolated areas on the East Coast to covering the contiguous lower 48 states discussed above and shown in Figs. 1 and 2. However, even after taking into account the increased inputs of land, labor, capital, and fuel used by the railroads, it remains clear that output per unit of input – total factor productivity (TFP) – also increased markedly. Indeed, Fishlow's preferred estimates of TFP show annual average growth rate of 3.5% over the entire period from 1839 to 1909 – a very respectable rate – albeit a dramatic slowdown from the productivity growth experienced in the 1840s of almost 10% per year (Table 3). Rates of productivity growth generally slowed, decade by decade, but most notably in the first decade of the twentieth century. Although Fishlow reports no later data in any detail, he does include a chart showing productivity growth through 1953 which suggests that this slowdown in the rate of productivity growth continued through at least the first half of the century (Fishlow 1966, Chart 1, p. 627). This pattern is consistent with estimates reported by others (Caves et al. 1981).

The productivity growth in supplying railroad services was partially reflected in declines in passenger fares and freight rates. In 1839, passengers paid about 5 cents per mile (a substantial price at a time when the daily wage was a dollar or less for a 10–12 hour day), and freight was paid for an average of 7.05 cents per ton-mile, a price that restricted carriage to relatively high value-to-weight items like manufactures (Fishlow 1965, 1966), meaning that there was plenty of space for other carriers, especially water. Eighty years later, passenger fares had been cut by more than 60% and freight rates by more than 90%. Freight rates for water carriage declined similarly, but the mix of freight was shifting in favor of higher value-to-weight items which favored rail shipment. Road carriage, however, experienced no real change in

Table 3 Average annual rate of growth in total factor productivity in American railroads by period, 1839–1909

	Ending year						
Starting year	1849	1859	1869	1879	1889	1899	1909
1839	9.8%	5.7%	5.5%	4.6%	4.2%	3.9%	3.5%
1849		1.8%	3.4%	3.0%	2.8%	2.8%	2.5%
1859			5.1%	3.6%	3.1%	3.0%	2.7%
1869				2.1%	2.2%	2.3%	2.1%
1879					2.3%	2.5%	2.1%
1889						2.7%	2.1%
1899							1.4%

Source: Computed from (Fishlow 1966, Table 10, p. 626)

productivity during the nineteenth century and so lost out whenever there was an alternative means of shipping. Moreover, the price changes in rail understate the real magnitude of the decline in rates as quality, measured in dimensions such as comfort, speed, and the certainty of timely delivery, also accompanied these falling prices.

Fishlow did not allocate railroad productivity growth between specific sources because of the lack of detailed information regarding the timing and extent of use of productivity enhancing devices and techniques, although particularly detailed information is available earlier on in the writings of a Czech engineer Franz Gerstner (Gerstner and Gamst 1997) acting on a commission from Czar Nicholas I of Russia. Fishlow does, however, suggest broad determinants. Consider intensity of use – more freight and passengers per track-mile. Demand for travel and shipping services eventually catches up with the capabilities of the rail system – the country grows into the transportation system that it has. Fishlow's data show intensity of use increased dramatically during the Civil War thanks to the needs of war, while track mileage itself increased only modestly. This probably accounts for a good portion of the much faster rate of productivity during the War decade relative to most other periods. On the other hand, productivity growth in the 1850s was much slower – a period when track mileage was increased sharply especially in areas that were less densely settled than others, like the Midwest.

Fishlow suggests a couple of relatively simple, if blunt, ways to capture the effects of this increased intensity of use. One of these is to recompute productivity assuming that the capital-labor ratio throughout was at its 1909 level. The other assumes that the capital-labor ratio actually increased (as it did in the economy in general) by holding the capital-output ratio constant. In the former case, annual productivity growth over the entire period is reduced from 3.5% to 2.9% – that is, by about 18%. In the latter case, the effect is about twice as large.

This greater intensity of use was also facilitated by various innovations in railroading. For example, the use of the telegraph to control and coordinate train movements (DuBoff 1980; Field 1992), block signaling to control specific stretches of track, and the construction of more sidings and bypasses made it possible to put more trains on each mile of America's predominantly single-track rail system. Moreover, the replacement of iron rails with steel enabled heavier trains to run at higher speeds for years longer before rails needed to be replaced.

These heavier trains and higher speeds were in turn made possible by the development of more powerful locomotives, burning more energy dense fuel (coal) and equipped with flexible trucks that helped guide them around tighter curves without jumping the track. Higher speeds and heavier trains, however, also called for improvements in braking. This was provided by adopting the Westinghouse air brakes which also eliminated the need to have individual brakemen positioned along the train. Getting trains moving and keeping the power applied to all the driving wheels efficiently while dealing with uneven track both front to back (as on a slope) and side to side (as a result of imperfect grading) was accomplished (from an early date) by equalizing levers which transferred loads from side to side and front to back. Fishlow suggests that these and other innovations like automatic couplers (which also drastically reduced mortality among railroad yard workers) collectively

accounted for perhaps half of the measured productivity growth, but without further subdivision and allocation.

So long as American railroads relied upon steam, however, there were some inherent limitations that could not entirely be overcome. Engines burned fuel – initially wood, then coal, and eventually even some oil – to boil water for steam. That steam provided the actual driving force but was exhausted upon use, and eventually the water that it represented had to be replaced. Indeed, locomotives needed more water than fuel, taking on wood or coal only every other stop – which might be as frequent as every 5–10 miles. Efforts to even out the consumption of these two vital components led to experiments with water troughs fed by streams built into the track that scooped water up to replenish the train's supply without it having to stop. By hauling larger water tenders and fuel bunkers (the two often combined with the water tank like an upturned saddle cradling the fuel), this range could be extended 10–20-fold but only by reducing the number of passengers or weight of goods being hauled.

The need for trains to carry their own fuel could be eliminated entirely by electrification. Indeed, this solution was adopted for urban and interurban electric transit systems but was prohibitively expensive for the vast US rail system. There were about 2100 miles of interurban electric track in operation by 1900, and construction boomed up through World War One, by which time there were 15,500 miles of track, principally in seven states, making possible extensive suburbanization around cities such as New York and Chicago as well as smaller cities, such as Indianapolis and Cleveland (Bogart 1906; Hilton and Due 1960). Thereafter, the interurban electric declined in popularity as access to private automobiles increased. For the longer distances involved on trunk line rails, the solution lays in the adoption of the diesel-electric locomotive, where the power to move the train comes from electric motors supplied with energy from diesel-powered generators. These first proved themselves in passenger service in the 1930s but eventually gained acceptance as freight engines (Lytle 1968; Mansfield 1963; Stover 1970). The adoption of electric power in railroading dramatically reduced labor requirements and spurred labor productivity growth in the industry – for example, multiple engines can be controlled from a single location with no need for firemen.

The Social Savings of Railroads

The success of early railroads spelled the beginning of the end for canals. Whereas canals added only about 400 miles during the 1840s, rail mileage grew by over 5000 miles (Carter et al. 2006), and during the 1850s, more miles of canal would be abandoned than were built, but railroad mileage tripled. Railroads simply proved to be superior in the eyes of consumers – whether as a means of shipping or of travel. The question is how big was this superiority? Fogel proposed to determine this by measuring the loss to the economy if the nation had been forced to rely on preexisting means of transportation – canals, steamboats, and wagon – by comparing the actual world with that from a counterfactual state of affairs (Fogel 1964, 1967).

In practice, however, he focused on just canals and wagons, devising a metric that he called "social saving." It was this work that the Nobel Committee cited in its award of the (joint) Nobel Prize in Economics to Fogel in 1993.

Fogel measured this social savings as the consumer surplus associated with switching from shipping by rail to the next best alternative means of shipping while assuming that the supply of shipping services (whatever the sources) was perfectly elastic at the prevailing rates and that the demand for shipping services was completely inelastic. The resulting consumer surplus is the area of a rectangle defined by the difference in the cost of shipping between the best and the next best means of shipping and the quantity shipped and is an upper bound on the actual consumer surplus. Any price elasticity in the demand for shipping would reduce this consumer surplus.

Of course, "the devil is in the details," such as what was shipped and how far. To make the problem more tractable, Fogel restricted his analysis to agricultural commodities – still the dominant sector of the economy until the very end of the nineteenth century in terms of value of output, employment, and movement of goods. He also divided his analysis between long-distance trade – specifically interregional trade – and that closer to home, intra-regional trade. Because good data were available from various government agencies, including the ICC, the Census, and the USDA, Fogel referenced his estimates to 1890, a date by which the railroad was firmly established as the premier mode of transportation (but the picture was not yet clouded by the advent of the internal combustion engine) and its impact should therefore be "large."

The expectation was that interregional trade would dominate. After all, the railroad was most famous for its transcontinental scope and routes like the Illinois Central connecting Chicago with the Gulf Coast. Moreover, resources – and thus trading opportunities – differ more at a distance than close to home. As it worked out, however, the social savings of the railroad in interregional trade were initially calculated as negative – that is, using the railroad imposed a net cost upon the economy rather than a benefit. The reason was simple: the low-cost alternative to rail transportation in long-distance trade was the canal. These offered lower shipping costs than the railroad, and some – like the Erie Canal – were still very much in business in 1890 even if other canals like the Wabash and Erie had been abandoned, hence the negative estimate of the consumer surplus, initially put at -$38 million (Fogel 1964, p. 47, fn. 57). However, this estimate ignored a wide variety of other costs associated with reliance upon water including cargo losses from spoilage and water damage, additional transshipping costs and the costs of shipping from farms to more distant canals, and added inventory costs due to slower shipping and the effects of spring flooding, summer drought, and winter freeze (these issues receive additional critical scrutiny in David 1969). Collectively, Fogel estimated that the net effect of these was to tip the social savings on interregional trade to a positive $73 million.

On the other hand, Fogel determined that the social savings in intra-regional trade were positive just based upon a comparison of the published freight rates without even considering those other factors that made shipping by rail "better." This was

because the next best alternative to rail in intra-regional trade was high-cost, slow wagon transportation – a mode of travel essentially unchanged from the eighteenth century to the early twentieth century. A combination of more expensive and extensive wagon shipping in combination with some limited water shipment would have raised the costs of intra-regional transportation of agricultural commodities by $337 million. This might have been reduced to a social savings of "just" $248 million through a modest, yet feasible, expansion of the canal system that would have brought more of the nation's agricultural land within 40 miles of a water route (Fogel 1964, p. 92, Table 3.10).

By having access to the railroad rather than being forced to rely upon just wagon and canal transportation for the shipment of agricultural commodities in 1890, the United States saved a total of $321 million (=$73 million on interregional trade + $248 million on intra-regional trade). This amounted to about 2.7% of 1890 GNP, a sum that Fogel considered relatively small (certainly in light of all the hyperbole accorded railroads from their inception). More recent estimates of the social saving on agricultural commodities by Donaldson and Hornbeck (2016) revise this figure upward, albeit only modestly to 3.22% of 1890 GNP. The real thrust of Donaldson and Hornbeck's analysis, however, is to estimate the impact that the coming of the railroad (or its loss) had upon agricultural land values (a calculation that Fogel also made). To make this calculation, they rely upon the trade model of Eaton and Kortum (2002) to develop a measure of the railroad's impact upon "market access," based upon least-cost shipping between markets and the importance of those markets (like population), a metric that has found increasingly widespread use. Based upon the change in market access, they conclude that the elimination of railroads in 1890 would have decreased land values in the United States by 60% – a large amount. Indeed, they call the railroads "critical" to the agricultural sector in 1890. Moreover, in contrast to Fogel's arguments, they do not believe that these losses could have been offset in any substantial way by extending the country's network of waterways.

Donaldson and Hornbeck's estimates, like those by Fogel, suffer from the limitation that they are designed to reflect the gain to agriculture from access to railroads rather than all goods, although as producers of relatively low value-to-weight items, agriculture would bear the heaviest burden of higher transport costs. Moreover, they ignore the impact of the railroad upon passenger traffic, which was an important source of revenue for the railroads and gain to the economy before motor vehicle traffic siphoned off that trade. Boyd and Walton (1972) estimate passenger losses at about $344 million in 1890 – more than doubling the loss estimated by Fogel. Thus, these losses combined with Fogel's own rough estimate of aggregate freight traffic savings would raise the total social savings to over $900 million or about 7.3% of GNP.

Such losses begin to look "large." Indeed, in their work, Donaldson and Hornbeck conclude that in the absence of the railroad, declines in population and worker utility would have generated substantial losses in GNP and aggregate welfare, even without considering the effects upon agglomeration and other externalities. Contemporaneous estimates to Fogel's made by Fishlow for 1860 concluded that the social savings

amounted to perhaps $155 million on (all) freight and $70 million on passengers. This loss would have been about 5% of 1860 GNP – and likely would have been much larger by 1890 – perhaps as much as 15% of GNP (Fishlow 1965).

The uncertainties and margins of error surrounding social savings estimates arise from both practical and theoretical issues. Among the practical concerns are the role of externalities such as speed, comfort, and safety (Lebergott 1966), the impact of inclement weather on city food supplies – and the ability of large cities to survive – in a world without railroads, and whether various biases cancel each other out (David 1969). Perhaps the most serious concerns are theoretical and concern the general equilibrium effects, not captured in the comparative statics approach of the counterfactual, of eliminating the railroad. Efforts to model these (e.g., by Williamson 1974, 1975) result in larger estimates than made by Fogel – as one might expect. Calibration is key, and other general equilibrium approaches (Donaldson and Hornbeck 2016; Kahn 1988) have resulted in estimates that lie somewhere between those by Fogel and Williamson.

Whether the railroads were "decisive" in terms of American growth and development seems unlikely, that they were important and their contribution large, however, seems more certain. Much of Fogel's critique was rhetorical, aimed at destroying a straw man that he termed the "axiom of indispensability" and moderating the most extreme hyperbolic claims of both contemporaries like Ralph Waldo Emerson (1903 364), who called the railroad "a magician's rod in its power to evoke the sleeping energies of land and water," and earlier generations of historians (up to and including Rostow) who described the railroad as "essential to the development of capitalism" and "responsible for much of the territorial division of labor" (Bolino 1961, 175).

Concluding Remarks

Although not decisive or critical to American growth and development in the way that generations of observers had claimed prior to Fogel's analysis and critique, railroads were important – arguably more important than Fogel's rhetoric would lead one to believe. Not only did they shape – or rather reshape – trade flows within and from the country for more than a century, but they played a major role in inter- and intra-regional specialization. This promoted economies of scale and agglomeration economies among other externalities by improving the means of transportation and communication. They would also have other, more diverse, and less obvious impacts, especially in areas of business that have not been a focus of cliometrics – for example, management (e.g., the emergence of the multidivisional business), accounting (both cost accounting for internal control and broader financial reporting for investors), and labor relations (e.g., labor unions like the United Brotherhood of Sleeping Car Porters and tracking and monitoring the performance of a geographically dispersed labor force). These aspects receive fuller but still largely qualitative treatment, in the writing of business historians, especially Alfred Chandler (Chandler 1965, 1977, 1979; John 2008) although some aspects of business organization are beginning to attract cliometricans (see, e.g., Guinnane et al. 2007; but especially Hilt 2008).

References

Adams CF, Adams H, Walker FA (1871) Chapters of Erie, and other essays. J. R. Osgood and company, Boston

Agricola G, Hoover H, Hoover LH (1912) Georgius Agricola De re metallica. Published for the translators by The Mining magazine, London

Arthur WB (1989) Competing technologies, increasing returns, and lock-in by historical events. *Econ J* 99(394):116–131. https://doi.org/10.2307/2234208

Arthur WB (1994) Increasing returns and path dependence in the economy, economics, cognition, and society. University of Michigan Press, Ann Arbor

Atack J (2013) On the use of geographic information systems in economic history: the American transportation revolution revisited. *J Econ Hist* 73(2):313–338

Atack J (2015) Historical geographic information systems (GIS) database of U.S. Railroads. Vanderbilt University. Available from https://my.vanderbilt.edu/jeremyatack/files/2015/12/RR1826-1911Modified122715.zip

Atack J (2018) Creating historical transportation shapefiles of navigable rivers, canals, and railroads for the United States before World War I. In: Gregory I, DeBats D, Lafreniere D (eds) The Routledge companion to spatial history. Routledge, Milton Park, pp 169–184

Atack J, Bateman F, Haines M, Margo RA (2010) Did railroads induce or follow economic growth? Urbanization and population growth in the American midwest, 1850-1860. *Soc Sci Hist* 34 (2):171–197

Atack J, Brueckner J (1982) Steel rails and American railroads, 1867-1880. *Explor Econ Hist* 19 (October):339–359

Atack J, Margo RA (2011) The impact of access to rail transportation on agricultural improvement: the American midwest as a test case, 1850-1860. *J Transp Land Use* 4(2):5–18

Bastiat F, Stirling PJ (trans) (1922) Economic sophisms. G. P. Putnam's Sons, New York

Bathe G, Bathe D (1935) Oliver Evans: a chronicle of early American engineering. Historical Society of Pennsylvania, Philadelphia

Bleakley H, Lin J (2012) Portage and path dependence. *Q J Econ* 127(2):587–644. https://doi.org/10.1093/qje/qjs011

Bogart EL (1906) Economic and social effects of the inter-urban electric railway in Ohio. *J Polit Econ* 14(10):585–601

Bolino AC (1961) The development of the American economy. C. E. Merrill Books, Columbus

Boyd H, Walton GM (1972) The social savings from nineteenth-century rail passenger service. *Explor Econ Hist* 9(1):233–254. https://doi.org/10.1016/0014-4983(71)90059-3

Carter SB, Gartner SS, Haines MR, Olmstead AL, Sutch R, Wright G, Cain LP (eds) (2006) Historical statistics of the United States millennial edition online. Cambridge University Press, New York

Caves DW, Christensen LR, Swanson JA (1981) Productivity growth, scale economies, and capacity utilization in U.S. railroads, 1955-74. *Am Econ Rev* 71(5):994–1002

Chandler AD (1954) Patterns of American railroad finance, 1830-50. *Bus Hist Rev* 28(3):248–263. https://doi.org/10.2307/3111573

Chandler AD (1965) The railroads, the nation's first big business; sources and readings. The forces in American economic growth series. Harcourt, New York

Chandler AD (1977) The visible hand: the managerial revolution in American business. Belknap Press, Cambridge, MA

Chandler AD (1979) The Railroads, pioneers in modern management. History of management thought. Arno Press, New York

Condit CW (1980) The port of New York. University of Chicago Press, Chicago

Craig LA, Palmquist R, Weiss T (1998) Transportation improvements and land values in the antebellum United States: a hedonic approach. *J Real Estate Financ Econ* 16(2):173–189

David PA (1969) Transport innovation and economic growth: professor Fogel on and off the rails. *Econ Hist Rev* 22(3):506–525. https://doi.org/10.1111/j.1468-0289.1969.tb00186.x

Davis JH (2006) An improved annual chronology of U.S. business cycles since the 1790s. *J Econ Hist* 66(1):103–121
Dickinson HW, Titley A (1934) *Richard Trevithick: the engineer and the man*. The University press, Cambridge
Donaldson D, Hornbeck R (2016) Railroads and American economic growth: a "market access" approach. *Q J Econ* 131(2):799–858. https://doi.org/10.1093/qje/qjw002
DuBoff RB (1980) Business demand and the development of the telegraph in the United States, 1844-1860. *Bus Hist Rev* 54(4):459–479. https://doi.org/10.2307/3114215
Duran X (2013) The First U.S. transcontinental railroad: expected profits and government intervention. *J Econ Hist* 73(1):177–200. https://doi.org/10.1017/S0022050713000065
Eaton J, Kortum S (2002) Technology, geography, and trade. *Econometrica* 70(5):1741–1779
Emerson RW (1903) "The Young American." The Complete Works of Ralph Waldo Emerson. Riverside Press, Cambridge
Fels R (1959) American business cycles, 1865–1897. University of North Carolina Press, Chapel Hill
Fernald RH (1918) Is our fuel supply nearing exhaustion. *Eng Eng* 25(April):835–836
Field AJ (1992) The magnetic telegraph, price and quantity data, and the new management of capital. *J Econ Hist* 52(2):401–413
Field AJ (1978) Sectoral shifts in antebellum Massachusetts: a reconsideration. *Explor Econ Hist* 15 (2):146–171
Fishlow A (1965) American railroads and the transformation of the antebellum economy, Harvard economic studies, vol 127. Harvard University Press, Cambridge, MA
Fishlow, Albert. 1966. "Productivity and technological change in the railroad sector, 1840–1910." In Output, employment, and productivity in the United States after 1800, Dorothy, Brady, 583–646. New York: Columbia University Press for NBER
Fleisig H (1975) The Central Pacific railroad and the railroad Land Grant controversy. *J Econ Hist* 35(3):552–566
Fogel RW (1960) The Union Pacific railroad; a case in premature enterprise. The Johns Hopkins University studies in historical and political science. Johns Hopkins Press, Baltimore
Fogel RW (1962) A quantitative approach to the study of railroads in American economic growth: a report of some preliminary findings. *J Econ Hist* 22(2):163–197
Fogel RW (1964) Railroads and American economic growth: essays in econometric history. Johns Hopkins Press, Baltimore
Fogel RW (1967) The specification problem in economic history. *J Econ Hist* 27(3):283–308
Gerstner FA, Gamst FC (1997) Early American railroads: Franz Anton Ritter von Gerstner's Die innern Communicationen (1842–1843). Stanford University Press, Stanford
Gross DP (2016) The ties that bind: railroad gauge standards and internal trade in the 19th century U.S. Harvard Business School working paper: 17-044
Guinnane T, Harris RON, Lamoreaux NR, Rosenthal J-L (2007) Putting the corporation in its place. *Enterp Soc* 8(3):687–729
Hallberg MC (2004) Railroad database. Old Railroad History. http://oldrailhistory.com/
Harley CK (1982) Oligopoly Agreement and the timing of American railroad construction. *J Econ Hist* 42(4):797–823
Higgs R (1987) Crisis and leviathan: critical episodes in the growth of American government. Oxford University Press, New York
Hilt E (2008) When did ownership separate from control? Corporate governance in the early nineteenth century. *J Econ Hist* 68(3):645–685
Hilton GW, Due JF (1960) The electric interurban railways in America. Stanford University Press, Stanford
Hughes JRT (1991) The governmental habit redux: economic controls from colonial times to the present, 2nd edn. Princeton University Press, Princeton
Jenks LH (1944) Railroads as an Economic Force in American Development. *J Econ Hist* 4(1):1–20
John RR (2008) Turner, Beard, Chandler: progressive historians. *Bus Hist Rev* 82(2):227–240

Kahn C (1988) The use of complicated models as explanations: a re-examination of Williamson's late nineteenth century America. *Res Econ Hist* 11:185–216
Kent DH (1948) The Erie War of the gauges. *Pa Hist* 15(4):253–275
Lebergott S (1966) United States transportation advance and externalities. *J Econ Hist* 26:437–465
Libecap GD, Hansen ZK (2002) "Rain follows the plow" and Dryfarming Doctrine: The Climate Information Problem and Homestead Failure in the Upper Great Plains, 1890-1925. *J Econ Hist* 62(1):86–120
Liebowitz SJ, Margolis SE (1994) Network externality: an uncommon tragedy. *J Econ Perspect* 8 (2):133–150
Liebowitz SJ, Margolis SE (1995) Path dependence, lock-in, and history. *J Law Econ Org* 11 (1):205–226
Lytle RH (1968) The introduction of diesel power in the United States, 1897-1912. *Bus Hist Rev* 42 (2):115–148
MacAvoy PW (1965) The economic effects of regulation; the trunk-line railroad cartels and the Interstate Commerce Commission before 1900. The MIT Press, Cambridge, MA
Mansfield E (1963) Intrafirm rates of diffusion of an innovation. *Rev Econ Stat* 45(4):348–359. https://doi.org/10.2307/1927919
Martin A (1971) Enterprise denied; origins of the decline of American railroads, 1897–1917. Columbia University Press, New York
Mercer LJ (1969) Land Grants to American railroads: social cost or social benefit? *Bus Hist Rev* 43 (2):134–151
Mercer LJ (1974) Building ahead of demand: some evidence for the land grant railroads. *J Econ Hist* 34(2):492–500
Mitchell BR (2007) International historical statistics, 5th edn. Three volumes: Europe, The Americas and Africa, Asia, and Oceania volumes. Palgrave Macmillan, Basingstoke
Modelski AM (1987) Railroad maps of North America: the first hundred years. Bonanza Books, New York
O'Sullivan M (2007) The expansion of the U.S. Stock Market, 1885–1930: historical facts and theoretical fashions. *Enterp Soc* 8(3):489–542
Olmstead AL, Rhode PW (2008) Creating abundance: biological innovation and American agricultural development. Cambridge University Press, New York
Paxson FL (1914) The railroads of the "Old Northwest" before the Civil War. *Trans Wisconsin Acad Sci Arts Lett* 17(1):247–274
Puffert DJ (1991) The economics of spatial network externalities and the dynamics of railway gauge standardization. Unpublished PhD thesis, Microform, Department of Economics, Stanford University
Puffert DJ (2000) The standardization of track gauge on North American Railways, 1830-1890. *J Econ Hist* 60(4):933–960. https://doi.org/10.2307/2698082
Puffert DJ (2009) Tracks across continents, paths through history: the economic dynamics of standardization in railway gauge. University of Chicago Press, Chicago
Raban J (1996) *Bad land: an American romance*. Pantheon Books, New York
Riegel RE (1931) Western railroad pools. *Miss Val Hist Rev* 18(3):364–377. https://doi.org/10.2307/1891405
Rostow WW (1956) The take-off into self-sustained growth. *Econ J* 66(261):25–48. https://doi.org/10.2307/2227401
Rostow WW (1960) The stages of economic growth, a non-Communist manifesto. Cambridge University Press, Cambridge
Rostow WW (1963) The economics of take-off into sustained growth; proceedings of a conference held by the International Economic Association. Macmillan, London
Schiffman DA (2003) Shattered rails, ruined credit: financial fragility and railroad operations in the Great Depression. *J Econ Hist* 63(3):802–825
Schumpeter JA (1939) Business cycles; a theoretical, historical, and statistical analysis of the capitalist process, vol 2 vols, 1st edn. McGraw-Hill, New York

Sellers C (1886) Oliver Evans and his inventions. *J Frankl Inst* 122(1):1–16. https://doi.org/10.1016/0016-0032(86)90114-6
Stover JF (1970) The life and decline of the American railroad. Oxford University Press, New York
Stover JF (1975) History of the Illinois Central Railroad, Railroads of America. Macmillan, New York
Stover JF (1987) History of the Baltimore and Ohio railroad. Purdue University Press, West Lafayette
Swisher SN (2017) Reassessing railroads and growth: accounting for transport network endogeneity. Cambridge, Cambridge Working Papers in Economics: 1718
Taylor GR (1951) The transportation revolution 1815–1860. Holt, Rinehart & Winston, New York
Taylor GR, Neu ID (1956) The American railroad network, 1861–1890. Studies in economic history. Harvard University Press, Cambridge, MA
Temin P (1964) Iron and steel in nineteenth-century America, an economic inquiry. MIT monographs in economics. The MIT Press, Cambridge, MA
Thorp WL, Thorp HE, Mitchell WC (1926) Business annals: United States, England, France, Germany, Austria, Russia, Sweden, Netherlands, Italy, Argentina, Brazil, Canada, South Africa, Australia, India, Japan, China, Publication of the National bureau of economic research, incorporated, vol 8. National Bureau of Economic Research, Inc., New York
Turner GE (1953) Victory rode the rails; the strategic place of the railroads in the Civil War, 1st edn. Bobbs-Merrill, Indianapolis
Ulen TS (1980) The Market for Regulation: The ICC from 1887 to 1920. *Am Econ Rev* 70 (2):306–310
United States. Census Office (1864) Agriculture of the United States in 1860: compiled from the original returns of the eighth census. Government Printing Office, Washington
United States Congress House (1873) Report of the select committee. . .appointed under the resolution of January 6, 1873, to make inquiry in relation to the affairs of the Union Pacific railroad company, the credit mobilier of America, and other matters. US Government Print Office, Washington
White JH (1968) American locomotives; an engineering history, 1830–1880. Johns Hopkins Press, Baltimore
Wicker E (1960) Railroad investment before the Civil War. In: Parker WN (ed) Trends in the American economy in the nineteenth century. Princeton University Press for the NBER, Princeton, pp 503–546
Williamson JG (1974) Late nineteenth-century American development: A general equilibrium history. Cambridge University Press, New York
Williamson JG (1975) The railroads and midwestern development 1870-1890: a general equilibrium history. In: Klingaman DC, Vedder RK (eds) Essays in nineteenth century economic history: the old Northwest. Ohio University Press, Athens, pp 269–352

Clio on Speed

A Survey of Economic History Research on Transport

Dan Bogart

Contents

Abstract

Cliometrics has made major advances in the historical analysis of transportation through better measurement, economic modelling, and estimation. This essay surveys several topics and recent research. First, the revolutionary changes in transport over the last 300 years are reviewed, including analysis on the rate and sources of productivity growth. It is argued that macro-inventions, like steam power, were important but many incremental innovations mattered too. Second, the effects of transport improvements on market integration, trade, urbanization, and aggregate income are examined. Much research focuses on questions of magnitude. In other words, how big are the effects of transport improvements like railroads? Perhaps surprisingly, there is still disagreement about the relative importance of transport even with the applications of new tools and Geographic Information Software. Two novel areas of research concern mortality and

D. Bogart (✉)
Department of Economics, University of California, Irvine, CA, USA
e-mail: dbogart@uci.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_67

persistence. Studies show that transport improvements contributed to higher mortality and influenced population density long after transport technologies became obsolete. These studies yield new perspectives on the effects of transport. Third, this essay examines why transport services were more efficient in some economies with a focus on the role of institutions as a fundamental factor. Evidence suggests institutions influenced investment in transport networks and the degree of public and private ownership. Ownership mattered for transport efficiency, and government involvement sometimes improved outcomes. Overall, cliometrics research on transport offers many insights on a historically important sector.

Keywords
Transport · History · Market integration · Urbanization · Institutions · Agglomeration · Shipping · Railways

Introduction

Transportation has undergone dramatic changes through the ages. Scholars from a wide range of fields have sought to understand the implications and explain those changes. Cliometrics has made key contributions to this broader literature through innovations in measurement, economic modelling, and estimation. This essay reviews the contributions of cliometrics to seven key topics. The first topic concerns the so-called transport revolution. There is much research tracking transport costs and travel speeds over time, making it possible to quantify the rate of change. This research shows that transport has undergone dramatic productivity improvements over the last 300 years. What caused these changes? Were macro-inventions, like steam power, the primary driver? The literature generally supports the importance of macro-inventions, but at the same time many studies show that productivity changes were incremental and contextual. Several examples are given below.

The second section examines transport, market integration, and trade. The trends in market integration are reviewed. Next the effects of transport improvements are examined. Perhaps surprisingly, some research has found that transport improvements were not the most important driver of market integration. But other research finds the opposite. Moreover, in one study, poorly designed transport regulation is argued to be a main driver of market disintegration. The connection between integration and transport is not fully settled.

The third section focuses on the contribution of transport improvements to aggregate incomes. The main question is the following: how much economic growth was attributable to transport improvements? The traditional cliometric methodology calculates the social savings. Recently, cliometrics has developed new data and econometric methods. As will be discussed below, the new research sometimes yields different estimates of transport's contribution. But in some cases, it appears that the social savings provide a reasonable approximation to the aggregate incomes gains from transport.

The fourth section examines the external effects of transport. Greater urbanization is one key example. Due to lower transport costs, individuals may change their location and affect their neighbors' productivity. Several works analyzing the effects of railroads on city population are reviewed. All find that cities with better railroad access increased their population relative to cities with poor access. An outstanding issue concerns possible reorganization effects, where transport causes population loss in areas with worse transport access. The effects of highways on central city population provide a good example of reorganization.

Higher mortality is another external effect of transport improvement. Pollution-related health problems are a major concern in the twenty-first century. This section also discusses how transportation-related pollution was an issue in the nineteenth century.

The fifth section examines the connection between transport and the persistence of economic activity across space. Populations tend to settle where there are natural advantages, like a navigable river. Transport improvements can create new locational advantages, which encourage population resettlement. As time goes on some transport technologies become obsolete and one might think that population would resettle elsewhere. However, the literature shows migration from obsolete transport hubs does not necessarily happen.

The sixth section examines how institutions affect transport. Institutions have long been highlighted in economic history as a fundamental cause of development. In the transport sector, institutions have been linked with the extent and quality of infrastructure networks. Several works are discussed which show that political institutions had a large effect on network development. This suggests that transport is a key channel through which institutions affected development in the past and today.

The seventh and final section discusses public and private ownership in transport. There are theoretical arguments in favor of both forms of ownership. Studies on the impacts of ownership and cost efficiency are discussed. Also, the role of institutions in determining railroad nationalizations in the nineteenth and early twentieth centuries are analyzed. This ties with current debates about when government is more effective in delivering transport services.

The Transport Revolution

The transport revolution is one of the most significant developments in human history and equals the industrial revolution in significance. According to historians (e.g., Taylor 2015; Bagwell 2002), the transport revolution involved dramatic reductions in the time and costs of moving goods and people and in the cost of communication. Britain's transport revolution provides an example of these significant changes. The real freight rate by rail in 1870 was 5% of the real freight rate by road in 1700. The speed of road travel in 1700 was only 8% of the speed by train in 1870 (Bogart 2014). Dispatch times for letters in 1860 were 33% of what they had been in 1820 (Kaukiainen 2001). For comparison, one can look at the falling price of high-valued manufactured goods, which were impacted by mechanization during the British industrial revolution. In 1840, the real price of fine printed cloth, known as calicoes,

was 25% of their price in 1770, implying a 2% annual rate of decline (Harley 1998). The overland freight rates mentioned above declined at an annual rate of 1.7% and dispatch times declined at an annual rate of 2.7%. These and other examples reveal that the transport sector was revolutionized to the same extent as manufacturing.

Much attention has focused on dating the transport revolution in various countries and for inland versus coastal and ocean shipping. In some cases, it is hard to identify the timing of significant change due to the gradual process of improvements. For example, there is evidence that shipping freight rates started to decline in the European trades during the early 1600s (Menard 1991). There is also evidence for greater productivity and lower sailing times in the transatlantic and East Indian trades during the late 1700s (Rönnbäck 2012; Solar 2013). The big changes in shipping productivity and freight rates arguably occurred in the mid-1800s with the introduction of steam power. Steamships were a global technology and brought lower freight rates to much of the world's trade, including India and China. Real shipping freight rates continued to decline between the late nineteenth and early twentieth century. The peak for productivity in pre-World War I ocean shipping was 1913 (Mohammed and Williamson 2004). For the next 50 years, shipping freight rates fluctuated significantly, mainly due to wars and policy decisions. More recently, that is in the last 50 years (1960–2010), freight rates declined once again as ships got larger and accommodated containers. The downward trend in freight rates in the last 50 years looks like the 1870–1913 era.

Inland transport costs also evolved over time. As an illustration, Walton and Rockoff (2013) summarize freight rates in the USA up to 1900. Around 1800 there were two main forms of inland transport: wagon and river. Wagon freight rates were 34 cents a ton mile, while river transport was between 2 and 10 cents a ton mile depending on whether boats travelled upstream or downstream. Significant changes occurred starting in the 1820s. The upstream river rate dropped to 1 cent a ton mile with the arrival of the steamboat. Building canals brought cheap water transport to areas without navigable rivers (around 5 cents a ton). Wagon rates fall to 15 cents a ton mile due to better roads. The introduction of railroads around 1840 was the next phase. It brought low freight rates and greater speed to areas that did have canals because of geography. Freight rates by rail started at 6 cents a ton mile and declined to less than 1 cent a ton mile by 1900. Note that freight rates continued to fall on railroads and canals after they were introduced. These technologies unleashed a period of exceptionally high productivity growth in the USA.

These patterns of lower inland freight rates and travel times were not seen worldwide, however. In a few economies, mostly northwestern Europe and the USA, there were improvements in inland transport starting in the 1700s (De Vries 1981; Bogart 2005; Gerhold 2014). But in other economies, including India, China, and Japan, it was only with the railway that inland freight rates and travel times fell (see Huenemann 1984; Ericson 1996; Kerr 2007). One explanation is that these economies did not improve their roads and waterway networks to the same degree as northwestern Europe and the USA in the late eighteenth and early nineteenth centuries.

The twentieth century has witnessed another transport revolution associated with the automobile and airplane. Their economic impact has been large since the

1950s. Air travel is the most recent example. It led to dramatic declines in the cost of passenger travel and improvements in communication. The average speed of air passenger travel went from 180 mph in 1950 to 408 mph in 1990 (Carter et al. 2006). That is, speeds increased by 2.25 times in 40 years. One effect in the USA has been the concentration of employment near cities with airports (Brueckner 2003).

Over the longer term, consider the following facts. In 1700, most wealthy individuals travelled on coaches at speeds less than 3 mph. In 2000, the wealthy travel in airplanes at more than 500 miles an hour. That implies long distance travel speeds are 166 times faster today than 300 years ago for the rich! In 1700, the poor did not usually travel, but when they did it was generally by foot, moving at 2–2.5 mph. In 2000, the poor often travel at speeds ranging between 20 and 60 miles an hour depending on congestion, roads, and vehicle types. Therefore, travel speeds for the poor are 10–24 times faster than 300 years ago. In the long-run perspective, it is evident that transportation has radically changed, but the rich have enjoyed more gains than the poor. Transport improvements are *not* neutral in their distributional effects.

This brief review should raise questions about the causal factors behind the transport revolution. One issue concerns the importance of large-scale technological changes (i.e., macro-inventions) versus incremental changes (i.e., micro-inventions). The thrust of much research in cliometrics points to the importance of incremental changes without minimizing macro-changes like steam power, the ship container, and the internal combustion engine.

Shipping provides a good example of this research. One narrative is that the productivity of sailing ships changed little over the centuries in comparison to advances associated with steamships. Harley (1988) made this argument starkly in the case of British shipping. Using a real freight rate series, Harley argued that little changed from 1740 to 1830, but after 1830, when steamships became more common, there was a precipitous decline in real freight rates (see Fig. 1). Harley's series challenged those of previous scholars like North (1958) who suggested that transatlantic freight rates declined as much in the age of sail as the age of steam. Harley used a different freight rate series and price deflator. As is often the case in cliometrics, measurement matters.

While steam was clearly important, subsequent research has shown that Harley over-states stagnation in the age of sail. From the 1780s, copper-sheathing protected wooden hulls from ship worm, significantly prolonging ships' lives and incidentally making ships faster and more maneuverable. This new technology reduced freight costs by about a third in the trade between Europe and Asia and led to a significant fall in mortality in the slave trade in the late eighteenth century (Solar 2013; Solar and Rönnbäck 2015). Menard (1991) and Shepherd and Walton (1972) show that freight rates in the transatlantic trade declined in real terms in the seventeenth and eighteenth centuries. In the case of the Chesapeake to London tobacco trade, real freight rates declined at an annual rate of 2.1% up to 1774. These authors attribute over 80% of the decline in freight rates to lower port times and specifically better packaging.

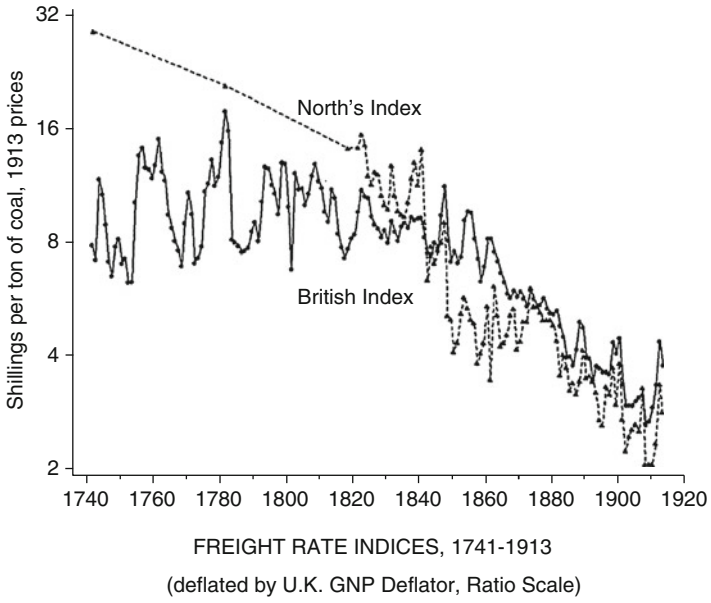


Fig. 1 Two shipping Freight rate indices. (Source: Harley 1988)

It is striking that something seemingly simple, like packaging, could make a key contribution to the transport revolution. The emphasis on packaging relates to an important point about transport costs: the product being shipped matters. Tobacco, cotton, and sugar were amendable to packaging improvements in the pre-industrial world. Other products like coal were not. The combination of products was also important. In another article, Harley (2008) shows that nineteenth century freight rates on grain became very low between New York and Liverpool in part because grain was used as ballast on ships transporting live cattle. Ballast is necessary because it gives stability to the ship. Usually ballast is transported for free or at very low freight rates, because they enable the more valuable good to be shipped. The implication is that absent the live meat trade, and the need for ballast, freight rates on grain would have been much higher.

Overall, the case of shipping suggests that large-scale technological changes were moments of discontinuous productivity advances, but they were often equaled by many incremental changes. Arguably a similar statement applies to productivity growth more generally. In that sense, transport is not different from manufacturing and some services.

Transport Improvements, Market Integration, and Trade

Greater market integration is one consequence of the transport revolution. Market integration relates to the co-movement of prices in different markets as defined in the law of one price. The law states that the price difference for a commodity in two

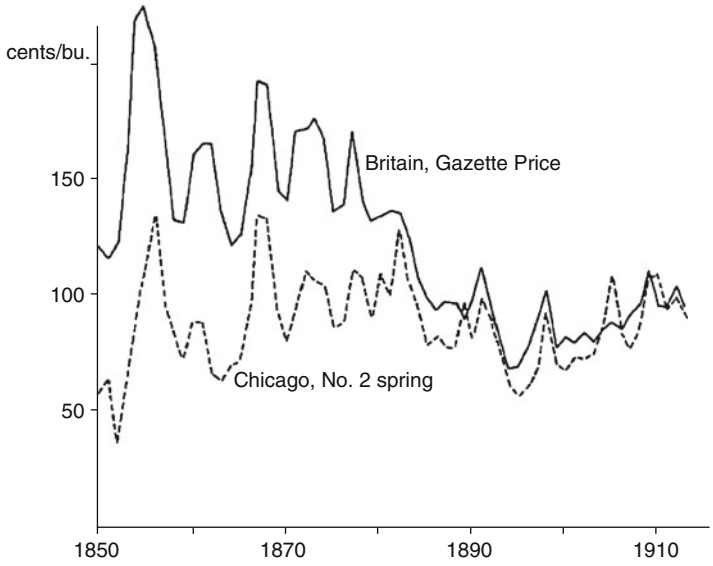


Fig. 2 Wheat prices in Britain and Chicago, 1850–1913. (Source: Harley 1980)

markets should be smaller or equal to transport and transaction costs, the so-called no-arbitrage gap. If the price difference is greater than the gap, trade will occur (commodities will move from the low to the high price market) and the price differences will shrink to the gap. The so-called "speed of adjustment" is the rate at which prices converge to the gap (Federico and Persson 2006).

Perhaps the most striking evidence of integration is the nineteenth-century transatlantic wheat market. As an illustration, Harley (1980) documents the difference in wheat prices between Britain and Chicago between 1850 and 1913 (see Fig. 2). The wheat price difference was high in the 1850s and 1860s (around 50 cents a bushel). By the 1890s and through the early 1910s, the price difference was negligible (around 5 cents a bushel). There was a large growth in trade between the USA and Britain accompanying the greater integration of wheat markets. In fact, global trade grew so much that historians have dubbed the late nineteenth century the first "era of globalization." Based on the wide diffusion of railroads and steam ships in the late nineteenth century, it would appear they were the main drivers of this market integration and international trade.

But in cliometrics reasonable theories are not always supported by the data. Jacks (2006) provides evidence that railroads and canals were not the main drivers of integration. This important paper aims to quantify the contribution of various factors to market integration. It proposes an empirical model to explain integration in wheat markets. The model has variables for technological factors, like railroads and canals, and institutional variables, like tariffs and common monetary regimes. The main specification is the following:

$$\text{Market integration}_{ijt} = \alpha \cdot \text{technology}_{ijt} + \beta \cdot \text{institutions}_{ijt} + \eta \cdot x_{ij} + \delta_t + \varepsilon_{ijt}$$

where market integration$_{ijt}$ is the estimated transaction costs or speed of adjustment between city pair i and j in year t, technology$_{ijt}$ and institutions$_{ijt}$ include the key explanatory variables. The remaining variables are controls including year fixed effects, distance, and common borders. Jacks finds a striking conclusion: institutional factors were more important in explaining the decline in price gaps. For example, having a railroad connection reduced the transaction cost by 0.02 standard deviations, while shared adherence to the gold standard reduced the transaction cost by 0.22 standard deviations. In short, institutional variables were the key driver of market integration during the late nineteenth century according to this model.

This essay will return to the topic of institutions later, but it is worth noting that other studies continue to argue that transport was an important driver of market integration and trade. For example, Pascali (2017) argues that the steamship had a major impact on global trade from 1850 to 1900. The introduction of the steamship led to an asymmetric change in trade costs among countries. Before this invention, trade costs depended on wind patterns, and some countries were favorably located. But wind patterns mattered much less after the steamship was introduced. Pascali uses this fact to identify large effects of steamships on trade. Donaldson (2018) reaches a similar conclusion for railways in colonial India using different data and a different identification strategy.

There is another study which also shows the importance of transport policies, but the effects are towards market disintegration. There is evidence for disintegration of markets in the inter-war period between 1920 and 1940. For example, Federico and Sharp (2013) show there was considerable disintegration of agricultural markets in the USA. They argue that disintegration was largely due to an increase in railroad freight rates. Why did that happen? According to Federico and Sharp, freight rates rose because of the 1920 Transportation Act and legislation that followed. There was a shift from a generally pro-competitive policy among US regulators to one in which railroad companies were subsidized through high regulated rates and were protected from competition by the emerging trucking sector. Matters were exacerbated by the price deflation occurring in the 1930s and the failure of regulators to lower rail rates as input prices fell. Arguably, it was the lobbying pressure of the railroad companies that kept the rates high. This case provides an excellent illustration of regulatory capture in transport. More generally, this study and the previous ones suggest that more research is needed to understand the complex relationship between transport, market integration, and trade.

Transport Improvements and Income Gains

Dramatic reductions in transport costs have potentially large implications for incomes and living standards. Many historians in the early twentieth century assumed transport improvements, like railroads, were indispensable to the growth in incomes during the nineteenth century. Several decades later, some of the most

famous cliometricians questioned whether this was true. Fogel (1964) and Fishlow (1965) approached this issue using the "social savings" methodology. The social savings are approximated by the formula: $(P_w - P_r)*Q_r$, where P_w is the price of transport using the old technology, P_r is the price of transport with railroads, and Q_r is the quantity or ton-miles shipped by railroads. The social savings are meant to approximate the consumer surplus created by railroads at some benchmark date (see Fig. 3). The price P_r is the marginal cost of delivering railroad services and P_w is the marginal cost of some alternative technology like wagons. The marginal costs equal the prices charged under the assumption of perfect competition in transport markets. The shaded region is the social savings, which is $(P_w - P_r)*Q_r$ after a correction for the elasticity of demand. Notice that a more elastic demand (i.e., a smaller slope for curve D) implies a lower social savings. A similar methodology has been used to calculate time savings from transport. Leunig (2006) provides one example for British railways.

The social savings from railroads are often compared with national income to assess the quantitative significance. Fogel and Fishlow estimated that national income in the USA would have been only 3–4% lower in 1860 or 1890 without railroads. In other words, railroads accounted for only a small portion of the income gains between 1840 and later decades. Many cliometricians have followed in their footsteps. For example, Summerhill (2005) and Herranz-Loncán (2014) provide a summary of social savings estimates for railways in Latin America. The savings equal about 25% of total GDP in Argentina, Mexico, and Brazil by 1910–1913. These are huge effects, and much larger than the USA. However, Herranz-Loncán (2014) finds that the social savings were less than 5% in Uruguay, Peru, and Columbia. In the case of Peru and Columbia, their railway networks were small, and their services were poor in quality.

The social savings methodology has some virtues. Ease of calculation is one. All that is needed are aggregate traffic volumes and estimates of average freight rates or speeds. For railroads, the social savings can be estimated in most economies by 1913 because national statistics are usually available by that date. This allows for cross-

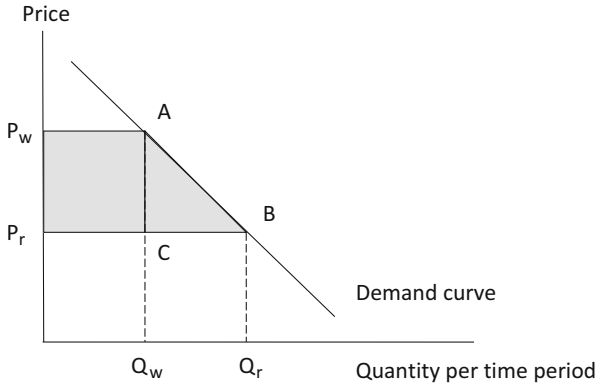


Fig. 3 The social savings of railroads depicted as a measure of consumer surplus. (Source: Author's design)

county comparisons of the gains from railroads and investigation of the factors that led to different impacts.

Nevertheless, the social savings methodology is controversial. Critics point to several problems (see McClelland 1972). First, it is not clear what the price of the alternative transport would have been in the absence of railways (P_w in the formula $(P_w - P_r)^* Q_r$). Congestion would have likely increased on wagon roads and canals if it had to handle the traffic volume associated with railways. The cost of using alternative transport modes is arguably underestimated as a result. Second, the social savings calculation omits backward and forward linkages. Railways increased demand for iron and steel as inputs and thus they fostered the development of these important industries. There are also changes in economic geography to consider. Lower transport costs can lead to agglomeration, which has positive effects on productivity.

Recently, cliometrics has used new data and methods to analyze the contribution of transport improvements to income. One good example is Atack and Margo (2011), who study the link between railroads and land values in the USA. Land values are informative because most externalities, like agglomeration, should be reflected in land income as it is the fixed factor of production. Atack and Margo's railroad data is derived from digitized nineteenth century maps and census materials revealing farm outcomes like land area, land which has been improved, and values in dollars. They cleverly employ Geographic Information Software (GIS) to generate spatial variables of interest. Similar GIS data are also employed to study European railways (see Martí-Henneberg 2013).

Atack and Margo use a difference-in-differences estimator (DID) to measure the impact of the railroad. In this case, DID compares the change in farm outcomes between 1850 and 1860 in a treatment group of counties – identified as those getting railroads during the period – versus a control group that did not. Their estimating equation is similar to the following:

$$\text{Log land improved}_{it} = \alpha \cdot \text{Railroad}_{it} + \eta_i + \delta_t + \varepsilon_{it}$$

where the dependent variable is the natural log of acres of improved farmland in county i, η_i is a fixed effect for the county, δ_t is a fixed effect for the census year (1860), railroad_{it} is an indicator equal to 1 if county i has a railroad in year t, and ε_{it} is the error term. Through the county fixed effects, this specification controls for unobservable factors that are specific to a county across time, and through the 1860 fixed effect it controls for factors specific to all counties in that decade relative to 1850. There is an issue, however, in that railroads were anticipated some years in advance. Farmers in counties that would get railroads may have started to improve their lands prior to railroad arrival. More generally, counties that would get railroads are likely to be different from other counties. For example, they could have a more skilled population and be more likely to develop even if railroads did not arrive. Thus, the assumption that the railroad indicator variable is "as good as random" conditional on the fixed effects is questionable.

Atack and Margo address endogeneity using an instrumental variables (IV) strategy. The idea is to find a variable that predicts railroad access but

does not affect land values in any way except through railroad access. Railroad plans from the early 1820s and 1830s are used to identify likely starting and endpoints for future railroads. The starting and end points are often places of economic or military significance. Atack and Margo then use GIS to create straight lines connecting the start and end points. Counties on this line have no special characteristics except that they are located on a favorable route for building a railroad between places of significance. They estimate a much larger effect of railroads on improved land using their IV strategy compared to ordinary least squares (OLS). They argue that railways were anticipated, leading to a downward bias in the estimated treatment effect. The upshot is that Atack and Margo create a new estimate for railroads' impact on agricultural improvement. Their estimates imply that in the 1850s, there was a 13.8% increase in improved farmland due to gains in rail access, which is nearly all the increase in improved farmland during that decade. Atack and Margo's findings imply that Fishlow's social savings estimate under-states the contribution of railroads. Fishlow argued that railroads could account for only a small portion of total economic growth between 1840 and 1860.

The DID approach to measuring the effects of transport improvements has received some criticism. Recall that DID compares the change in outcomes in a treatment group of counties – identified as those getting transport access – versus a control group that did not. But one might argue that a rail "yes or no" classification does not incorporate network effects. Spatial general equilibrium trade models suggest that a railroad linking a farmer to a major market should have a greater effect than a railroad linking a farmer to a minor market. In the former case, the farmer will get cheaper or better manufactured goods because the major market has more competition and variety. Also, the farmer will get more land rent because it can more easily sell in a market with high demand and hence high agricultural prices. In the language of trade models, a railroad linking a farmer to a major market will provide greater "market access."

The baseline formulation of market access is the following: $MA_i = \sum_{j=1,\neq i} pop_j / tc_{ij}$, where MA_i is the market access of location i, pop_j is the population of any other location j, and tc_{ij} is the transport cost from i to j, and $\sum_{j=1,\neq i}$ is the sum over all other locations j except i. Notice that market access is high when there are low transport costs to more populous locations (i.e., tc_{ij} is low and pop_j is high). This implies that getting close to a railroad line will have different effects depending on its connections. If the railroad creates lower transport costs to large centers, then market access will increase more than if the railroad creates lower transport costs to small or medium centers. It also implies that railroads can have effects on locations very far away through the network structure.

There are several works in the historical literature estimating market access. One of the most well-known is Donaldson and Hornbeck (2016). It is worth examining their approach because it reassesses the effects of railroads on American agricultural development. They build their analysis from a trade model. They derive an expression for equilibrium land rental rates, which are log-linear in one endogenous economic variable, market access, and several exogenous variables like fixed land productivity. This implies that with panel data one can estimate the equation:

$$\text{Log land value}_{it} = \alpha \cdot \text{MA}_{it} + \eta_i + \delta_t + \varepsilon_{it}$$

where MA_{it} is market access of county i in year t, η_i and δ_t are fixed effects for the county and census year. Notice the similarity with Atack and Margo's model. The key difference is that Donaldson and Hornbeck use market access rather than the indicator variable for the county having any railroad. As they point out, one can include both variables in the same specification and compare their effects. Another feature worth noting is that endogeneity is less of a concern for market access variables compared to an indicator variable for having a railroad in a county. Market access is largely determined by network structure, which in many contexts is driven by decisions beyond the county under consideration.

The measurement of market access is another innovation of Donaldson and Hornbeck's study. They calculate the lowest county-to-county freight rate for all combinations of county pairs (tc_{ij} in the market access formula). The calculation relies on three components. Each is worth commenting on. The first uses transportation cost parameters that apply to a given unit length of each transportation mode (railroad, waterway, and wagon). The transport cost parameters consist of freight rates per mile. These are often reported in the literature, but one might wonder about their accuracy. Stated freight rates could be based on a small number of observations or locations. Practically speaking, freight rates could be mis-measured, leading to biases in the estimates.

The second step in calculating market access uses a transportation network GIS database that maps where freight could move along each transportation mode. The GIS data are highly detailed and offer tremendous potential. However, there is an issue in geo-rectifying railroad and waterway maps, which are the under-lying data. As Atack (2013) shows the sources are not always consistent, and researcher choices need to be made. The third step involves the computation of lowest-cost freight routes along the network for given cost parameters. The computation comes from network analysis software. Computation time can be a problem if there are large numbers of counties and many transport modes, although this problem is quickly diminishing with better computing power.

In their analysis, Donaldson and Hornbeck find large effects of railroads on market access and on agricultural land values. Their counterfactual estimates imply that in the absence of railroads, agricultural land values would have decreased by 60% in 1890. In terms of income lost, their estimates imply that Gross National Product (GNP) would be 3.22% lower in 1890. Donaldson and Hornbeck's estimated impact is larger than Fogel (1964), who argued that without railroads the loss in agricultural land values would total 2.7% of GNP in 1890. While this may sound like a small difference, consider that the loss of *all* agricultural land income could at most lower US GNP by 5.35% in 1890. Therefore, the impact of railroads is capped at 5.35% of GNP in the approach focusing on agricultural land. Future research may shed light on the urban land values or capital and labor income. Perhaps the main lesson is that the "market-access" approach offers new insights and is a useful tool for evaluating the effects of transport improvements on income.

Transport Improvements and External Effects

There is a large historical literature examining what can be termed "external" effects of transport improvements. It considers the effects of using new transportation services on changing residence from a rural to urban area, on participating in new trade, or exposure to pollution to name a few. Several interesting papers examine external effects and are worth discussing in some detail.

Greater urbanization was a feature of many advanced economies in the nineteenth and twentieth century. Urbanization was associated with increased consumption opportunities and is considered a good measure of living standards. Urbanization also had effects on productivity through agglomeration. In theory, as more people locate and work in urban areas, the productivity of the average worker increases, leading to higher incomes. There are compelling reasons to think that transport improvements increased the concentration of economic activity for most economies up to the mid-twentieth century. In economic geography models with increasing returns, agglomeration forces are stronger when transport costs change from high to intermediate (Krugman and Venables 1995). In the intermediate range, food is shipped more cheaply, consumers spend more on manufactured goods, which are more productively made in a single location, and workers tend to locate in a single location. The implication is that with better transport innovations, like the railroad, urbanization should have increased.

Several papers examine the effects of transport improvements on urbanization. One example is Hornung (2015), who studies railroads and urbanization in Prussia. In Prussia, urbanization increased significantly between 1841 and 1871. This was quite early by international standards. The other cases of early urbanization were Great Britain and the Netherlands. Prussia was also an early adopter of railroads. Most up to 1850 were built between large cities. This feature was likely due to most early railroad projects being privately owned, financed, and operated. After 1850, the Prussia government began to subsidize railroad construction and the network extended. The reasons for Prussian government involvement are complex. National defense was one key factor.

What were the effects of railroads on Prussian urbanization? Hornung studies many cities and employs various specifications to answer this question. One specification involves a panel regression of log city-population on an indicator for having railroad access. Endogeneity is addressed using straight lines which connect important cities through the railroad network. Hornung's instrument is very similar to the one used by Atack and Margo (2011). The rationale is that lines were mostly built linearly due to high construction costs. Consequently, ordinary cities located on a direct line between important cities gained access to the railroad more by chance. In contrast, cities whose location deviated from the straight line could gain access only for reasons potentially endogenous to the city's growth. Hornung's conclusion is that gaining railroad access increased a Prussian city's population by approximately 15% after 15 years.

Did railroads have a similar effect on urbanization elsewhere? Berger and Enflo (2017) study a similar issue in nineteenth century Sweden. They found that early

railroad access increased city population by 25–30% over several decades. Interestingly, the specifications and data in these two studies on Prussia and Sweden are quite similar, yet the estimated effects of railroads are different. As one reads through several papers of this kind, it becomes clear that the effects of railroads generally differ. A better understanding of heterogeneity is a subject that deserves greater attention in future research.

Another outstanding issue in this area concerns "reorganization" effects, where transport improvements increase population in one area at the expense of lowering it elsewhere. The suburbanization process in the USA during the mid-twentieth century provides a good example of reorganization. Baum-Snow (2007) shows that the aggregate population of US central cities declined by 17% between 1950 and 1990, despite population growth of 72% in metropolitan areas as a whole. Why did this happen? At this time, the US Federal Government funded a massive highway expansion, known as the inter-state highway program. The 1956 Interstate Highway Act committed the federal government to pay 90% of the cost of construction for the 41,000-mile highway system. The rest was financed by state and local governments. National defense is one of the main reasons the US Federal Government increased its contribution. Many interstate highways passed through the central business districts (CBDs) of major cities. In 1956, the CBDs were densely populated in part due to the construction of railroads a century earlier. As highways passed through the CBD, they made it easier for households to live at the edge of the city and to commute to the CBD for work. Baum-Snow estimates that one new highway passing through a central city area reduced its population by 18%. A counterfactual estimate implies that central city population would have grown by about 8% had the interstate highway system not been built. These estimates imply large effects of highways on population reorganization. Similar impacts have been found in Europe. For example, Garcia-López et al. (2015) find that each highway caused an 8–9% decline in central city population between 1960 and 2011.

The Baum-Snow analysis takes other factors into account, namely, the effects of income levels and inequality on suburbanization. The estimates suggest that greater income reduces central city population, which makes sense because households want to buy more housing space with higher income and space is cheaper in the suburbs. Baum-Snow also finds that greater income inequality reduced central city population. This would make sense if rich households wanted to isolate themselves from poor households leading to large differences in schooling and other amenities.

Building on these results, one might wonder how reorganizing the population in space affects welfare. On the one hand, if transport innovations led to more racial segregation, as in the US case, there might be negative welfare consequences. It could also lead to dispersion in production, reducing agglomeration benefits. On the other hand, moving population to suburban areas allowed more housing consumption which is clearly desired by many families. Not surprisingly, the welfare effects are quite complicated and depend on a range of pecuniary and non-pecuniary externalities.

There is a growing interest in the negative externalities associated with transport. Pollution is one of the largest concerns in the twenty-first century and many of

the emissions come from transportation vehicles (Winston 2013). The historical literature has begun to examine this issue. Britain was perhaps the most polluted economy before World War I. Significant amounts of coal were burned for home, industrial, and transportation activities. Were the amounts significant enough to cause poor health? This is a difficult question to answer because air pollutants were not measured before the twentieth century. However, innovative research by Beach and Hanlon (forthcoming) infers pollution using local industrial structure and coal usage by industry in Britain during the nineteenth century. They estimate that a one standard deviation increase in coal use raised infant mortality by 6–8%. As it turns out, transport was not the main consumer of coal in mid-nineteenth-century Britain. It was the home and manufacturing sectors. This would suggest that railroads were not a direct driver of pollution related health problems in this setting.

This conclusion may not hold elsewhere however. Tang (2017) focuses directly on the negative health effects of railroads in the late nineteenth century. An example provides an illustration. In 1886, a cholera epidemic swept through Japan and killed 108,405 people, accounting for 1 out of 9 deaths that year. Strikingly, prefectures with railroad access in 1886 had a higher incidence of mortality, 336 deaths per 100,000, compared to the 245 deaths in prefectures without rail access. Many factors could potentially explain the railroad-mortality connection, but Tang is able to isolate the effects of railroads using matching and DID regression models. The specification has similarities to studies examining the effects of railroads on land values:

$$\text{Mortality rate}_{it} = \alpha \cdot \text{Railroad}_{it} + \eta_i + \delta_t + \varepsilon_{it}$$

where the dependent variable is deaths per 100,000 people in prefecture i in year t, Railroad_{it} is an indicator for prefecture i having a rail connection in year t, and the remaining variables are fixed effects. Tang estimates that rail access accounts for a 5.5% increase in mortality between 1884 and 1893. Moreover, rail-associated mortality represents about 66% of the total increase in mortality pre and post adoption of railroads. Tang also uses official causes of death to show that 75% of rail-associated deaths were due to communicable diseases, like tuberculosis and influenza. The implications of this finding are large. One of the benefits of transport improvements is to increase urbanization and productivity. In fact, in another paper (2014), Tang demonstrates that railroads contributed to industrialization in Japan. But as Tang (2017) shows, in some circumstances greater communication can also lead to disease transmission and higher mortality. The field of cliometrics needs more research on the negative health consequences of transport, as they loom large in the twenty-first century.

Persistence and Long-Run Impacts of Transport

Most of the analysis thus far has focused on the short-run effects of transport improvements. For example, the contribution of railways to increasing land values and urbanization in the nineteenth century. But there are theoretical reasons to argue

that the impact of railroads and other transport innovations matter over the long-run. The channel is through persistence of locational choices. If a population settles in an area because of some natural or human-made advantage in the past, then the population may become "locked-in" for the future. Lock-in could persist even as the original natural or human advantage which determined the settlement becomes largely irrelevant.

Transportation advantages are one of the main factors identified in the persistence literature. Bleakley and Lin (2012) provide one of the classic studies on persistence and the role of transport in creating lock-in. They focus on portage sites in the USA during the early 1800s. Much overland transport went by lakes and rivers at this time. There were some sections of rivers that were not navigable, which meant that traders had to stop and carry their products and canoes along the river until navigation could resume. At portage sites commercial services were provided. Thus, population density tended to be higher at portage sites. Transportation options in the US clearly changed during the nineteenth and twentieth centuries making river transportation far less important. One might imagine that early portage sites would gradually lose population to locations which had other advantages. Bleakley and Lin show this did not happen. They analyze the following model

$$\text{Log}(\text{population in } 2000_i) = \alpha \cdot \text{portage site}_i + \beta \cdot x_i + \varepsilon_i$$

where $\text{Log}(\text{population in } 2000_i)$ is the log of 2000 population in a geographic unit like a county or census tract, portage site_i is an indicator for whether the unit was a portage site in the early nineteenth century, and x_i is a set of city-level controls. Note that portage is defined as a fall line or river intersection. The use of the geographic variables to identify portage makes this specification akin to a reduced form regression, in which endogeneity is not a concern. Bleakley and Lin estimate that portage units have 90 logs points higher population density in 2000. That is equivalent to 145% increase in population density. One might think that portage sites were about 145% larger in the nineteenth century when portage was most valuable. However, this is not true. Nearly half of the population advantage of portage sites emerged between 1900 and 2000, which is long after canoes became a hobby vessel.

There are other papers which document persistence in outcomes associated with early transport investment. Jedwab and Moradi (2016) study colonial Ghana, where the British built railroads linking the coast to sparsely populated mining areas and the hinterland. Using a data set on cities, they look at the effects railroads had on the distribution of population from 1891 to 2000. Jedwab and Moradi first establish the importance of railways for initial urban location. They estimate the following model:

$$\text{Urban population in } 1931_i = \alpha \cdot \text{Railroad in } 1918_i + \beta \cdot x_i + \varepsilon_i$$

where 1931 urban population in a location is regressed on indicators for having rail access in 1918 plus controls x_i. Instruments are introduced to address endogeneity for railroads and are ignored for the present discussion. Jedwab and Moradi estimate that the urban population in 1931 was significantly higher if the location

had a railroad in 1918. They argue that the decrease in internal trade costs encouraged the local cultivation of cocoa, which became a leading export in Ghana. Rural populations increased along the lines as cocoa cultivation required more labor in cocoa-producing villages. Urban populations then increased because villages used the towns as trading stations.

After the 1970s, rail transport in Ghana was far less important. Railroad tracks were not maintained, and road transport increased. Did this mean that areas close to railroads lost population? The answer is no. Jedwab and Moradi estimate a second set of regressions.

$$\text{Urban population in } 2000_i = \alpha \cdot \text{Railroad in } 1918_i + \beta \cdot x_i + \varepsilon_i$$

$$\text{Urban population in } 2000_i = \alpha \cdot \text{Railroad in } 1918_i + \eta \cdot \text{Urban population in } 1931_i + \beta \cdot x_i + \varepsilon_i$$

The first regression identifies whether railroads in 1918 affected urban population in 2000. This would represent a persistent effect because railroads had lost their original significance in transport. The second equation examines whether railroads' effect runs through the urban population in 1931, a lagged dependent variable. Jedwab and Moradi show that railroads in 1918 had a strong effect on 2000 population, but the coefficient on the rail indicator is small once the urban population in 1931 is included in the specification. In other words, the spatial equilibrium became stable after railroads were built. Subsequent transportation technologies (i.e., trucks) did not relocate population from near railroads. The Ghanan case is quite striking because the end of colonialism represented a large economic and political change, but it was not sufficient to wipe out the persistent effects of railroads.

How are these persistence patterns understood? Bleakley and Lin (2012) propose a simple model where an individual's utility is a function of the population density at their location. Two factors affect the utility according to density. One is the strength of congestion. Congestion increases with density and lowers the utility of the individual. The second is the strength of agglomeration. Agglomeration increases with density and raises the utility through greater productivity. Natural advantages, like portage sites, are another factor in the model. They increase the utility of locating in one place versus another. Thus, natural advantages can be crucial at some point in time, leading to a density in one location. If the natural advantage disappears, say because of general technological progress, then an individual might leave a location because it is congested. However, when the agglomeration forces are strong, it may be optimal to stay. Is the strength of agglomeration sufficiently powerful to lock-in locations? While compelling, there are cases like post-war Japan where cities were wiped out by bombs and yet they came back to their original size (Davis and Weinstein 2002). If agglomeration was strong, then these cities should not have remerged. No doubt, the issue of lock-in is far from settled.

Institutions and Transport Development

Much of the existing literature on transport analyzes its effects on income and development. But there is a broader question one could ask. What makes some societies effective in delivering transport services? This is closely related to the broader question of what makes some societies rich and others poor. One popular view is that institutions are a fundamental cause of economic development. Institutions are the humanly devised constraints that structure political, economic, and social interaction. In their formal sense, institutions are constitutions and legal systems, but less formally they include norms and beliefs. As originally emphasized by North (1991), institutions matter because they affect transaction costs and hence incentives to invest and innovate. Extending this logic to transport, one might imagine that institutions affect transport infrastructure investment. Here there are large fixed costs, raising the risk of expropriations or misallocation.

In the literature, there is some evidence that transport infrastructure became more developed in countries with stronger democratic institutions and with greater state capacity. A comparison of England and France in the eighteenth century provides one illustration. In England, road infrastructure was provided through a mixture of local initiative and parliamentary oversight. Local groups submitted petitions requesting the right to form a "turnpike-trust" and improve a stretch of road in their area. Parliament usually granted these requests and named the petitioners as trustees with the authority to levy tolls and improve the road. Parliament set the maximum tolls and gave local officials powers to settle disputes between property owners and trustees. Turnpikes were established on all highways linking London with major provincial cities (Bogart 2005). France had a different approach to their highways. The crown designated some highways as primary roads (royal routes) and others as secondary roads. The primary roads received some funding from the central government in Paris and were constructed and maintained by the Ponts-et-Chaussees, an elite engineering school. Secondary roads were maintained by French communes, generally through the use of corvee labor.

How did these different paths affect the size and efficiency of road networks in England and France? The data suggest that England had more km of paved roads per capita than France. England also had faster travel speeds (Szostak 1991). One reason is that English political institutions placed significant limits on the ability of parliament or the crown to arbitrarily change tolls or to substantially reduce the rights of turnpike trusts. This high degree of regulatory commitment encouraged trustees and private groups to make investments in the road because they expected that parliament would uphold their rights (Bogart 2011). It is unlikely the French crown could make such commitments, and thus private investors would have hesitated in making such investments. A second reason is that the English parliament devised mechanisms for compensating landowners when their land was taken. Parliament empowered juries who had local knowledge. They also gave county Justices of the Peace powers to oversee juries in case they were too generous to landowners. Thus, local knowledge was combined with checks and balances. France had no such

system before the French Revolution and it struggled to overcome local resistance (Rosenthal 2009).

France further developed its road network after the French Revolution. The central government increased its funding of national routes. There was also substantial change in the funding and organization of secondary roads. A French law of 1836 gave communes the right to levy 5% surtax from the four main direct taxes in their jurisdiction. It also allowed departmental councils to impose financial contributions on communes located along roads of regional importance (Price 2017). The changes following the French Revolution provide further testimony to the importance of institutions.

The financing of railways also illustrates the importance of institutions. Railway construction often required financial support from governments. However, many relied on tax revenues collected from trade, which often depended on railways. Bignon et al. (2015) argue there was a two-way feedback between government revenues and railways, with a potential for multiple equilibria. In other words, countries could end up in a "poverty-trap" with few railways and little revenue.

Bignon et al. develop their argument for Latin America. They document that railway and fiscal development were low in some Latin American countries, like Columbia and Ecuador, and high in others, like Argentina and Uruguay. They propose a system of equations to account for the co-evolution between railways and tax revenues. A discussion of their model is useful because it speaks to dynamic effects. The first system models government revenues as a function of trade.

$$\text{Government revenues}_{it} = \alpha \cdot \text{trade}_{it} + \beta \cdot x_{it} + \eta_i + d_t + \varepsilon_{it}$$

where government revenues$_{it}$ is the log of tax revenue of country i in year t, trade$_{it}$ is the log value of trade, x_{it} is a vector of controls, and the rest are country and year fixed effects. The variable trade$_{it}$ is endogenous and Bignon et al. use railways to instrument for trade. The effect of railways on trade is plausible, but do railways affect government revenues in other ways? The authors argue they do not because no taxes were directly levied on railways. There could be indirect effects of railroads, which violates the exclusion restriction for an instrumental variable. Nevertheless, some assumption is needed to identify the effect of trade and railways are a plausible instrument.

The second equation in their model specifies the target size of the rail network measured in the natural log of rail track miles. The equation is:

$$\text{rail}^*_{it} = b \cdot \text{government revenues}_{it} + c \cdot x_{it} + \eta_i + d_t + \mu_{it}$$

where rail*$_{it}$, the target network size, is a function of government revenue, control variables specific to the country, and fixed effects. Bignon et al. argue for a dynamic equation to describe rail development because the target is never achieved in any period. The growth of the rail network from t-1 to t is the difference between the target and the reality in the previous period.

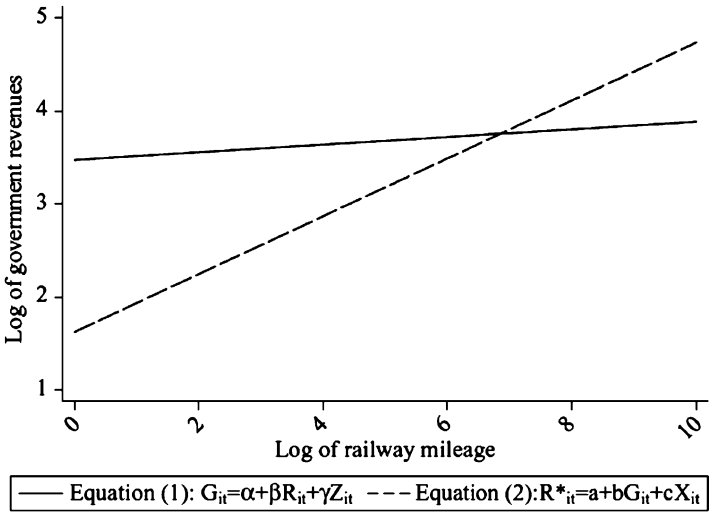


Fig. 4 Estimated relationship between tax revenues and railway miles in Latin America. (Source Bignon et al. 2015)

$$(\text{rail}_{it} - \text{rail}_{it-1}) = \delta(\text{rail}^{*}_{it-1} - \text{rail}_{it-1})$$

In the last step, they substitute the target rail network rail^{*}_{it} into the dynamic equation to get the following specification.

$$(\text{rail}_{it} - \text{rail}_{it-1}) = \delta \cdot \text{rail}_{it-1} + \delta \cdot b \cdot \text{govt.revenues}_{it-1} + \delta \cdot c \cdot x_{it-1} + \delta \cdot \eta_i + \delta \cdot d_{t-1} + \delta \cdot \mu_{it-1}$$

Their aim in this equation is to estimate the effect of government revenue and lagged rail network size on the growth of rail networks. With these two estimates, they can recover the structural parameter b. Importantly, government revenue is endogenous according to their framework. They instrument for revenues with each country's total level of diplomatic representation abroad and an index of legislative effectiveness. Here the authors feature the role of institutions explicitly. Legislative effectiveness is a measure of institutional quality, which arguably leads to more taxation. This assumption is supported by much research on state capacity (e.g., Dincecco 2015).

Bignon et al.'s estimation yields insightful results (Fig. 4). They find evidence for a two-way feedback between railways and government revenues. Their estimates are shown in the following graph for the average Latin American country. More railway mileage is estimated to increase government revenue (the solid line). Also more tax revenues are estimated to increase railway mileage (the dotted line). The equilibrium is reached where these two lines intersect. Notably for history, the equilibrium can change if there is an exogenous shock to either railways or tax revenues. For example, suppose legislative effectiveness improves due to political changes in a

country. This would lead to more revenues and more rail infrastructure (imagine the solid line shifting up). More generally, this analysis illustrates how better institutions can help countries escape a poverty trap through more transport infrastructure.

Public and Private Sector Involvement

Historically both the public and private sectors have been deeply involved in the transport sector. Various scholars have tried to explain the duality (see Newberry 2002; Milward 2005). One view sees some level of public provision as necessary because the private sector would fail to provide the efficient quantity of transport services. One reason is the natural monopoly features of transport. Very large fixed costs, like building a port or railroad, imply that total costs are minimized if there is one supplier in the market. But the one supplier can charge monopoly prices, which is inefficient. A government owner of transport would not necessarily charge the monopoly price however, because governments generally consider more than just profits when making decisions. There is an alternative view that private provision is generally preferred because of government "failure." Interest groups are a potent source of such failure because they can appeal to various layers of government to support socially undesirable infrastructure projects or to kill desirable policy reforms (Winston 2013). Ultimately it is an empirical question whether public or private provision has been beneficial or not.

Railway history provides insights on where the boundary between public and private was drawn. Interestingly, most railways built between 1830 and 1880 were owned by private companies. Afterwards there was a trend to more government ownership. In some cases, governments started to build railways and in other cases nationalizations placed large railway assets in public hands.

There are various reasons why railway nationalizations occurred in some countries and not others before 1913. For the sake of brevity and continuity with the previous section, this essay focuses on political and legal factors. In theory, limited government, embodied by strong constraints on the executive, should work against nationalizations. Constraints made it harder for the government to expropriate private property like railroad tracks and rolling stock. Common law legal systems had a similar effect in theory (see La Porta et al. 2008). In common law countries, courts have traditionally required that expropriations satisfy a "public use" and that owners receive "just compensation." The public use requirement might have made it difficult for public officials to nationalize railroads because they had to explain why government ownership satisfied a public benefit. Moreover, the just compensation clause might have prevented governments from imposing a low price on companies. In civil law countries, railroad companies could also appeal to courts, but they may not have been as effective if governments could intervene and ensure that decisions went in their favor.

Bogart (2009) provides evidence on the role of institutional factors in explaining the likelihood or extent of nationalizations across countries between 1870 and 1913. The author creates panel data on the incidence and extent of nationalizations across

more than 1200 country-year pairs. This is combined with other cross-country data, including constraints on the executive branch of government, the degree of democracy, legal origin, and a host of other variables. Cross-country estimates for 1910 reveal that the incidence of nationalizations (i.e., exceeding a minimum threshold of nationalized rail mileage) was more likely in countries with French and German civil law legal systems compared to common law legal systems and Scandinavian civil law legal systems. Nationalization was also more likely in countries with weak constraints on the executive branch or with less democracy.

The Swiss case provides some insights on why legal systems mattered for railway nationalizations (see Bogart 2009). Up to 1890, most of the railways in Switzerland were built and owned by French companies. The original concessions created a special board of arbitrators, who would settle disagreements between the French companies and the Swiss Federal government in the event of a proposed nationalization. The arbitrators were separate from the Swiss federal court system, and they presumably provided some assurance to French companies that government policy decisions would be made without political interference. By the 1890s, foreign ownership of railways became generally unpopular in Switzerland. But there were public concerns that a government takeover of railways would be quite costly to the Swiss taxpayer because French investors would demand high compensation. In 1896, one year before a major Swiss railway nationalization law was passed, the Federal Council and Assembly passed a law nullifying the authority of arbitrators and requiring that disputes over railways be settled in federal courts. The close timing of the arbitrator law and nationalization suggests they were connected. This example illustrates how government intervention and legal systems interact.

What impact did nationalizations have on the railway sector? Bogart (2010) examines the efficiency of railway operation. In other words, the money an economy spent supplying x ton miles and y passenger miles given wage and fuel prices and stocks of railway capital. The cross-country evidence suggests that the average nationalization lowered the cost efficiency of railway systems. However, there are also some economies where railway nationalizations increased cost efficiency. Colonial India is one interesting example. The initial construction and management of the Indian rail network was done by private British companies operating with a public guarantee. If net earnings (i.e., gross earnings minus working expenses) as a proportion of capital outlay yielded less than the guaranteed return of 5% in any year, the colonial government in India compensated the company with the difference. The guarantees proved costly to the colonial government. Years later it decided to nationalize all formerly private railway lines and become the majority owner. By 1910 the government had nationalized all the private railway companies that built the main trunk lines up to 1880.

Importantly, railway companies in India were nationalized on the 25th or 50th year of their original contract. This fact is useful because the timing of nationalization was determined by the date of the original concession contract and other exogenous events many decades earlier. Bogart and Chaudhary (2012) exploit this fact when they analyze the effects of Indian railway nationalizations on working

expenses (i.e., variable costs). They construct a panel dataset on inputs, outputs and costs for the primary standard and metre gauge railway systems between 1874 and 1912. Regression analysis shows that working expenses were 13% lower on average following a change to colonial government ownership. The cost declines are not driven by firms anticipating takeovers, poor quality, changes in reporting standards, or long-run trends. Rather, the evidence suggests the government reduced operational costs by cutting labor costs. These results are surprising given that nationalizations generally raise costs. Bogart and Chaudhary suggest that the colonial government in India had unique incentives to operate railways well. It had weak fiscal capacity and railways were one of the few industries it could tax effectively, provided it had majority ownership. This case points to the value of detailed studies on institutions, which are often missed in the cross-country analysis.

Conclusion

In summary, cliometrics has made major advances in the historical analysis of transportation through better measurement, economic modelling, and estimation. This essay surveys several topics and innovative research studies. There are several general conclusions. First, there were revolutionary changes in transport over the last 300 years generating high rates of productivity growth. Macro-inventions, like steam power, were important but many incremental innovations mattered too. Second, there is evidence that transport improvements contributed to greater market integration, urbanization, and aggregate income. There is also evidence that transport improvements influenced population density long after their original function became obsolete. But there is still disagreement in the literature about the relative importance of transport and there are also some studies which argue that transport improvements contributed to higher mortality. Teasing out the positive and negative effects of transport improvements on welfare is likely to remain an important area of research. Third, institutions were a fundamental factor in determining why transport services were more efficient in some economies. There is evidence that institutions influenced investment in transport networks and the degree of public and private ownership. There is also evidence that ownership mattered for transport efficiency, and in some cases, government ownership improved outcomes. Overall, cliometric research on transport offers many insights on this historically important sector.

Cross-References

- ► Market Integration
- ► Railroads
- ► Spatial Modeling
- ► Travel and Tourism

References

Atack J (2013) On the use of geographic information systems in economic history: the American transportation revolution revisited. J Econ Hist 73(2):313–338
Atack J, Margo RA (2011) The impact of access to rail transportation on agricultural improvement: the American Midwest as a test case, 1850–1860. J Transp Land Use 4(2):5–18
Bagwell P (2002) The transport revolution 1770–1985. Routledge, London
Baum-Snow N (2007) Did highways cause suburbanization? Q J Econ 122(2):775–805
Beach B, Hanlon WW (forthcoming) Coal smoke and mortality in an early industrial economy. Econ J. https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12522
Berger T, Enflo K (2017) Locomotives of local growth: the short-and long-term impact of railroads in Sweden. J Urban Econ 98:124–138
Bignon V, Esteves R, Herranz-Loncán A (2015) Big push or big grab? Railways, government activism, and export growth in Latin America, 1865–1913. Econ Hist Rev 68(4):1277–1305
Bleakley H, Lin J (2012) Portage and path dependence. Q J Econ 127(2):587–644
Bogart D (2005) Turnpike trusts and the transportation revolution in 18th century England. Explor Econ Hist 42(4):479–508
Bogart D (2009) Nationalizations and the development of transport systems: cross-country evidence from railroad networks, 1860–1912. J Econ Hist 69(1):202–237
Bogart D (2010) A global perspective on railway inefficiency and the rise of state ownership, 1880–1912. Explor Econ Hist 47(2):158–178
Bogart D (2011) Did the Glorious Revolution contribute to the transport revolution? Evidence from investment in roads and rivers. Econ Hist Rev 64(4):1073–1112
Bogart D (2014) In: Floud R, Humphries J (eds) The transport revolution in industrializing Britain: a survey in Cambridge economic history of Britain 1700 to 1870, 3rd edn. Cambridge University Press, Cambridge
Bogart D, Chaudhary L (2012) Regulation, ownership, and costs: a historical perspective from Indian railways. Am Econ J Econ Pol 4(1):28–57
Brueckner JK (2003) Airline traffic and urban economic development. Urban Stud 40(8):1455–1469
Carter SB, Gartner SS, Haines MR, Olmstead AL, Sutch R, Wright G, Cain LP (eds) (2006) Historical Statistics of the United States Millennial Edition. Cambridge University Press, New York
Davis DR, Weinstein DE (2002) Bones, bombs, and break points: the geography of economic activity. Am Econ Rev 92(5):1269–1289
De Vries J (1981) Barges and capitalism: passenger transportation in the Dutch economy, 1632–1839, vol 4. HES Publishers, Utrecht
Dincecco M (2015) The rise of effective states in Europe. J Econ Hist 75(3):901–918
Donaldson D (2018) Railroads of the Raj: estimating the impact of transportation infrastructure. Am Econ Rev 108(4–5):899–934
Donaldson D, Hornbeck R (2016) Railroads and American economic growth: a "market access" approach. Q J Econ 131(2):799–858
Ericson SJ (1996) The sound of the whistle: railroads and the state in Meiji Japan, vol 168. Harvard University Asia Center, Cambridge, MA
Federico G, Persson KG (2006) Market integration and convergence in the world wheat market, 1800–2000. In: Hatton T, O'Rourke K, Taylor A (eds) New comparative economic history, Essays in honor of Jeffrey G. Williamson. MIT Press, Cambridge, MA
Federico G, Sharp P (2013) The cost of railroad regulation: the disintegration of American agricultural markets in the interwar period. Econ Hist Rev 66(4):1017–1038
Fishlow A (1965) American railroads and the transformation of the ante-bellum economy, vol 127. Harvard University Press, Cambridge, MA
Fogel RW (1964) Railroads and American economic growth. Johns Hopkins Press, Baltimore

Garcia-López M-À, Holl A, Viladecans-Marsal E (2015) Suburbanization and highways in Spain when the Romans and the Bourbons still shape its cities. J Urban Econ 85:52–67
Gerhold D (2014) The development of stage coaching and the impact of turnpike roads, 1653–1840. Econ Hist Rev 67(3):818–845
Harley CK (1980) Transportation, the world wheat trade, and the Kuznets cycle, 1850–1913. Explor Econ Hist 17(3):218
Harley CK (1988) Ocean freight rates and productivity, 1740–1913: the primacy of mechanical invention reaffirmed. J Econ Hist 48(4):851–876
Harley CK (1998) Cotton textile prices and the industrial revolution. Econ Hist Rev 51(1):49–83
Harley CK (2008) Steers afloat: the North Atlantic meat trade, liner predominance, and freight rates, 1870–1913. J Econ Hist 68(4):1028–1058
Herranz-Loncán A (2014) Transport technology and economic expansion: the growth contribution of railways in Latin America before 1914. Revista de Historia Económica-J Iber Lat Am Econ Hist 32(1):13–45
Hornung E (2015) Railroads and growth in Prussia. J Eur Econ Assoc 13(4):699–736
Huenemann RW (1984) The dragon and the iron horse: the economics of railroads in China, 1876–1937, vol 109. Harvard University Asia Center, Cambridge MA
Jacks DS (2006) What drove 19th century commodity market integration? Explor Econ Hist 43(3):383–412
Jedwab R, Moradi A (2016) The permanent effects of transportation revolutions in poor countries: evidence from Africa. Rev Econ Stat 98(2):268–284
Kaukiainen Y (2001) Shrinking the world: improvements in the speed of information transmission, c. 1820–1870. Eur Rev Econ Hist 5(1):1–28
Kerr IJ (2007) Engines of change: the railroads that made India. Greenwood Publishing Group, Westport
Krugman P, Venables AJ (1995) Globalization and the inequality of nations. Q J Econ 110(4):857–880
La Porta R, Lopez-de-Silanes F, Shleifer A (2008) The economic consequences of legal origins. J Econ Lit 46(2):285–332
Leunig T (2006) Time is money: a re-assessment of the passenger social savings from Victorian British railways. J Econ Hist 66(3):635–673
Martí-Henneberg J (2013) European integration and national models for railway networks (1840–2010). J Transp Geogr 26:126–138
McClelland PD (1972) Social rates of return on American railroads in the nineteenth century. Econ Hist Rev 25(3):471–488
Menard R (1991) Transport costs and long-range trade, 1300–1800: was there a European 'transport revolution' in the early modern era?. In: Political economy of merchant empires, pp 228–275. Cambridge University Press, Cambridge UK.
Millward R (2005) Private and public enterprise in Europe: energy, telecommunications and transport, 1830–1990. Cambridge University Press, Cambridge, UK
Mohammed SIS, Williamson JG (2004) Freight rates and productivity gains in British tramp shipping 1869–1950. Explor Econ Hist 41(2):172–203
Newberry DM (2002) Privatization, restructuring, and regulation of network utilities, vol 2. MIT Press, Cambridge, MA
North D (1958) Ocean freight rates and economic development 1730–1913. J Econ Hist 18(4):537–555
North DC (1991) Institutions. J Econ Perspect 5(1):97–112
Pascali L (2017) The wind of change: maritime technology, trade, and economic development. Am Econ Rev 107(9):2821–2854
Price R (2017) The modernization of rural France: communications networks and agricultural market structures in nineteenth-century France, vol 13. Taylor & Francis, London
Rönnbäck K (2012) The speed of ships and shipping productivity in the age of sail. Eur Rev Econ Hist 16(4):469–489

Rosenthal J-L (2009) The fruits of revolution: property rights, litigation and French agriculture, 1700–1860. Cambridge University Press, Cambridge
Shepherd JF, Walton GM (1972) Shipping, maritime trade and the economic development of colonial North America. Cambridge University Press, Cambridge, UK
Solar PM (2013) Opening to the East: shipping between Europe and Asia, 1770–1830. J Econ Hist 73(3):625–661
Solar PM, Rönnbäck K (2015) Copper sheathing and the British slave trade. Econ Hist Rev 68(3):806–829
Summerhill WR (2005) Big social savings in a small laggard economy: railroad-led growth in Brazil. J Econ Hist 65(1):72–102
Szostak R (1991) Role of transportation in the industrial revolution: a comparison of England and France. McGill-Queen's University Press-MQUP, Montreal
Tang JP (2014) Railroad expansion and industrialization: evidence from Meiji Japan. J Econ Hist 74(3):863–886
Tang JP (2017) The engine and the reaper: industrialization and mortality in late nineteenth century Japan. J Health Econ 56:145–162
Taylor GR (2015) The transportation revolution, 1815–1860. Routledge, Abingdon
Walton GM, Rockoff H (2013) History of the American economy. Cengage Learning, New York
Winston C (2013) On the performance of the US transportation system: caution ahead. J Econ Lit 51(3):773–824

Travel and Tourism

Thomas Weiss and Brandon Dupont

Contents

Abstract

Travel and tourism, once activities that were open only to the elite, have become increasingly accessible to many more people. As a result, tourism is now an important economic activity worldwide. In this chapter, we discuss the

T. Weiss (✉)
University of Kansas, Lawrence, KS, USA
e-mail: t-weiss@ku.edu

B. Dupont
Western Washington University, Bellingham, WA, USA
e-mail: Brandon.Dupont@wwu.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_74

conceptual issues in defining and measuring travel and tourism, efforts to econometrically estimate the demand for travel, and the possible links between tourism and economic growth. We also describe the rise and decline of seaside resorts in England and the cliometric history of tourism in the United States, particularly emphasizing the historical forces that shaped tourism in Hawaii. And while economic historians have largely neglected the topic, we identify issues on which the tools of cliometrics might be brought to bear.

Keywords

Tourism · Tourism demand · Seaside tourism · Britain · Mediterranean · Hawaii · United States · Tourism-led development · Economic growth · Demand elasticity · Income elasticity

Introduction

Before the widespread use of steam engines in ocean shipping in the last half of the nineteenth century, transoceanic travel was expensive and unpredictable. An Atlantic crossing by sailing vessel in the year 1800 would have taken roughly the same amount of time as the 35-day crossing by Christopher Columbus in 1492 (Fowler 2017). The unpredictable duration and the risks associated with trans-Atlantic voyages clearly limited European-North American commercial traffic, migration, and tourism. There were no more than 2000 Americans who traveled to Europe in any year during the late eighteenth and early nineteenth centuries, but the number increased as steamships replaced sailing ships by the mid-nineteenth century. While the number of Americans who traveled abroad was still small during the nineteenth century, their increase exceeded overall population growth by a healthy margin. Not all of these travelers were tourists, but many of them were.

Tourism has been an increasingly important part of the economy in most countries. According to the World Tourism Organization (WTO 2018a, b), "International tourism is the world's largest export earner and an important factor in the balance of payments of most nations."[1] In 1950, global international tourist arrivals represented only about 1% of the world's population; today, that figure is about 16% (UNWTO 2017). It is one of the world's most important sources of employment, and it is the largest industry in many countries. According to the United Nations Conference on Trade and Development (UNCTAD), tourism today generates an estimated 10% of the world's GDP and 30% of world trade in services. Among OECD countries, tourism accounts for 4% of GDP, 6% of total employment, and 21% of service exports. And there is increasing recognition that tourism might play an important

[1]The World Tourism Organization is an umbrella organization for world tourism that was started in 1925 as the International Congress of Official Tourist Traffic Associations and became the WTO in 1975.

role in economic growth as demonstrated by a recent United Nations report on "Tourism for Transformative and Inclusive Growth" in Africa (UNCTAD 2017).

Tourism is particularly important for some smaller developing economies, especially island economies and others with natural amenities like extensive coastlines. Consider, for example, that tourism represents 62% of GDP in Seychelles, 43% in Cabo Verde, and 27% in Mauritius; it is, however, also important – and increasingly so – for large developed economics. According to the World Tourism and Travel Council (WTTC), tourism accounts for nearly 10% of GDP in the United Kingdom, 9% in Germany, and 8% in the United States.[2]

Tourism's increased importance prompted the UNWTO to lead an international effort in the mid-1990s to produce a consistent metric for travel and tourism across countries. This effort, along with a 1995 Conference on Travel and Tourism at the White House, led to the construction of the Travel and Tourism Satellite Accounts, which complies with the UNWTO standards (Platzer 2014). Numerous countries now produce similar statistics, making cross-country comparisons feasible.

Defining and Measuring Tourism

One of the key conceptual issues is whether tourists should be distinguished from those traveling for other reasons and whether it is even possible to make such a distinction in practice. Even if everyone agreed that tourists traveled for pleasure and make up a subset of all visitors, it is not easy to determine the purpose of each trip. Travel for pleasure can be thought of as a type of final demand while travel for business purposes can be viewed as a derived demand for services that are necessary inputs into the production of some other good or service. But in practical terms, both types of travelers purchase a "composite product involving transport, accommodation, catering, natural resources, entertainment, and other facilities and services, such as shops and banks, travel agents, and tour operators" (Sinclair and Stabler 1997, p. 58). So, while the distinction between final and intermediate demand is important in determining tourism's contribution to GDP, more commonly tourism scholars, governments, and other organizations adhere to a broad definition. Tourism is travel away from home lasting longer than a day for any reason, most commonly for recreation or business, but includes other purposes such as visiting family or seeking health care.

Measuring tourism is no easier than trying to define it. Nevertheless, while it has been referred to as "statistically invisible," researchers have measured what tourists consume: tourists must travel so they can be counted through airline passenger miles or international border crossings, and their spending can be measured by bookings of hotel rooms or other accommodations. Surveys of travelers have also been

[2]This includes direct and indirect contributions according to the World Travel & Tourism Council: https://www.wttc.org/. The Bureau of Economic Analysis shows a slightly smaller share for the US economy.

conducted to obtain more direct information about spending and about their main purpose for traveling.[3] Because they are difficult to separate out, and because they sometimes overlap, business travelers are often included in statistics on tourism collected by most government agencies. The Bureau of Economic Analysis (BEA), for example, specifies that a tourist is someone who "either travels outside of his or her usual environment for a period of less than 1 year or who stays overnight in a hotel or motel."[4] The "usual environment" is the place of normal residence, leisure, study, and work, but in any case is defined by the area within some specified minimum distance of home.[5]

Research on the Economic History of Tourism

Although tourism has been around for a very long time, the earliest travelers were not thought of as tourists. Harold Vogel (2016, p. 235) characterized them as "nomads, warriors, pilgrims, and the elite" whose trips were not primarily for leisure purposes. Some traveled on pilgrimages for religious reasons, while others journeyed for medical care, or were commercial travelers. There were some, however, who as long ago as Greek and Roman times (i.e., around 500 B.C.) fit the mold of what today we think of as a tourist (Casson 1974; Towner 1966).

However inchoate that ancient travel might have been, tourism in Europe began in earnest in the seventeenth century. Initially this meant for the most part, British tourists trekking to the continent, especially to France and Italy, on the Grand Tour in the seventeenth and eighteenth centuries. While we do not know precisely how many people went on the Grand Tour, it was large enough to have been seen as something that upper class young men were expected to do by the mid-eighteenth century (Burk 2005).[6] Beyond its role in the rise of mass tourism, the Grand Tour may have also played an important role as a mechanism by which ideas spread during the European Enlightenment.

The numbers of travelers on the Grand Tour were small at the start. Scholars have estimated around 20,000 such travelers in the eighteenth century, perhaps as high as

[3]The purpose of traveling has been measured by the *Survey of International Air Travelers*, administered by the U.S. Department of Commerce, which asks travelers about the main purpose of their trip. This survey has been administered to a random sample of international air travelers to or from the US (excluding Canadians) on a monthly basis since 1983.

[4]Okubo and Planting (1998, p. 11). The BEA prefers the term visitor to tourist because it is more descriptive of the travel activities included in the satellite accounts (Ibid, p. 8).

[5]For the United States, the BEA defines this area as 50–100 miles of home, whereas the Consumer Expenditures Survey of the Bureau of Labor Statistics uses 75 miles, while the American Travel Survey of the Bureau of Transportation Statistics uses 100 miles, and private surveys use 50–100 miles (Ibid, p. 11). In any case, the distance would be determined by the existing technology and costs of transportation, so would surely vary over time.

[6]Early travel guides were developed as the Grand Tour became more widespread (Burk 2005).

40,000 near the end of the century, and 50,000 in the 1830s (Towner 1985). By that time, a parallel development in domestic travel was proceeding apace. Spas and seaside resorts, which had been the province of the elite in the eighteenth century, were becoming more accessible to the middle classes, increasingly so once the first travel agencies were established in the mid-nineteenth century. By the last quarter of the nineteenth century, tourism was well-established in Europe especially among the English.

The number of tourists arriving in Europe swelled considerably with the advent of steam powered ocean liners in the latter half of the century, as well as continual increases in the speed and quality of the ships, and increases in carrying capacity, driven largely by immigrant traffic. Arrivals from the US alone tripled from 40,000 around 1870 to 120,000 in 1900, doubling again before World War I. After that war, travel resumed where it had left off, with arrivals from the US increasing by about 50% before a substantial downturn occurred during the Great Depression, and a more serious interruption during World War II and the years leading up to it. Travel recovered quickly after the war; by 1950, the number of tourist arrivals in Europe ran around 25 million, and had increased 20-fold by 1990 or thereabouts (Shaw and Williams 1997).

Despite that long history, the study of tourism by historians is a more recent development. According to Towner and Wall (1991, p. 73), before 1990 there was "very little mainstream historical research" on tourism. And work by economic historians was even scarcer. Since then, there has been a vast outpouring of work on tourism originating from a variety of disciplinary perspectives. While most of this work has been done by sociologists, geographers, and those in cultural studies, there is also a large body of scholarship on the economics of tourism and the history of tourism. Many of the economics articles are practical and business oriented, with titles such as "Preventing tourists from canceling in times of crises" (Annals of Tourism Research, 2016), but there are quite a few studies that focus on change and growth. Most are not economic history per se, as they cover recent and relatively short time periods, and are typically policy-oriented, but they are empirical investigations on topics similar to those studied in cliometrics. Although they do not all use the analytical tools of cliometricians, they do make use of pertinent economic statistics.

It would be impossible to summarize all the work that might be related to cliometric type investigations, so we have narrowed down our survey in two ways. We first look at work on two topics – the demand for tourism and tourism's impact on economic growth – regardless of country. We then review the economic history of tourism in two specific cases. In one instance, we focus on the history of seaside tourism as it developed in Britain and Spain. The combined story in essence covers the history of seaside resorts from their origin as exclusive spaces for the wealthy to their present day status as mass tourism. The other story is that of Hawaii, one of the few states in the US in which tourism is of relatively greater importance than it is for the nation as a whole. In both of these histories, scholars have used economic statistics and reasoning, even though they did not make use of counterfactual analysis or more sophisticated economic analysis.

An Overview of the Economic History of Tourism

The conventional wisdom about the economic history of tourism is that tourism has an income elasticity greater than one, so increases in income that took place throughout the industrializing world in the nineteenth and twentieth centuries, and even earlier in England, stimulated the demand for travel. Of course, tourism depends on other things as well, in particular factors that affect the price and supply of tourist services, perhaps especially changes in transportation. Improvements in transportation that reduced the price and speed of travel, and improved the regularity and quality of travel as well, led to increased numbers of tourists which eventually made it possible for firms, including hotels, to take advantage of economies of scale.

One of the key supply-side developments was the packaged tour industry, which is usually thought to have begun with Thomas Cook's temperance tour from Leicester to Loughborough in 1841. He and subsequent packaged tour companies took advantage of the benefits of scale economies and increased transport capacity to offer lower-cost travel opportunities, thereby opening tourism to an even larger segment of society. Competition among tour operators served to further reduce prices and fuel the rise in the number of travelers. The second consequence was the emergence of a "tourist industry." Hotels came into greater use, replacing the private homes that had been the mainstay of the earlier and much wealthier tourists; and restaurants increased in number to serve these hotel patrons. With increased speed of travel by rail, hotels would locate at major railroad terminuses, replacing the more numerous smaller inns scattered along the way to more distant destinations that comprised the Grand Tour. As a result, new tourist destinations beyond the handful that had attracted travelers to the Grand Tour emerged as well. The third consequence was the emergence of "mass tourism," which revolutionized the nature of travel from the small, select, individually arranged trips of the seventeenth, eighteenth, and early nineteenth centuries.

However plausible this summary might sound, there has not been a great deal of evidence put forth to confirm the precise timing of the rise of mass tourism, the mechanism behind it, or the relative importance of the components of that mechanism. As described below, these sorts of questions have been addressed to some extent in quantitative fashion, but not as fully as cliometricians might like. Moreover, the analyses typically focus on the details of the growth process as it played out at the individual country, regional, or local level, not for an entire continent.

The Demand for Tourism

Tourists make a wide range of economic choices, including how much to budget for a trip, the amount of time to spend on it, and how to get there, economic decisions that are to some extent amenable to formal empirical analysis. Following developments in econometrics, scholars have used a variety of methods to estimate the demand for tourism, primarily for the purposes of generating income and price elasticities or to forecast demand for tourist destinations. While this work is

empirical in nature, it is not generally cliometric, primarily because history is not a prominent component of the analysis, and institutional factors impacting tourism are often neglected. Moreover, the research on demand estimation is often policy oriented, focused primarily on very recent trends.

Methodological Approaches

The earliest papers in which scholars estimated demand functions for travel date to the early 1960s, although there was relatively little empirical evidence on the extent to which travel responded to changes in income or other factors until fairly recently. As late as 1997, Shaw and Williams (p. 19), writing about tourism in Western Europe, stated that "there is little quantitative evidence on the elasticity of demand."

There are, however, a number of studies that use various econometric methods to estimate the demand for tourism, typically either to generate estimates of various elasticities or to produce forecasts of demand. While there are a wide range of variables included in these models, most of them regress tourism arrivals or expenditures on some combination of: the average cost of travel between countries; exchange rates; per capita income in the origin country; some measure of relative prices between countries; and in some cases dummy variables to control for various country-specific factors. Eighty-four of the 100 published studies reviewed by Lim (1997) included income as an explanatory variable, 73 used some measure of relative prices, 55 used transportation costs between the countries, and 25 used an exchange rate variable.[7] The particular specification used depends on the nature of the underlying data and the objective of the study, but the traditional tourism models transform all variables into natural logs, and then use OLS to estimate the elasticity values.[8]

Almost Ideal Demand System (AIDS) models, originally developed by Deaton and Muelbauer (1980), have also been used in some empirical work on tourism (for applications in the tourism literature, see Fujii et al. 1985; O'Hagan and Harrison 1984; Syriopoulos and Sinclair 1993).[9] The AIDS models are built on the theory of consumer choice and allow us to consider a multilevel budgeting process whereby travelers allocate spending to various travel-related "goods" (these could be country-destinations or specific travel-related expenditures within a given destination).

[7]Note that exchange rates are part of relative price measures when used, as described by Lim (1997). Also, as Lim (1999) and Song et al. (2010) point out, transportation costs and income are typically highly correlated, so many studies include only income.

[8]Lim (1997) found that 56 of the 100 papers she reviewed used only log-linear models while 14 others used both log-linear and linear models.

[9]One could use other models like the Linear Expenditure System (LES) developed by Stone (1953), the Rotterdam Model (Thiel 1965), or the translog model (Christensen et al. 1975), but most scholars in the tourism literature have used the AIDS models, which do not impose a priori restrictions on the elasticities as do the LES models. See Deaton and Muelbauer (1980) for a discussion of how the AIDS models compare to the Rotterdam model.

The AIDS model can be expressed as follows, where the consumer's budget share w_i for each good i = 1, 2, …, n is expressed as a function of price (p) and total expenditure (x):

$$w_i = \alpha_i + \sum_j \gamma_{ij} \log p_j + \beta_i \log\left(\frac{x}{P}\right)$$

Deaton and Muelbauer (1980) developed a nonlinear price index that is difficult to estimate in practice, so most empirical estimates use a simpler linear price index developed originally by Stone (1953).

The AIDS models are appealing in part because they are derived explicitly from, and thus conform to, the axioms of consumer choice, but also because they yield expenditure elasticities that are of interest in understanding income generated by various tourism-related activities.

A variety of time series specifications have also been used in the tourism literature (especially for forecasting demand), starting with the autoregressive integrated moving average (ARIMA) developed by Box and Jenkins (1970). Variations on the ARIMA models have been used in more recent research as well – Song and Li's (2008) review of 121 articles published between 2000 and 2007 shows that over two-thirds of them used some version of an ARIMA model. As Song and Li (2008) point out, inconsistencies in the forecast performance of ARIMA models has pushed scholars to explore a variety of alternative time series approaches including generalized autoregressive conditional heteroskedastic (GARCH) models (see Chan et al. 2005). Other approaches have used error correction (Song et al. 2000; Lim and McAleer 2001), vector autoregressive models (Wong et al. 2006), autoregressive distributed lag (ARDL) models, or time varying parameter models (Song and Wong 2003). The time varying parameter models are potentially quite important, given the evidence that elasticity estimates vary, perhaps substantially, over time (see Peng et al. 2015).

Gravity models, originally developed by Jan Tinbergen in the 1960s international trade literature, have also recently been used in modeling tourism demand (Keum 2010; Morley et al. 2014; Culiuc 2014).[10] As applied to tourism, the standard gravity model suggests that tourism flows depend on: the size of the origin and destination countries, typically measured with GDP per capita; the geographical distance between the origin and destination as a proxy for transportation costs; and in many cases a variety of country-specific policy measures or dummy variables.

In practice, these models are typically estimated using OLS in the following form:

$$\ln T_{ij} = \alpha_0 + \alpha_1 \ln M_i + \alpha_2 \ln M_j + \alpha_3 \ln D_{ij} + \varepsilon_{ij}$$

where $\alpha_1, \alpha_2 > 0$ and $\alpha_3 < 0$.

[10] See also Kimura and Lee (2006) who found that trade in services is better predicted by the gravity model than is trade in goods.

M_i and M_j represent "economic mass," typically measured by GDP per capita in countries i and j, respectively, and D_{ij} is a measure of distance between countries i and j. The assumption is that travel between two countries depends positively on their GDP and is inversely related to the distance between them (or some other measure of trade costs including common language or currency, for example).

Panel data methods have also been used in more recent scholarship as panel datasets have become more widely available. For recent applications in the tourism literature, see Yazdi and Khanalizadeh (2016) and Culiuc (2014).

Elasticity Estimates

Income elasticity estimates vary considerably across studies, and obviously depend on the model specification, the variables included, and the countries/regions in question, but nearly all of them indicate that overseas travel is a luxury good with income elasticity values higher than one.[11] Based on a meta-analysis of 80 studies of international tourism demand, Crouch (1995) concluded that most income elasticity estimates were between 1 and 2, similar to other studies by Song et al. (2010) and Peng et al. (2015).

There is evidence of considerable variation in income elasticity across origin-destination pairs. Crouch (1995) found income elasticity values ranging from 0.3 for Latin America to 4.4 for the developed Asian economies (apparently mostly driven by income elasticity in Japan, as that is the focus of most of the Asian studies). Peng et al. (2015) find that income elasticity is particularly high among Europeans and Americans destined for Africa (the average income elasticity values are 3.25 and 5.84, respectively), but even for the more common European visits by Americans, the average income elasticity of 1.81 is fairly high.

There is also some evidence that income elasticity has fallen over time. Song et al. (2000) showed that the income elasticity of American and UK tourists destined for South Korea fell from 8.0 and 5.0, respectively, in the 1970s to 2.5 and 2.0, respectively, in the 1990s. Other studies examining different origin-destination pairs similarly show that income elasticity has declined over time. Gunter and Smeral (2016) explore the reasons for falling income elasticities between 1977 and 2013, arguing that structural macroeconomic changes have led to increased precautionary savings and liquidity constraints that limited spending on luxuries (and may have shifted traveler preferences away from international and toward domestic trips).

The length of the trip is another source of variation in income elasticity. Since international tourist travel is a luxury good, particularly for longer trips abroad, tourists may be more likely to cut back during periods of economic stress by taking

[11]Whether one uses per capita income, total income, or some other measure will depend on the purpose of the study; these differences can account for some of the variation in the income elasticity measures. Peng et al. (2015, Table 4) also found that there is some variation in income and price elasticity estimates depending on the general type of model used and the frequency of the underlying data.

shorter trips. But it is possible that long-haul tourists are simply a fundamentally different category of traveler altogether as Anastasopoulos (1984) suggests. If long-haul tourists are high-income, it is conceivable that long-haul income elasticities are lower than that for short-haul travelers. Crouch (1994) proposes that since long-haul travel is no longer accessible only to the wealthy, the income elasticity of demand for long-haul travel likely exceeds that for short-haul trips. His meta-analysis of 80 empirical studies of international tourism suggests that income elasticity does in fact depend in part on the length of the trip, and that while both long- and short-haul trips have income elasticity values higher than 1, travelers on long-haul trips are more sensitive to changes in income. In the studies he reviewed, the average income elasticity for long-haul trips was 2.99 compared to 1.98 for short-haul trips.

As with income elasticities, price elasticity estimates vary across studies because of differences in methods of estimating them, the time periods under consideration, and different origin-destination pairs. Nevertheless, there are some general findings that emerge from this literature. Peng et al. (2015) analyze 195 different studies published between 1961 and 2011 and find an average price elasticity of −1.28. As we would expect, there is a considerable variation in these values across regions. According to Crouch (1995), price elasticity values range from −0.4 in Northern Europe to −0.8 in Latin America.[12] Gatt and Falzon (2014) found that the own-price elasticities derived from an AIDS model for British tourists to Mediterranean countries varied from −0.76 for Turkey to −3.44 for Cyprus. Peng et al. (2015) found that American and European tourists are most price sensitive with respect to Africa as a destination (the average price elasticity is −3.08 for Americans and − 2.19 for Europeans) and least price sensitive with respect to Oceania (the average price elasticity is −0.675 for Americans and − 0.449 for Europeans).

As was the case with income elasticity, Crouch (1994) showed that price elasticity also depends partly on the length of the trip: price elasticity is somewhat higher for short-haul trips (average of −0.6) as compared to long-haul trips (average of −0.48). These differences may partly reflect that tourists are less aware of foreign prices in more distant destinations or it may simply be that tourists see more distant destinations as more irresistibly appealing, and thus less sensitive to changes in price for those destinations.

Most of the empirical studies focus on income and/or price elasticities, but there is some evidence on traveler responsiveness to other factors as well. There is less variation in exchange rate elasticity across regions and studies, with the average elasticity reported by Crouch (1994) of about −1.0 (the exchange rate is expressed as the ratio of units of origin country currency per unit of destination country currency). There is evidence that travelers going to Northern Europe are significantly less sensitive to exchange rates as compared to those going to Southern Europe and the Mediterranean. North American tourists and those from Oceania are the most sensitive to transportation costs according to Crouch (1994), with average elasticities

[12]The regions referred to here are for the region of origin. Crouch (1995, p. 112) also reports results for region of destination.

of −1.52 and − 1.46, respectively. Tourists from Northern Europe and the developed Asian economies appear to be least sensitive to transportation costs, presumably because many of the international trips taken by Northern Europeans are relatively short distance (although the reasons for the low transportation cost elasticities for Japanese and other developed Asian economy tourists are less clear). Costa (1997) found that recreational expenditure elasticities for Americans, which were about 2.0 at the start of the twentieth century, were cut in half by the end of the century. She attributes this change to the falling price of recreation, investment in public recreational goods, and rising income levels.

The Impact of Tourism on Economic Growth

Whether tourism has stimulated economic growth, either nationally or locally, is the sort of question that should interest economic historians. It is certainly of interest to many developing countries and some developed ones as well. Has tourism been a source of economic growth? Have revenues from tourism exports stimulated economic growth in the same way that exports of manufactured goods or some agricultural commodities are thought to have done? Has tourism played as important a role in shaping the economy of some countries or regions as say cotton exports did for the United States before 1840 as described in Douglass North's (1961) analysis? Scholars of tourism have studied some of these issues, although the research is typically policy-oriented and focused on recent time periods in developing countries. The analyses tend to ignore how the tourist industries came to their current state, and do not consider what other paths toward faster growth might have been taken.

There have been more than 100 studies that have addressed the question of whether tourism leads to economic growth, and Brida et al. (2016) have surveyed 95 of them. A few of the studies, mostly those that examined groups of countries, used cross-sectional or panel data analysis, but most were time series analysis using primarily VAR and VECM methods to test for Granger causality. The consensus of these studies is that in the long run, the tourism-led growth hypothesis was validated in many different countries and regions. Only in 19 cases was economic growth found to have led tourism, or there was no Granger causality at all. Tourism-led growth was found in 7 out of 10 cases in Africa and the Middle East; 13 out of 13 cases in the Americas (mostly the Caribbean, South and Central American countries)[13]; 25 out of 32 Asian and Pacific destinations; and in 18 out of 22 European destinations. In the 16 studies of groups of countries, the tourism led hypothesis fared very well. Only one study, which covered 140 developing countries, found no support for the hypothesis. In 11 of the other cases, tourism was found to lead economic growth in all countries in the group, and in four studies, tourism led economic growth in at least some of the countries in the group.

[13]The single US case (Tang and Jang 2009) gave only a short run result, which showed that economic growth led tourism development.

These Granger causality tests make a prima facie case that tourism, especially international tourism, might have fostered economic growth. Strictly speaking, the results demonstrate that tourism is more likely to have caused economic growth than that economic growth caused the growth of tourism exports; or that growth of tourism receipts preceded the growth of income, rather than the growth of a nation's income preceding the arrival of tourism receipts. But it is not clear what to make of all this.

Although these studies present the results as bearing on long-term economic growth, the length of the time periods covered are almost all shorter than 50 years, the length of time that Simon Kuznets thought it would take to demonstrate that a country had achieved modern economic growth. The mode is 27 years, and there are only 11 unduplicated cases covering 40 or more years. Moreover, many of these studies have looked at countries where tourism is one of the more important sectors of the national economy.[14] As Brida et al. point out (2016, p. 424) ". . . there appears a sample bias since the countries, for which the TLGH is tested, are destinations characterised by a high tourism propensity and thus the weight of the tourism sector in such economies is sufficiently prominent to exert a positive impact on economic growth."

More substantively, few studies go beyond testing for the likely direction of causation. An oft-cited study, by Balaguer and Cantavella-Jordà (2002), used quarterly data for Spain for the period 1975–1997 in a regression of real gross domestic product on real foreign exchange earnings from tourism and the real effective exchange rate (a proxy for external competitiveness). They found that tourism had a strong positive impact on the growth of the economy, via a multiplier effect of foreign exchange earnings. They estimate that "a 5% sustained growth rate in foreign exchange earnings from tourism would imply an estimated increase of almost 1.5% in domestic real income in the long run" (p. 881). Policymakers would likely salivate over such prospects, but some skepticism is in order. As the authors point out, in the period before their study, tourism never represented more than 5% of overall Spanish income, which makes it unlikely that it could have such a large impact. Moreover, they also found that external competition was essential. It is the combination of the three variables – real GDP, tourism earnings, and the real exchange rate – that yields a reliable long-run relationship. If the real exchange rate is dropped, they find no cointegrating vector between economic growth and tourism, which seems to suggest that international competition may have been a more important force for economic growth.

In a study of tourism in Nicaragua, Vanegas and Croes (2007) found that tourism exports had a larger impact on economic growth from 1980 to 2005 than did exports of coffee (the main agricultural crop) and manufactured products: a 5% increase in tourism exports produces about a 3% increase in economic growth compared to a 2.5% impact on growth from coffee exports and a 1.6% impact from manufactured

[14]Oh (2005) claims that when tourism is a small share of GDP, we are more likely to see the "causality" run from GDP to tourism.

goods exports. On the other hand, Sinclair (1998) found that tourism had a rather limited multiplier effect, at least for developing countries; income multipliers ranged from 0.78 for the Bahamas to 1.59 in Sri Lanka, and only 3 of the 8 countries in her study had multipliers that exceed 1.0. This appears to be largely because tourism dollars are "channeled to developing countries via tour operators, located and owned in industrialised countries. . ." (Sinclair 1998, p. 29).[15] For more developed economies, income multiplier estimates are generally higher; for instance, Sinclair points to an estimate of 1.7 for the UK (based on Richards 1972).

A more fundamental question is how tourism exports stimulate economic growth? What is the transmission mechanism? Although receipts from international tourists are invisible exports, they can have the same effects as visible exports of agricultural or manufactured goods, by generating foreign exchange that allows a country to import more, especially capital goods, than would otherwise be the case (see Balaguer and Cantavella-Jordà 2002).[16] Local firms that supply tourist services might become more efficient because of competition with other tourist destinations. And the tourist trade might provide the opportunity for local firms to take advantage of economies of scale (see McKinnon 1964, Krueger 1980, and Bhagwati 1988). Other effects, perhaps less widely accepted, are that tourism might stimulate investments in new infrastructure, such as airports, roads, and hotels, and in human capital, increasing knowledge and professional services (Sakai 2009; Blake et al. 2006). And as Marrocu and Paci (2011) pointed out, tourism can enhance regional total factor productivity because it is one way for information to flow across borders.

A study by Capo et al. (2007) used a Solow growth model to estimate the sources of growth in the Balearic and Canary Islands, both of which are major tourist destinations in Spain. Both destinations grew more rapidly than the country as a whole between 1965 and 2000 – the Canaries grew a full percentage point faster while the Balearics grew faster by about ¾ of a percentage point – suggesting tourism contributed to growth for the entire country. But their results show that the chief source of output growth over the 35 years was the growth of capital, which accounted for 41% of the growth in the Balearics and 57% in the Canaries, and productivity (i.e., the Solow residual), accounted for 25% in the Balearics and 18% in the Canaries. More notably, the contribution of productivity advance was greater at the start of the period than at the end; in the last 5 years, productivity advance was negative in the Balearics and zero in the Canaries. They attribute this to the islands' failure to invest in long-term determinants of growth.[17] The tourism industry on these islands did not, and perhaps did not need to, invest in productivity-enhancing

[15]She also points out that tourism often carries substantial costs on infrastructure or other tourism-promoting activities.

[16]The Marshall Plan recognized this possibility as a way to help war-torn countries recover after World War II.

[17]This result is reminiscent of North's (1961) story about US economic growth. The export proceeds entered the US income stream in the South, but the proceeds were used to demand food from the West and manufactured goods and services from the North, thereby stimulating economic growth and development in the rest of the country, but not in the exporting region.

innovation or human capital, with the consequence that their growth slowed down over time relative to those regions that did invest in these sorts of things.

Another question is what impact did tourism's growth have on other industries? Did the growth of tourism crowd out investment and growth in other industries? This has been addressed in computable general equilibrium models that are increasingly being used in tourism studies. Their most frequent use is in economic impact studies, such as the effect of a terrorist attack or a special event such as the Olympic games (see Dwyer 2015), but they have also been used to assess the net impact of a change in tourist demand on a nation's economy. Examples include Australia, Fiji, and Singapore (Adams and Parmenter 1995; Narayan 2004; and Meng 2014). The general result is that the gains from the growth of the tourist-related industries were offset by losses in other industries, such as nontourist exports, because of an appreciation of the country's currency due to the increased tourist demand. As Dwyer concludes, "Unless there is significant excess capacity in tourism-related industries, the primary effect of an economy-wide expansion in inbound tourism is to alter the industrial structure of the economy rather than to generate a large increase in aggregate economic activity" (Dwyer 2015, p. 115). This would suggest that tourism may not be as much of a key to economic development as has been suggested by the Granger causality or input-output models.

And tourism can be on the other side of this process. Forsyth et al. (2014) used a Computable General Equilibrium assessment (CGE) to examine the impact of a boom in mining exports from 2004 to 2011 on the Australian tourist industry. They found that the upsurge in mining exports resulted in Australia suffering from "Dutch disease" manifested by the slowing of growth in the tourist sector. The mining boom resulted in an appreciation of the Australian dollar, which led to a slowing down of tourist arrivals, as well as an increase in outbound tourism, both of which contributed to slowing growth of the domestic tourism sector, an effect which was not completely offset by the income effect of the mining boom.

The Economic History of Seaside Resorts

Tourism may have begun with the Grand Tour, or even earlier with pilgrimages or the travels of Greeks and Romans, but it was not until the eighteenth century that the sort of travel we now consider tourism became well established, even if not yet commonplace. It took the guise of taking the waters at spas and going to the seashore. Travel to spas took place in many locations in the eighteenth century – even in colonial America – but going to the seaside was especially important in Britain. It was the preeminent example of this type of tourism, and a harbinger of things to come at seasides elsewhere. Despite the emergence of competing seaside resorts – first in France and the low countries in the late eighteenth century; then elsewhere in France, northern Germany, Scandinavia, and the Atlantic coast of Spain in the early nineteenth century; and later in the nineteenth century in a few locations in Italy – the British resorts as a whole thrived for roughly two centuries, and indeed are still functioning. Today, however, the seaside resorts in the Mediterranean are

more popular, and not only because tourists have come to prefer warmer climates. That success stems in part from transportation improvements, from the nature or types of seaside resorts that arose in England centuries ago, and from the response of British resorts to the competition of large-scale tourism that emerged in the Mediterranean.

The history of that transformation from the traditional British seaside resorts to the mass tourist resorts in Spain and elsewhere provides a fertile area for cliometric research on matters that are common to the development of other industries. To date, scholars have evaluated the appropriate measures of tourism, assessed the available data, and compiled pertinent statistics. While these are important advances in our understanding of this major shift in the resort industry, there is still much that cliometricians could do. Quantitatively assessing the relative importance of the various factors that shaped the growth and distribution of the seaside resort industry is one such example. Further examination of how the historical and institutional characteristics of various seaside resorts influenced the response to changes in income, transport costs, and hotel prices seems called for as well.

There are elements of path dependence in the development of this particular tourism industry that may not be fully explained by data on prices and tourism flows. Certain decisions made by economic agents and local governments in these various resort areas appear to have had a more profound impact than might be expected. And this is surely not the only tourist industry to have been shaped by path dependence. A better understanding of the broader context in which this particular industry developed will facilitate analysis of the history of other tourist industries. And, at the same time, this sort of analysis is a useful complement to the more technical, statistical approaches described earlier in the chapter.

The earliest British seaside resorts, those in the South of England at Brighton, Margate and Weymouth, as well as Whitby and Scarborough in the North, began in the eighteenth century by attracting an aristocratic clientele. In the early nineteenth century, upper middle class resorts, such as Torquay and Bournemouth, were established, and in the late nineteenth century there was more rapid growth in the number of both resorts and tourists, fed by the extension of the railway system and designed to serve the working class and day trippers. Blackpool and Hastings may be the most well-known, but by the early twentieth century, resorts had sprung up on almost every coastal region in England and Wales. The industry continued to grow rapidly through the early 1970s, save for interruptions by World War I, the Depression, and World War II, although not all seaside resorts fared well (Walton 1997, p. 24).

The seaside resort still provided the most popular form of domestic holiday up through 1974, which marked the zenith of Britons taking domestic holidays (40.5 million holidays of 4 nights or more) of all types, not only seaside holidays (Demetriadi 1997). Only sometime in the 1970s did international competition from seaside resorts outside the UK and competition from other types of tourism, both domestically and abroad, bring about a decline in British seaside holidays. From there on, the shift of the seaside resort industry from the UK to other locations, most notably the Mediterranean, may be as remarkable an economic

change as the relocation of the textile industry from New England to the Southern United States in the late nineteenth century.

UK tourists were drawn to other European destinations in large numbers beginning in the 1960s, first to the Mediterranean seaside, especially Spain, and then to Alpine winter resorts as both became packaged and produced for mass consumption.[18] In the 1980s, some traveled farther to Florida, Australia, Thailand, and Gambia, but Europe remained the most popular international destination for UK residents. This international travel posed problems for British seaside resorts because foreign tourists to the UK, while increasing in number (four million in 1967 vs. 16.6 million in 1992), came for traditional landscape and heritage tourism, and especially for urban tourism in London where 54% of all foreign tourists went in 1992 (Williams and Shaw 1997).

As plausible a story the above seems to be, it has not been fully documented by statistics and analysis. The reason for this is the same shortcoming faced by historical research on tourism everywhere – an absence of appropriate data, even for the most basic variables, such as the number of tourists.

The data problems and efforts to deal with them are described in *The Rise and Fall of British Coastal Resorts* (Shaw and Williams 1997). Because of the lack of data as well as dissatisfaction with some of the existing data, tourism in the British resorts is measured differently over time in this volume. It is variously measured by the resident population of the seaside towns for the period 1900–1950, by the number of holidays of 4 or more nights for the period 1950–1974, and the decline after 1974 is charted by the number of holiday nights spent in seaside resorts. On the Mediterranean side of the story, the editors of *Europe at the Seaside* (Segreto et al. 2009) argue that there had been no study of tourism for the Mediterranean region as a whole primarily because of a lack of statistics. "The aim of this book is to fill this gap," which they do with some success (p. 2).

John Walton (1997), writing about British seaside tourism in the period 1900–1950, argues that there is no plausible evidence on visitor numbers, and obtaining them would require much research into local sources that would not likely be very reliable. Walton is skeptical of the estimates produced by individual resorts in the 1930s and uses census figures on the permanent population of the resort towns, rather than numbers of visitors, even though he recognizes the problems with the census data.

Data on resident population show that the British seaside resort industry was well established by the beginning of the twentieth century. Between the census dates 1881 and 1911, the off-season (permanent) population of the seaside resorts increased by around 60% and accounted for 4.5% of the population of England and Wales, and the share rose to 5% in 1931 and 5.7% in 1951 (Shaw and Williams 1997; see Walton 1983 for further detail). Walton's analysis focuses on population change in a sample of 116 of the most prominent resorts based on reputation and advertising visibility . . . "rather than on more demanding 'scientific' criteria" – and excludes those that

[18]The internationalization of British tourism began in the previous decade with the first air package holiday to Corsica (Williams and Shaw 1997, p. 3).

posed a serious problem of boundary change. He examines their performance in two ways: the absolute increase in numbers of people and the percentage increase in the population.

Based on numbers of people, the large, established resorts, such as Brighton, Bournemouth, and Blackpool, experienced large increases in numbers and retained their preeminence. The notable exceptions were Great Yarmouth, Dover, and Hastings, which recorded declines or only a small increase. Based on growth of population, the fastest growing resorts were those that were small at the start of the period, and were boosted by the development of holiday home sites and holiday camps. These attracted vacationers from East London, who could not have afforded holidays in the more established resorts, as well as Bohemian and artistic middle classes, and a growing number of year-round, elderly residents.

In the end, Walton recognizes that population growth cannot be the measure of success for all seaside resorts. Census data indicated that by 1951 a number of smaller places, such as Grange-over-Sands and Lyme Regis, that had been successful resorts due to their climatic advantages chose to remain small; local authorities and landowners restricted development and discouraged amusements. Others grew slowly but retained their architectural amenities and gardens, and succeeded in other dimensions; Eastbourne, for example, had the fifth highest figure for retail spending per head in 1950. And yet others like Torquay grew impressively but went down-market in the process. Such divergence among the objective function of individual resorts poses a problem for analyzing the causes of tourism growth and its consequences in any country.

Walton argues that regardless of which measure of tourism growth is used, the causes of that growth were to a large extent the same. It was important to have the right conditions for growth, which included "a welcoming attitude to the expanding popular market, as suggested by openness to plotland and holiday camp incursions" (p. 38). Resorts that sought to safeguard a high "social tone" did not grow because they could not, or did not, move down market, at the same time that their traditional, upper class visitors were lured abroad. Accessibility to transport was important to attract visitors as well as enable growth of the resident population; seaside commuting and retirement were important in the interwar years. Whatever the sources of growth, the seaside industry by 1951 had recovered from World War II.

The period from 1950 to 1974 has been labeled the golden years for seaside resorts, although the term might be better reserved for only the decade of the 1950s. That decade had full employment, there was little or no competition domestically or abroad for seaside resorts, and the 1950s was the first decade to experience the full impact of Holidays with Pay Act (Demetriadi 1997). It was still an age of railway travel, not private autos, so the existing resorts remained as accessible and preferred as they had been.[19] Domestic tourism continued its post-World War II rise up to

[19]As car ownership increased and roads improved, tourists were no longer captive to the resort near the railway terminal.

1974, part of which was an increase in second and third holidays; a phenomenon which in the longer run did not bode well for longer term stays at seaside resorts (Demetriadi 1997).

The Shift from British to Spanish Seaside Resorts

Even while the numbers of domestic holidays taken by British residents were rising in the 1960s and early 1970s, the British seaside resort industry struggled with both its own problems and with the emergence of newer and more fashionable seaside resorts elsewhere, especially in the Mediterranean region that drew large numbers of British tourists. By 1969, Spain was the leading destination by far for British overseas travelers, attracting 32% of them (Demetriadi 1997). And the shift continued, with British travelers going to Spain growing at 4.5% per year from 1970 to 1993.

There is a long list of factors that contributed to the decline in domestic seaside holidays. Cooper (1997) identified 20 threats facing British resorts after the Second World War, and he did not include basic economic variables, such as changes in exchange rates, that cliometricians would tend to favor.[20] So what did cause the shift to overseas destinations of seaside vacationers? Did it reflect a failure of domestic entrepreneurs in the seaside resort industry or the foresight, skill and judgment of the entrepreneurs in the Mediterranean? Was it the consequence of improved air transportation, which suddenly made the Mediterranean more accessible, or at least relatively more accessible compared with local transport to the seaside? Or, as many assert, was it due to the introduction of all-inclusive package tours, which took advantage of the improvements in air transportation, especially the introduction of jet airplanes, which made travel not only faster but also cheaper (Lyth 2009)?

Independent airlines were established after World War II to fly routes not served by scheduled airlines and to provide other airline service. They blossomed because they could easily acquire aircraft from the supply left after the War, and because for much of the 1950s, they were subsidized by government contracts to fly troops to war zones, especially to Korea. In 1965, a change in airline regulations gave independent airlines more freedom to compete with scheduled airlines, and they took advantage of the situation (Lyth 2009). Independent airlines in the UK realized before the nationalized airlines did that the demand for air travel was price-elastic, and that on faster flights travelers would not care much about the luxuriousness of the accommodations, and for lower prices would be willing to depart from lesser known airports. Travel agencies embraced these independent airlines and their low fares, and struck deals with hotels and resorts that enabled

[20] Absent to a large extent from discussions of this shift from seaside resorts in the UK to resorts in Spain is the impact of changes in taste. See Urry (1997) for a discussion of the impact of cultural changes.

them to offer all-inclusive package tours that could provide holidays in Spain cheaper than those that could be had in the UK.[21]

But there remain some questions as to the importance of tour packages and the timing of their impact on the transformation of the British seaside resorts. According to Demetriadi (1997), declines in seaside tourism in Britain before the mid-1970s were not due to the growth of overseas tour packages. Although they had begun to play a role in shaping the world market for tourist travel, they were not affordable to a broad segment of the population until after 1972. Overseas package tours did not cut into domestic resorts as evidenced by the small increase (0.75 million) in overseas travel between 1965 and 1970, but a much larger increase of 4.5 million in domestic trips. On the other hand, Cooper (1997) suggests that all-inclusive tours may have been of greater importance before 1974 than Demetriadi claimed, but in any case were among the key factors contributing to the decline after 1974.

The evidence on the prices of tour packages (Lyth 2009) lends some support to Cooper's notion that package tours may have been more important before 1974, but at the same time raises questions about their impact after 1974 or 1975. The nominal prices of tour packages rose by about 10% from 1966 to 1972, but when adjusted for inflation, real fares fell by about 5% between 1966 and 1970, and then fell considerably more from 1971 to 1972. From 1973 to 1977, however, package tour prices rose: more than doubling in nominal prices and rising by 27% in real terms. So, all other things being equal, tour package prices should have had more of an impact on increasing travel before 1974 than after.

Price was not the only consideration. There were other variables – such as changes in exchange rates, rising incomes, and increased holidays – that would have increased demand for tourism of all sorts; domestic as well as overseas, and seaside tourism as well as urban and heritage tourism. And there were also supply-side factors that came into play, in particular the availability and quality of tourist accommodations. Was the Spanish tourism industry better prepared to respond to an upsurge in tourist arrivals? Had they built ahead of demand and travel agents took advantage of that oversupply to construct all-inclusive tours? Why were the British seaside resorts unable – or unwilling – to respond in the same way to the rise of mass tourism?

The growth of tourism in Spain was due in part to changes on the supply side: "to the interaction of multiple actors, such as entrepreneurs, hotel groups, international travel agencies, specialized and non-specialized workers, local, regional and national governments" (Manera et al. 2009, p. 6). The nature and style of this assortment of economic agents varied across nations, particularly as regards Spain (and some other Mediterranean countries) vis-à-vis the UK (and other European countries). In the UK, as well as France and Germany, large, managerial enterprises with large bureaucratic and strong financial structures were more prominent, whereas medium- and large-sized groups of economic agents were more common in Spain (Manera

[21]Package deals also were a way to get around restrictions on the amount of currency that could be taken out of the country.

et al. 2009). Manera et al. argue those differences conditioned the response. In Spain, those business groups or associations responded by adapting the supply-side of tourism, changing the infrastructure, and constructing hotels and resorts geared toward mass tourism. In the UK, on the other hand, the suppliers of seaside tourism tried to cope with the increased demand by duplicating the same small-scale operations that had been so appealing and successful in the past. This reflects the baggage they were dealing with when the Mediterranean competition emerged.

Even when the UK's seaside resorts were thriving in the 1960s, there was a lack of quality accommodations. The older Victorian hotels were no longer fashionable, and remodeling and upgrading, as well as heating and lighting, were more expensive in the larger older hotels. This contributed to a gradual decline in the number of upper and middle income tourists (with some resorts being notable exceptions) who increasingly opted for overseas travel, which in turn led to some of the large hotels being sold or converted to blocks of flats.[22] The number of large-scale seaside hotels declined by 40% between 1961 and 1970. Moreover, the Second World War had opened up other jobs for women, so hotels had less cheap labor available, and the Catering Wages Act in the early1950s boosted wages. Seaside tourism was facing increased competition domestically from heritage sites and urban areas, especially London.

Cooper sees the British response in largely the same terms as Manera et al., but with a somewhat different perspective. The domestic resorts in the UK did little to upgrade their facilities because they had lots of older fixed capital and owners were unable or unwilling to adapt to the changes that affected the industry in the 1970s, in part due to the seasonality of the business. With a shorter season in the UK, resort hotels there did not appear as attractive an investment as urban hotels, or as attractive as hotels in the Mediterranean for that matter. Moreover, the existing capital stock was largely small scale, comprised mostly of independently owned hotels that could not afford to finance expansion.

The Spanish seaside resorts stood to benefit in either case; whether the response by the British resorts was an expansion based on replicating the same small-scale hotels and resorts, or by a lack of response for whatever reason. But Cooper argues further that the British-outbound tourism sector (airlines and travel agents) responded better to meet the increased demand arising from increased real incomes. And this response meant taking advantage of the advances in air travel and working with the newer and more fashionable Spanish hotels and resorts, many of which were established to attract large numbers of middle and lower income tourists (Lyth 2009), those that had become the mainstay of many of the British seaside resorts.

For whatever reason, the failed response in the UK stands in sharp contrast to the behavior of the hotel industry in Spain, perhaps because they were ready for mass

[22]See Demetriadi (1997). Seaside resorts were harmed as well by the rise of second and third vacations, which became more common and were shorter in duration. Travelers increasingly took their one long vacation overseas, and spent their shorter vacations and weekend breaks domestically and increasingly in locations that were more readily accessible than the seaside (Cooper 1997).

tourism, thanks to an earlier history of some seaside tourism. Alicante (Costa Blanca), Malaga (Costa del Sol), and Palma de Mallorca were famous resorts by the turn of the nineteenth century, with Alicante being known as the "Playa de Madrid," thanks to a direct rail link established in 1858 (Valenzuela 1998, p. 55). In any case, the existing hotel owners, especially on Majorca, responded much more adroitly than did their counterparts in Britain. Not only did they expand to accommodate more tourists on Majorca, they invested in hotels in other seaside locations, and continued from there to become some of the leading hotel companies in the world (Serra 2009).

Although this history indicates there was a massive shift in seaside tourism from Britain to Spain, and there is a widely held view in Britain that the domestic seaside resorts have been struggling ever since, the reality is not quite that dismal. Beatty and Fothergill (2003) found that between 1971 and 2001, employment and population in Britain's seaside towns grew slightly faster than the national average, a finding that harkens back to Walton's idea that, for some seaside cities, tourism was a means to attract more permanent residents. Beatty and Fothergill (2003) could not sort out employment in the tourist industries from all other employment, but a later study by Beatty et al. (2010) did so, and found that far from entering a spiral of decline caused by the loss of tourism business, the economy of Britain's seaside towns proved remarkably resilient over this period.[23] The study looked at employment, rather than numbers of visitors or resident population, and found that from 1999 to 2007, employment increased by a little more than 1% per year. Tourism in 2007 was still a major source of employment, equal in size to the telecommunications sector and bigger than the motor industry, aerospace, pharmaceuticals, or steel.[24] They conclude that "far from being in terminal decline as a result of the rise of foreign holidays, a substantial British seaside tourist industry remains alive and well" (Beatty et al. 2010, 10).

Economic History of Tourism in the United States

Tourism in the United States is not as quantitatively important as it is in Europe, or in most individual nations in Europe or elsewhere, but it is not a negligible part of the US economy. Value added by tourism in 2015 comprised 2.7 of GDP, larger than the share accounted for by utilities and mining (Osborne and Markowitz 2017, p. 4).[25] Tourism has accounted for 3.5–4.0% of the nation's employment

[23]They say it should be possible to apply their estimation techniques to earlier years in order to get a longer time series.

[24]Its output is low in relation to employment, which reflects the prevalence of low-wage and part-time employment in much of the industry (Beatty et al. 2010, p. 10).

[25]For those who ascribe to Douglass North's idea that American growth was propelled by the growth of cotton exports, it is worth noting that the income from those exports never exceeded 6% of GNP.

since the early 1990s.[26] Consumer demand for tourism output was equal to 90% of consumer spending on financial services and insurance, and nearly twice the size of spending on motor vehicles and parts.[27] Americans' demand for tourism consumption within the United States dwarfed their spending on outbound tourism and widely exceeded the demand by international travelers to the United States (OECD 2016, Tables 1 and 3).

The rise to its current importance took a long time. The history of tourism in America, as well as travel overseas by Americans, is similar to that for Europe. It was initially confined to the elite, but as incomes rose and transportation costs declined over time, travel became increasingly popular, but early on was still seen as extraordinary behavior. Eventually, so many could afford to travel, and did so, that tourism became a product – advertised widely and produced on a large scale. The absence of a reliable quantitative record of tourism's history, however, has resulted in an assortment of claims as to when American tourism, whether domestic or international, became so important or so large in scale that it could be deemed "mass tourism," and its services could be commodified. Although domestic tourism is larger than overseas travel by a wide margin, the latter is better documented and its long-term trends measured more reliably.

In the late nineteenth century, only around 100,000 Americans traveled overseas, and all of them did so by steamship. Today, there are in excess of 30 million overseas travelers, almost all of whom travel by air.[28] This long-term rise in overseas travel was not smooth and uninterrupted – it was punctuated with significant declines during periods of warfare and economic downturns and significant surges as in the 1950–1970 period – but it rose at an impressive average annual rate of about 5% between 1820 and 2000 (Dupont and Weiss 2013). There were essentially two phases to this long-term growth: prior to World War I when the increase in overseas travel was primarily driven by population growth, and the share of Americans traveling rose very little; and the post-World War II period when travel was driven by improvements in transportation technology and economic factors, such as rising incomes, and tourism rose noticeably faster than the population.[29]

The pace and timing of international overseas travel to the US were similar to those for Americans traveling overseas in the long run, but there were some noticeable differences in shorter time periods. And the volumes were different: the

[26]U.S. Bureau of Economic Analysis (2017) "Travel and Tourism Satellite Account" Sept. 2017, and "National Income and Product Accounts: GDP and Personal Income, Section 6". The earliest estimates cover 1992, 1996, and 1997 (Kass and Okubo 2000, Table 1, and an earlier and preliminary estimate for 1992 in Okubo and Planting 1998). The share rose to 4.3% in 1998–2000.

[27]U.S. Bureau of Economic Analysis (2017) "Travel and Tourism Satellite Account" Sept. 2017, and "National Income and Product Accounts: Gross Domestic Product Third Quarter 2017. Consumer demand for tourism is reported as Direct Tourism Output in the Satellite Accounts.

[28]See OECD (2016), Country Profiles-United States. There are also around 40 million who travel over land to Canada and Mexico.

[29]This statistical record covers only Americans traveling overseas, and does not include travel to places reached over land, mostly Canada and Mexico (Carter et al. 2006, series Dh324).

number of tourist arrivals was below the number of Americans traveling overseas, especially before World War II (see Fig. 1a), but even after the War when the former was less than half of the latter up to 1973. Not until the 1990s did the number of foreign travelers to the US equal or exceed the number of Americans traveling overseas (see Fig. 1b).

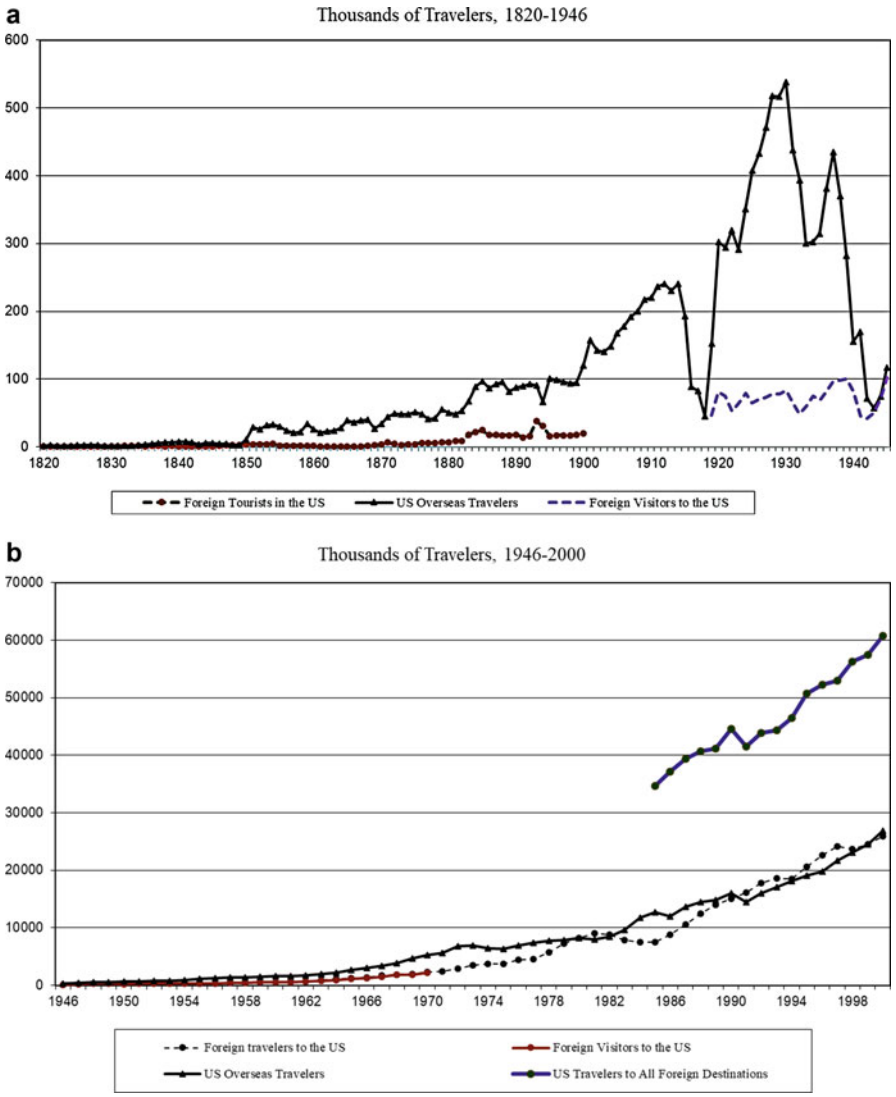


Fig. 1 (**a**) US overseas travelers: Dupont and Weiss 2013; Foreign tourists in the US and foreign visitors: Carter et al. 2006, Dh320 and Dh326. (**b**) US overseas travelers: Dupont and Weiss 2013; US travelers to all destinations, Carter et al. 2006, Dh324; Foreign travelers and visitors: Carter et al. 2006, Dh325–26

One of the biggest changes in overseas travel resulted from the considerable cost reductions that came with improved transport technology. These improvements meant that travel abroad, once available only to the wealthy elite, became increasingly open to middle-class Americans and foreign travelers as well. The most important technological shifts came first with the mid-nineteenth century transition from sailing to steam-powered vessels and then, roughly a century later, the transition from propeller to jet-propelled aircraft. The first transition to steam-powered vessels resulted in shorter passage times and more reliably predictable schedules, both of which reduced costs for passengers. Pan American initiated the mid-twentieth century transition in 1958, advertising "6 ½ magic hours to Europe." This transition reduced travel times further as the average speed of international aircraft increased from 224 mph to 482 mph between 1950 and 1970, although the time savings was much more modest than that generated by the shift from sail to steam. And it still takes 6 ½ hours to reach Europe!

Two papers (Dupont et al. 2011; Dupont and Weiss 2013) have attempted to sort out the relative importance of several factors as explanations of the long-term rise in overseas travel. Population growth alone would have led to an increase of about 1.9% annually from 1820 to 2000, accounting for about 46% of the growth in the number of overseas travelers prior to World War I, but only about 23% in the period from World War I to the year 2000. The remaining growth, due to the rising share of the population traveling abroad, was a small fraction of a percent in the early nineteenth century (between 0.01% and 0.02%), rose more than ten-fold by the early twentieth century, fell during the world wars and the Great Depression, but has risen steadily since the end of World War II. By the end of the twentieth century, the share of the US population traveling overseas was about 9%, some 30 times higher than it was at the end of World War II (Dupont et al. 2011, Table 1). By 2015, 74 million US citizens, about 23% of the population, traveled to international destinations.

The main reasons for the rising share of the population traveling overseas were rising incomes, changes in exchange rates, changes in passenger fares, and technological improvements, such as the shift from sailing to steam-powered vessels in the last half of the nineteenth century and the shift to jet-powered aircraft a century later. The empirical evidence indicates that Americans traveling to Europe responded to some degree to all of these factors, but were most sensitive to changes in per capita income and the price of travel (Dupont et al. 2011). Other factors mattered as well: improvements in hotels and restaurants, the publication of travel guide books, developments in package tours, and growth of official tourism offices abroad, but these are not easily quantified.

In contrast to the statistical evidence available about international travel, the history of domestic tourism in the US still rests largely on anecdote, at least up to the post-World War II era, and to some extent up until the BEA Satellite Accounts produced the first estimates for 1992. Even when scholars have written thoughtful histories about particular places, they have used only a handful of illustrative statistics. Based on that sort of evidence, it has been claimed that mass tourism arrived in the late nineteenth century (Brown 1995), between the world wars (Dulles 1965; Jakle 1985), or not until after World War II (Cocks 2001). Weiss (2004) compiled a crude time series up to World War II that suggests by 1930 tourism had

grown to include many middle-income Americans, but had not yet become mass tourism, which was postponed by two decades due to the disruptions of the Great Depression and World War II.

Clearly, we lack a statistically-based explanation of tourism's rise in the US. There are plausible explanations and identification of the factors that mattered, but no quantitative assessment of which factors were the more important. Cliometricians have not yet studied the growth of American tourism to the same extent as they have examined other industries, such as agriculture or finance, that are not as large or not much larger than tourism, or other components of consumer behavior and spending, such as the durables revolution. If cliometrics is thought of as having two components – the construction of new data series and the use of statistics instead of anecdote and narrative description, and the use of economic models to analyze the data – then so far cliometric research on American tourism has been focused more on the first of these.

Weiss's (2004) survey of tourism before World War II is an example of the first prong, an attempt to measure the rise in the numbers of tourists. Although his evidence has its shortcomings, is limited to a few benchmark dates, and does not extend beyond World War II, it offers a plausible depiction of the long-term trend for the nation and of the relative importance of several different types of tourism and of some specific tourist sites before 1950. Moreover, the derivation of the national totals is clearly specified, so others can replicate the results and revise them as new information comes to light.

For the period after World War II, there are some time series data available from federal and state sources. Estimates of expenditures by travelers have been made at the federal and state level for many years since 1960, but these figures have never been assessed by scholars outside the agencies that created them, nor have they been used in any long-term study of tourism. The consistency among the different sets of estimates has not been determined. Nevertheless, these various data sets could be assessed for consistency with one another and, with appropriate adjustments, combined to form a very useful time series for the post-World War II period.

Data Sets Available

In 1963, 1967, 1972, 1977, and 1982, the Bureau of the Census conducted a National Travel Survey as part of the Census of Transportation. Among the information collected was the number of trips, mileage traveled, and number of days away from home. The Federal Highway Administration conducted the National Household Travel Surveys beginning in 1969, but most of those covered only daily travel, not long distance travel.[30] Two of those surveys, however, the *American Travel Survey* conducted in 1995 and the *National Household Transportation Survey* taken

[30]Surveys were taken in 1969, 1977, 1983, 1990, 1995, and 2001. In the second survey, taken in 1977, long-distance travel was included but was confined to trips of 75 miles-or-more taken during the 14-day period preceding the survey.

in 2001, covered long-distance trips. The former included trips of 100 miles or more taken in 1995, while the latter included trips of 50 miles or more taken in 2001. None of these travel surveys, however, included travel expenditures, but some of the data collected were used by the United States Travel Data Center to reconstruct expenditures.

The US Travel Data Center[31] developed a Travel Economic Impact Model that enabled them to use the travel activity data collected by the Bureau of the Census and average costs of each unit of travel activity (e.g., cost per night by type of accommodation) obtained from other sources, such as the Census of Retail Trade, to estimate tourists' expenditures.[32] The first year for which the Travel Data Center made an estimate was 1974, but it was not until 1983 that the estimates were made readily available with publication in the *Statistical Abstract of the United States*. From 1983 to 2001, the figures were published annually, except in 5 years: 1988, 1991, 1992, 1997, and 2002.

Before 1982, agencies in various states collected data on travel, including travel spending, in their respective states. Some of these studies were conducted under the auspices of the Department of Transportation. The Business Research Division at the University of Colorado periodically gathered data from various state sources covering the period 1960–1982, and published compilations every 2 or 3 years from 1969 to 1984. While data are reported for most states in one or more years, there are some missing observations because not all states collected data in all years. The data for California, for example, are missing in about half the years in the period 1960–1982. Moreover, the consistency of the data across states and even over time within the same state appears questionable in some instances. For example, expenditures in New Jersey were reported as only $778 million in 1972, but $3.5 billion 2 years later.

Implications of These Data

If we use these series in their raw form – that is to say, we ignore any inconsistencies within the state-based series, and take those data as being comparable with those for the period after 1983 – they show several points worth noting. Growth in the aggregate for the nation was substantial in nominal terms, but much less so in constant prices. Between 1960 and 2003, travel expenditures rose at an average annual rate of 8.2% per year, but only 3.7% per year in constant prices of 1982–1984. Almost all the growth in real expenditures per capita for the nation took place before 1983 (see Table 1).[33] Over the entire period from 1960 to 2003, per

[31]The Travel Data Center was established in 1973 and is now called the U.S. Travel Association. https://www.ustravel.org/.

[32]Their model specified 15 different expenditure categories, such as commercial lodging and variable auto/truck costs, and related SIC business types (Frechtling 1976).

[33]The national totals for this period vary in terms of the states covered from year to year, but even if we made generous allowances for the states that were missing in these years, the increase in spending during this period would be greater than in the post-1983 period.

Table 1 Travel spending in the United States. (Million of dollars, except per capita figures)

	Nominal	Real	Real spending per capita
1960	$16,580	$56,015	$314
1967	$25,747	$77,088	$395
1983	$197,597	$198,393	$855
2003	$490,870	$266,777	$917
Average annual rates of change (%)			
1960–1983	11.38	5.65	4.45
1960–1967	6.49	4.67	3.33
1967–1983	13.58	6.09	4.94
1983–2003	4.65	1.49	0.35
1960–2003	8.20	3.70	2.52

Spending per capita is for the resident population, not the number of tourists. Real expenditures are in prices of 1982–1984. For sources, see the text.

capita expenditures in constant prices (1982–1984 = 100) rose by $603, or at an annual average rate of 2.5%. Between 1983 and 2003, the period for which the Travel Center's estimates are readily available, per capita spending increased by only $62. Between 1983 and 2003, there were ups and downs in travel spending, but only a sluggish rate of increase of 0.35% per year.

Along with the increases, there was a noticeable redistribution in the importance of tourism spending among states. For example, California, which today ranks at the top of the list, was only in 7th place in the 1960s, whereas Florida, which was number one in the 1960s, had slipped behind California by 1983. Also of note is that Hawaii, often seen as the premier tourist destination, has never made the top ten list. In fact, while it rose in importance between the 1960s and 1983, going from 34th to 27th place, it subsequently fell back in the rankings.[34]

Explaining the Rise of Domestic Tourism

Perhaps it is not surprising that there has been no cliometric work done to explain the rise of tourism in the United States as a whole. Tourism accounts for only 3–4% of GDP, and it is not an industry known for technological breakthroughs that might have spillover effects on productivity in other industries. Nevertheless, the issues are of greater interest in those states and cities where tourism is relatively more important. This would include states, such as Florida and Hawaii; cities, such as Orlando and Las Vegas; and specific sites, such as Niagara Falls and the Grand Canyon. Even at this level, the historical analyses have not been cliometric in nature, even though there are any number of instances in which counterfactual analysis would have been very appropriate and informative. The history of tourism in Hawaii offers an excellent case study of what we might call near-cliometric analysis.

[34]Space limitations preclude the inclusion of a fuller table showing the state rankings over time.

History of Tourism in Hawaii

Hawaii is one of the mostly highly rated tourist destinations in the world. As attractive as Hawaii now seems, tourism could not have reached its current heights without transportation innovations, especially jet planes, and might not have succeeded as much or as easily without the infrastructure that had been built to serve agricultural markets. Some of its early success in attracting and accommodating travelers, which laid a foundation for subsequent tourist growth, was due to its location on major shipping routes from North America to Asia and Australia. The private sector may have played the dominant role, but government took an active role in shaping the direction in which the economy should go, the extent to which tourism should be managed, and was concerned with "sustainable tourism" as early as 1976, well before that concept was defined in 1992 (Mak 2008).

Perhaps because the government was involved early on in the development of tourism in Hawaii, or because it was an island economy that could more easily monitor and measure some of its economic activity, it is a place for which one can find pertinent historical statistics on tourism. Data collection began around 1911, carried out by the Hawaii Promotion Committee, an organization founded by the Honolulu Chamber of Commerce in 1903 to promote tourism to Hawaii. Its successor organizations, the Hawaii Tourist Bureau (1919), the Hawaii Visitors Bureau (1945), and finally the Hawaii Visitors and Convention Bureau (1997), continued that work producing what were considered "the best tourism statistics in the world"[35] (Mak 2008, p. 111). And, in 1977, Robert Schmitt compiled the *Historical Statistics of Hawaii* to be a counterpart to the *Historical Statistics of the United States* covering the same demographic, economic, and social topics. The tables are complemented by discussion of the development of the statistics in each topical category, and by provision of pertinent definitions, notes, and qualifications, as well as reference to appropriate published sources. One can find estimates of population back to 1778–1779, annually from 1832 to 1976; income from the major export industries (sugar, pineapple, defense, and tourism) annually from 1910 to 1975; gross state product annually from 1939 to 1975; and visitor arrivals annually from 1922 to 1975. Moreover, many of these data, as well as other time series, and more recent statistics, can be accessed on the state government's website (Hawaii http://dbedt.hawaii.gov/).

These data reveal several notable points. Hawaii has always been heavily dependent on exports (see Fig. 2). Just prior to World War II, the value of Hawaii's four major exports – pineapple, sugar, military spending, and visitor spending – equaled 50–60% of the state's gross product, a dependence that suggests there would likely be a continual search for goods and services that could generate export revenues. Although visitor spending was the smallest of the four at that time – about 5% of gross product – the conditions were conducive to its expansion. And not long after

[35]In 1998, the research function of the Visitor and Convention Bureau was turned over to the state's Department of Business, Economic Development and Tourism (Mak 2008).

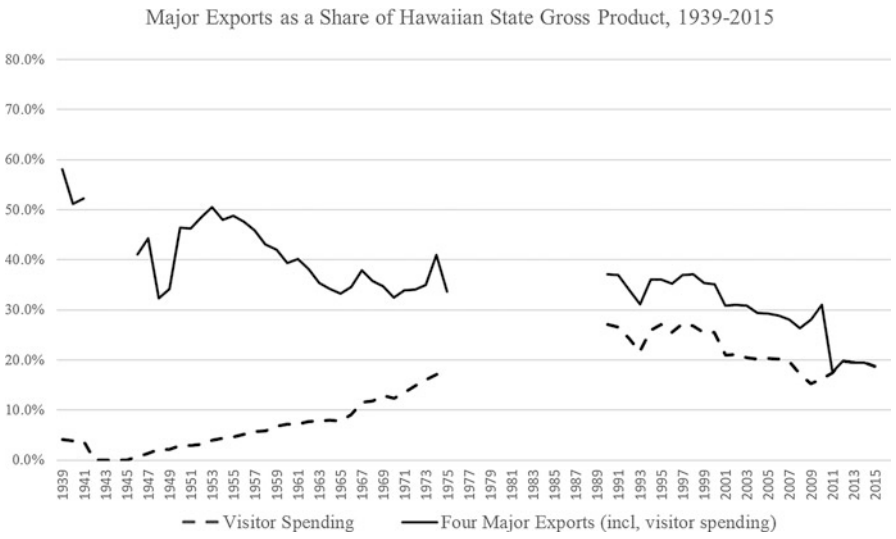


Fig. 2 The major exports are sugar, pineapple, defense spending, and visitor spending. Visitor spending includes spending by travelers from the continental US, so the total is not the value that would appear in the international accounts. (Mak 2015, Table 1, http://dbedt.hawaii.gov/economic/databook/db2016/)

World War II, the expansion began. Growth of visitor arrivals was quite rapid up through the 1980s, languished through the 1990s and the beginning of the twentieth century, and since 2010 has experienced another growth spurt (see Fig. 3). The value of visitor spending, which had fallen to zero during the War, rose to equal about 30% of state gross product and is the only one of the four major pre-War exports of any importance today (see Fig. 2).

These historical data have been used by economic historians to explain the rise of tourism, its leveling off for nearly two decades, the role of government, and tourism's impact on the local economy and population. Although these analyses are not full-blown cliometric studies, the authors rely heavily on statistics and use economic models, at least implicitly, to explain the major patterns in the state's tourism history.

According to James Mak (2008), tourism in Hawaii is essentially a post-World War II phenomenon. Early travelers to the islands were predominantly sailors with few visitors, most of whom were only passing through. What might be seen as a tourist trade can be said to have begun in 1867 when regular steamship travel commenced between Hawaii and the US mainland. Limited growth continued until the end of the century as steamship service improved; it was more frequent, more comfortable, and faster, although still involved traveling on cargo ships. Not until 1927 did the first passenger ship, the *Malolo* with a capacity of 650, make its maiden voyage. Pan Am Clipper airline service began in 1936, but carried only seven passengers at a fare of $356. Despite regular airline service, most visitors

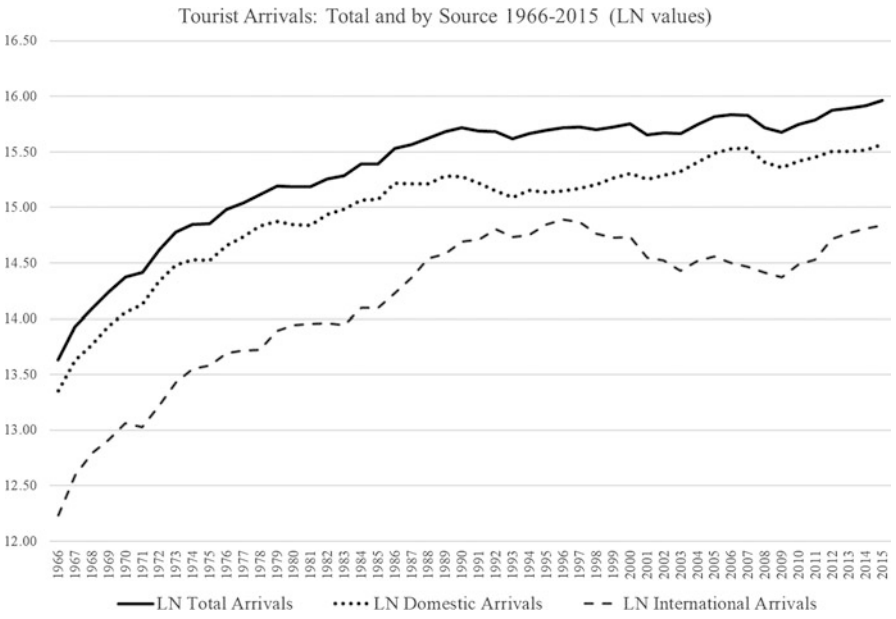


Fig. 3 Domestic arrivals are those from the mainland US. International arrivals are largely from Japan. (Mak 2015, Table 1 and Schmitt 1977, Table 11.7)

arrived by ship until well after World War II. The War shut down the tourist trade, and it took more than a year after the war to return to normal. Then, in 1947, United introduced DC-4 s, which were faster and cheaper ($225 fares) than the Clippers, and in 1948 Northwest initiated flights from Seattle and Portland.

All these events were only a prelude to the tourist boom that was to come and laid a foundation for the mass tourism that came after statehood was achieved in 1959. For example, when private capital was not interested in investing in a first-class hotel after the tourist trade began in the late 1860s, the Hawaiian monarchy did, building the Hawaiian Hotel in downtown Honolulu. The hotel was completed in 1872 and shortly thereafter a small annex in the Waikiki area was built and enlarged in 1894, making it the first major beach hotel in the islands (Mak 2008). Accommodations do not appear to have been a constraint on the growth of tourism before statehood. The number of tourists rose from about 2000 in the 1880s to around 9676 in 1922 (Schmitt 1977, Table 11.7). While that growth may not have propelled the Hawaiian economy, it showed that the state could play a role in developing tourism.

With statehood in 1959 came “a huge surge in tourist travel to Hawaii from the US mainland” (Mak 2008, p. 16) due primarily to increases in airline capacity and reduced airfares. In 1959, Pan Am began jet clipper service and by the mid-1960s almost all propeller airplanes had been replaced by jets, which cut travel time from days in the case of ships to around 5 h. In the1960s, fares set by the CAB fell 46%, most of it during the 1960–1965 period. Yet, with all these favorable conditions, it is

a little surprising to see that the rate of growth of tourist arrivals was about the same in the 10 years before statehood as the 10 years after.[36] What then was the impact of the decline in airfares, the increase in airline capacity, and increases in real income in the US after 1959? Why did tourism growth not accelerate in the 1960s instead of continuing along the same trend? Was tourism growth held back by other forces, such as supply constraints on the islands? These seem like questions ripe for counterfactual analysis, but economic historians have not addressed them.

Transport improvements continued to foster tourism. In 1969, the Civil Aeronautics Board allowed five more passenger airlines to fly from the US mainland to Hawaii, which meant more cities had easier access to Hawaii either directly or via connections. In 1970, Pan Am introduced the 747 jumbo jet, which still took 5 h, but carried many more passengers. And in the late 1970s, the airline industry was deregulated, which resulted in further declines in air fares, especially on long-haul routes and for recreational travel.

The rise in the annual volume of tourists from 242,000 in 1959 to 6.7 million in 1990 was due in part to a large increase in Japanese tourists that began in the late 1960s. That upsurge was due to a relaxation of Japanese prohibition on outbound tourism until 1964, and easing of restrictions on the number of dollars that could be spent overseas (Mak 2008). Once restrictions were lifted, conditions were favorable. Jumbo jets had already been introduced in 1970 and packaged tours arose to take advantage of that airline capacity. With Japan running surpluses in their balance of trade, the yen increased in value and made travel that much less expensive.

Japanese arrivals continued to rise until 1997, except for a dip in 1993–1994. Nevertheless, the Hawaiian tourism industry struggled throughout 1990s. As Grandy (2002, p. 40) put it, beginning in 1991 "the bottom had fallen out of westbound arrivals," i.e., arrivals from the US mainland (see also Fig. 12, p. 22). And just as US arrivals started to rise again near the end of the decade, Japanese arrivals declined between 1997 and 2003, a phenomenon Mak (2008, p. 23) describes as "somewhat of a mystery." Overall, the 1990s were a troubled period, due to a series of events, beginning in 1991 with the first Gulf War, the collapse of the Japanese bubble, and a US recession, especially in California. Then in 1992, Hurricane Iniki hit in September, and a fare war among airlines for travel within the US mainland made flights to Hawaii relatively more expensive. Finally, the Asian Financial crisis of 1997–1998 dampened demand.

There may have been other reasons as well: Waikiki was shabby and in need of renovation, while other destinations – Japan, Australia, New Zealand, and Canada – had become more competitive and were reducing Hawaii's share of the tourist market. The new century did not fare any better at the start with the terrorist attack on the World Trade Center in 2001 and SARS epidemic in Asia in 2003. But after the

[36]See Mak (2008). The rate in the 10 years before statehood was 22% per year from 1949 to 1959 and for the 10 years after it was 20% from 1959 to 1969 (calculated from data in Schmitt (1977, pp. 273–74, Table 11.7).

initial shock of 9/11, Hawaii may have benefited from Americans' fear of traveling abroad and showed a recovery after 2003 (Mak 2008).

The high rates of growth of tourists in the 1950s and 1960s appear to have fostered economic growth at that time. In 1959, income per person in Hawaii was 20% below the national average; by 1970, it equaled the average and, according to Mak (2008, p. 30) ". . .this convergence was largely due to the extraordinary growth of tourism in Hawaii."[37] At the beginning of the period, tourism accounted for 16% of the four major exports, but it had increased to 38% by 1970. Over this period, visitor spending rose by 400%; if one adheres to an export-driven model of economic growth, one would surmise that tourism's impact must have been substantial.

The rapid growth of tourism in the 1960s led some residents and politicians to question whether unchecked tourism growth was good for the state. Even on Maui, which had only just begun to experience tourism growth, there was concern that it could "'spoil the charm and beauty of Maui'" (Blackford 2001, p. 24). Further resentment arose in the 1970s even though tourist arrivals increased at only 8.8% per year. The "Babbie Report" prepared by a public interest group tried to show that the growth of tourism and its induced growth of the resident population resulted in declines in the quality of life (Babbie 1972). It argued further that the quality of life could be stabilized only by halting in-migration, which meant halting growth of jobs, which meant halting growth of tourism (Mak 2008). Others argued as well that too much growth would erode the Aloha spirit, which would significantly reduce tourism to the islands.

By the 1980s, there were further concerns that the state's economy had become too dependent on tourism. There were a variety of responses, including the creation of the Maui Economic Development Board, which was charged with trying to create a more diversified economy by attracting high-tech companies to the area, and develop a more diversified agriculture (Blackford 2001). Neither of those plans worked out very well, at least not by the end of the century, despite having invested nearly $50 million on high-tech ventures. Another effort came when the Office of State Planning established a policy requiring developers to create one nontourism-related job for each hotel room the developer wanted to build, but Mak says the policy did not have much of an impact on the development of tourism.[38] And as described above, it may have been a moot issue as other events slowed the growth of tourism.

Despite these efforts to diversify the economy and a continued slowing of tourist arrivals in the 1990s, tourism remained the dominant export at the end of the decade; at nearly $10 billion, it was 2.5 times the size of the other three major exports and equal to 35% of the state's gross domestic product. Given tourism's greater importance to the economy, its slowing down and stagnation were a more serious concern.

[37]The comparison used current price output values, implicitly assuming that prices rose in Hawaii at the same rates as for the US.

[38]On principle, the developer had the alternative of paying the equivalent of a nontourist job in cash ($25,000); in practice, the requirement could be met in a number of other ways as well.

As the economy continued to struggle in the 1990s, the state attempted to revitalize the tourism industry.[39] Its chief effort was to create the Hawaii Tourism Authority, which would be funded by a hotel tax (Mak 2008). The authority's first plan put forth in 1999 was to promote tourist arrivals, diversify the tourist product, and increase visitor spending per person per day. The latter meant offering higher quality tourism products and services, especially hotels and resorts. Some evidence on average spending for lodging and in total indicates that different islands took different paths, with Maui choosing the high end of the market (Mak 2008). Other evidence, notably the efforts to attract more Chinese tourists, suggests that maximizing tourist arrivals was still a higher priority than attracting higher-spending tourists.

The state and local governments took other actions as well, including the construction of a conference center in Honolulu, increasing funding for marketing, and adopting tax incentives to encourage hotel renovations and build a world-class aquarium (Grandy 2002). Each of these efforts appears to be an event or policy ripe for cliometric examination. Interestingly, Mak provides a narrative of what a counterfactual analysis would have to consider in order to assess one aspect of this government effort to revitalize tourism. He asks if the fairly large increase in spending on the promotion of tourism in 1999 was the reason for the rise in tourism that occurred afterward. But he then begs off answering the question, arguing that it is difficult to determine because there were too many factors that came into play. These included economic conditions in the US and Japan, the yen/dollar exchange rate, lingering effects of the Asian financial crisis, and crises yet to come, such as the September 11 terrorist attack in 2001 (Mak 2008).

Overall, Mak downplays the role of government plans and policies, in part because the government did not impose tight controls over tourism development. Land use laws were the chief means by which the government could control development, and that was not very inhibiting.[40] So, why was the private sector able to be so successful in an industry with substantial external benefits and costs? There seem to be two general reasons. Tourism planning brought all the stakeholders together, helped to inform residents about the role and value of tourism, and helped to form a shared vision for the state's tourism product. Perhaps more important and tangible is that some major private tourist developments in Hawaii have taken into account what we normally think of as externalities, such as preserving or enhancing scenic beauty. The internalization of these benefits was accomplished because of the high concentration of land ownership. Only 72 private owners controlled 46% of all the land in Hawaii. At the extreme, one company owned 98% of all the land on the island of Lanai; an island that was once a major producer of pineapples could, with little government involvement or regulation, be converted into one of the world's most exclusive resort destinations. The owners had incentive to do so because they could capture all the external benefits (Mak 2008).

[39]See Grandy (2002) for a discussion of these efforts.

[40]See Blackford (2001, esp. ch.3) for a discussion of land tax issues as well as Native Hawaiian land issues.

Conclusion

In the cocktail scene from *The Graduate*, a 1967 movie starring Dustin Hoffman, a friend of the family pulls the new graduate aside to give him advice about his future and says, "I want to say one word to you. Just one word. . ..plastics. . .there's a great future in plastics." Today, if they remade the movie, or at least that scene, the friend offering advice would more likely say – tourism! While not identified as an industry in standard industrial classifications, tourism today is an important economic activity in every country, and especially so in some. Although its importance is largely a post-World War II phenomenon, it has been around for centuries, so it has a long economic history. And yet, economic historians and cliometricians have not done much research in the area.

There is great potential for cliometric research in the field. For the period after the Second World War, data are becoming increasingly abundant as tourism's importance increases, so there is plenty of raw material to work with. There is a scarcity of data for the period before the War, but this is the sort of problem cliometricians love to tackle, both to expand the statistical database and to apply imaginative techniques to the sparse data.

Other economists have engaged in tourism research on topics of general interest, such as estimation of price and income elasticities and testing the validity of the hypothesis that tourism growth can generate broader economic growth. Although that work is not economic history, it has identified pertinent time series and data sources, and highlighted current issues that would benefit from a longer term historical perspective.

We also tried to demonstrate that there are interesting long-term studies to be done. We focused on the history of seaside tourism, which began with the rise of British seaside resorts and the subsequent shift of much of that market to the mass tourist resorts in Spain and elsewhere; and on the rise of tourism in Hawaii and its impact on the state's economy. In both cases, scholars have evaluated the appropriate measures of tourism, assessed the available data, and compiled pertinent statistics. Nevertheless, there is much that cliometricians could do. Quantitatively assessing the relative importance of the various factors that shaped the growth and distribution of the seaside resort industry is just one such example. There are also opportunities to conduct counterfactual analyses of tourism. Would Hawaii have prospered more if it had not shifted so heavily to tourism? Which British seaside resorts, if any, might have been better off without having entered the resort business?

There are many other topics in tourism history that we have not addressed, but which should appeal to cliometricians. What have been the consequences for labor productivity and wages of an economy's shift to tourism? What has been the impact on female labor force participation? Has a shift to tourism reduced or increased the likelihood of environmental damage? Can tourism growth be sustained indefinitely, or does it inevitably stagnate because of its impact on the resident population? In all its aspects, creating new databases, applying the latest statistical techniques to their analysis, and using economic theory, cliometrics is well poised to contribute to our knowledge of the tourism industry.

Cross-References

- ▶ Historical Measures of Economic Output
- ▶ Political Economy

References

Adams PD, Parmenter BR (1995) An applied general equilibrium analysis of the effects of tourism in a quite small, quite open economy. Appl Econ 27:985–994
Anastasopoulos PGE (1984) Interdependencies in international travel: the role of relative prices: a case study of the Mediterranean region. PhD. Dissertation, New School for Social Research
Babbie E (1972) The Maximilian report. Citizens for Hawaii, Honolulu
Balaguer J, Cantavella-Jordà M (2002) Tourism as a long-run economic growth factor: the Spanish case. Appl Econ 34:877–884
Beatty C, Fothergill S (2003) The seaside economy: the final report of the seaside towns research project. Sheffield Hallam University, Centre for Regional Economic and Social Research
Beatty C, Fothergill S, Gore T, Wilson I (2010) The seaside tourist industry in England and Wales. Sheffield Hallam University, Centre for Regional Economic and Social Research
Bhagwati J (1988) Export-promoting trade strategy: issues and evidence. The World Bank Research Observer 3(1):27–57
Blackford M (2001) Fragile paradise: the impact of tourism on Maui, 1959–2000. University Press of Kansas, Lawrence
Blake A, Sinclair TM, Campos-Soria JA (2006) Tourism productivity: evidence from the United Kingdom. Ann Tour Res 33(4):1099–1120
Box G, Jenkins G (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco
Brida JG, Cortes-Jimenez I, Pulina M (2016) Has the tourism-led growth hypothesis been validated? A literature review. Curr Issue Tour 19(5):394–430
Brown D (1995) The twentieth-century tour: the decline of the great hotels. In: Tolles B (ed) Historical New Hampshire. New Hampshire Historical Society, Concord, pp 125–140
Burk K (2005) The grand tour of Europe. April 5 lecture at Gresham college. https://www.gresham.ac.uk/lectures-and-events/the-grand-tour-of-europe. Accessed 27 Jan 2018
Business Research Division (1969, 1971, 1973, 1975, 1978, 1981, 1984) Travel trends in the United States and Canada. University of Colorado, Boulder
Capo J, Riera FA, Rossello NJ (2007) Tourism and long-term growth: a Spanish perspective. Ann Tour Res 34(3):709–726
Carter S, Gartner SC, Haines MR, Olmstead AL, Sutch R, Wright G (eds) (2006) Millennial edition of historical statistics of the United States. Cambridge University Press, New York
Casson L (1974) Travel in the ancient world. George Allen & Unwin, London
Chan F, Lim C, McAleer M (2005) Modelling multivariate international tourism demand and volatility. Tour Manag 26(3):459–471
Christensen LR, Jorgensen DW, Lau LJ (1975) Transcendental logarithmic utility functions. Am Econ Rev 65(3):367–383
Cocks C (2001) Doing the town: the rise of urban tourism in the United States, 1850–1915. University of California Press, Berkeley
Cooper C (1997) Parameters and indicators of the decline of the British seaside resort. In: Shaw G, Williams A (eds) The rise and fall of British coastal resorts: cultural and economic perspectives. Mansell, London, pp 79–101
Costa D (1997) Less of a luxury: the rise of recreation since 1888. NBER working paper 6054
Crouch GI (1994) Demand elasticities for short-haul versus long-haul tourism. J Travel Res 33 (2):2–7
Crouch GI (1995) A meta-analysis of tourism demand. Ann Tour Res 22(1):103–118

Culiuc A (2014) Determinants of international tourism. IMF Working Paper
Deaton A, Muelbauer J (1980) An almost ideal demand system. Am Econ Rev 70(3):312–326
Demetriadi J (1997) The golden years: English seaside resorts 1950-1974. In: Shaw G, Williams A (eds) The rise and fall of British coastal resorts. Mansell, London, pp 49–78
Dulles A (1965) America learns to play. Appleton-Century-Crofts, New York
Dupont B, Weiss T (2013) Variability in overseas travel by Americans, 1820–2000. Cliometrica 7(3):319–339
Dupont B, Gandhi A, Weiss T (2011) The long-term rise in overseas travel by Americans, 1820–2000. Economic History Review 65(1):144–167
Dwyer L (2015) Computable general equilibrium modelling: an important tool for tourism policy analysis. Tourism and Hospitality Management 21(2):111–126
Forsyth P, Dwyer L, Spurr R (2014) Is Australian tourism suffering Dutch disease? Ann Tour Res 46:1–15
Fowler WM Jr (2017) Steam titans: Cunard, Collins, and the epic battle for commerce on the North Atlantic. Bloomsbury, New York
Frechtling D (1976) Proposed standard definitions and classifications for travel research. Marketing travel and tourism, seventh annual conference proceedings. Travel Research Association, Boca Raton
Fujii E, Khaled M, Mak J (1985) An almost ideal demand system for visitor expenditures. JTEP 19 (2):161–171
Gatt W, Falzon J (2014) British tourism demand elasticities in Mediterranean countries. Appl Econ 46(29):3548–3561
Grandy C (2002) Hawai'i becalmed: economic lessons of the 1990s. University of Hawai'i Press, Honolulu
Gunter U, Smeral E (2016) The decline of tourism income elasticities in a global context. Tour Econ 22(3):466–483
Hawaii Department of Business, Economic Development & Tourism. http://dbedt.hawaii.gov/. Accessed 12 July 2018
Jakle J (1985) The tourist: travel in twentieth-century North America. University of Nebraska Press, Lincoln
Kass DI, Okubo S (2000) U.S. travel and tourism satellite accounts for 1996 and 1997. Surv Curr Bus 80:8–24
Keum K (2010) Tourism flows and trade theory: a panel data analysis with the gravity model. Ann Reg Sci 44(3):541–557
Kimura F, Lee H (2006) The gravity equation in international trade in services. Rev World Econ 142 (1):92–121
Krueger A (1980) Trade policy as an input to development. Am Econ Rev 70(2):288–292
Lim C (1997) Review of international demand models. Ann Tour Res 24(4):835–849
Lim C (1999) A meta-analytic review of international tourism demand. J Travel Res 37(3):273–284
Lim C, McAleer M (2001) Cointegration analysis of quarterly tourism demand by Hong Kong and Singapore for Australia. Appl Econ 33(12):1599–1619
Lyth P (2009) Flying visits: the growth of British air package tours, 1945-1975. In: Segreto L, Manera C, Pohl M (eds) Europe at the seaside: the economic history of mass tourism in the Mediterranean. Berghahn Books, New York, pp 11–30
Mak J (2008) Developing a dream destination: tourism and planning in Hawaii. University of Hawai'i Press, Honolulu
Mak J (2015) Creating 'paradise of the Pacific': how tourism began in Hawaii. University of Hawai'i at Manoa, working paper no. 15-03
Manera C, Segreto L, Pohl M (2009) The Mediterranean as a tourist destination: past, present, and future of the first mass tourism resort areas. In: Segreto L, Manera C, Pohl M (eds) Europe at the seaside: the economic history of mass tourism in the Mediterranean. Berghahn Books, New York, pp 1–10
Marrocu E, Paci R (2011) They arrive with new information: tourism flows and production efficiency in the European regions. Tour Manag 32(4):750–758

McKinnon RI (1964) Foreign exchange constraints in economic development and efficient aid allocation. Econ J 74(294):388–409
Meng S (2014) The role of inbound tourism in the Singaporean economy: a computable general equilibrium (CGE) assessment. J Travel Tour Mark 31(8):1071–1089
Morley C, Rossello J, Santana-Gallego M (2014) Gravity models for tourism demand: theory and use. Ann Tour Res 48:1–10
Narayan P (2004) Economic impact of tourism on Fiji's economy: empirical evidence from the computable general equilibrium model. Tour Econ 10(4):419–433
North D (1961) The economic growth of the United States, 1790–1860. Prentice-Hall, Englewood Cliffs, NJ
O'Hagan JW, Harrison MJ (1984) Market shares of US tourist expenditure in Europe: an econometric analysis. Appl Econ 16(6):919–931
OECD (2016) OECD tourism trends and policies 2016. OECD Publishing, Paris https://doi.org/10.1787/tour-2016-en
Oh C (2005) The contribution of tourism development to economic growth in the Korean economy. Tour Manag 26(1):39–44
Okubo S, Planting M (1998) US travel and tourism satellite accounts for 1992. Survey of Current Business July:8–22
Osborne S, Markowitz S (2017) US travel and tourism satellite accounts for 2013-2016. Survey of Current Business June:1–6
Peng B, Song H, Crouch GI, Witt SF (2015) A meta-analysis of international tourism demand elasticities. J Travel Res 54(5):611–633
Platzer MD (2014) US travel and tourism: industry trends and policy issues for congress. Congressional Research Service, Washington DC
Richards G (1972) Tourism and the economy: an examination of methods for evaluating the contribution and effects of tourism in the economy. University of Surrey, Surrey
Sakai M (2009) Public sector investment in tourism infrastructure. In: Dwyer L, Forshyth P (eds) International handbook on the economics of tourism. Edward Elgar, Cheltenham, UK
Schmitt RC (1977) Historical statistics of Hawaii. University Press of Hawaii, Honolulu
Segreto L, Manera C, Pohl M (eds) (2009) Europe at the seaside: the economic history of mass tourism in the Mediterranean. Berghahn Books, New York
Serra A (2009) The expansion strategies of the Majorcan hotel chains. In: Segreto L, Manera C, Pohl M (eds) Europe at the seaside: the economic history of mass tourism in the Mediterranean. Berghahn Books, New York, pp 125–143
Shaw G, Williams A (eds) (1997) The rise and fall of British coastal resorts: cultural and economic perspectives. Mansell, London
Sinclair MT (1998) Tourism and economic development: a survey. J Dev Stud 34(5):1–51
Sinclair MT, Stabler M (1997) The economics of tourism. Routledge, London
Song H, Li G (2008) Tourism demand modelling and forecasting: a review of recent research. Tour Manag 29(2):203–220
Song H, Wong KF (2003) Tourism demand modeling: a time-varying parameter approach. J Travel Res 42(1):57–64
Song H, Romilly P, Liu X (2000) An empirical study of outbound tourism demand in the UK. Appl Econ 32(5):611–624
Song H, Kim JH, Yang S (2010) Confidence intervals for tourism demand elasticity. Ann Tour Res 37(2):377–396
Stone R (1953) Cost and production functions. Princeton University Press, Princeton
Syriopoulos TC, Sinclair MT (1993) An econometric study of tourism demand: the AIDS model of US and European tourism in Mediterranean countries. Appl Econ 25(12):1541–1552
Tang C-H, Jang S (2009) The tourism-economy causality in the United States: a subindustry level examination. Tour Manag 30(4):553–558
Thiel H (1965) The information approach to demand analysis. Econometrica 33(1):67–87
Towner J (1966) Tourism history: past, present and future. In: Seaton AV (ed) Tourism: the state of the art. John Wiley & Sons, Chichester, pp 721–728
Towner J (1985) The grand tour: a key phase in the history of tourism. Ann Tour Res 12(3):297–333

Towner J, Wall G (1991) History and tourism. Annals of Tourism 18(1):71–84
UNCTAD (2017) Economic development in Africa: tourism for transformative and inclusive growth. United Nations, New York and Geneva
Urry J (1997) Cultural change and the seaside resort. In: The rise and fall of British coastal resorts: cultural and economic perspectives. Mansell, London, pp 102–113
U.S. Bureau of Economic Analysis (2017) https://www.bea.gov/industry/index.htm. Accessed Sept 2017
U.S. Census Bureau (various years) Statistical abstract of the United States. Government Printing Office, Washington, DC
U.S. Department of Commerce (various years) Survey of international air travelers program. Retrieved from https://travel.trade.gov/research/programs/ifs/index.asp
U.S. Travel Data Center https://www.ustravel.org/. Accessed 12 July 2018
Valenzuela M (1998) Spain: from the phenomenon of mass tourism to the search for a more diversified model. In: Williams A, Shaw G (eds) Tourism and economic development, 3rd edn. Wiley, New York, pp 43–74
Vanegas M, Croes RR (2007) Tourism, economic expansion and poverty reduction in Nicaragua: investing co-integration and causal relations. Department of Applied Economics, staff paper series #P07-10, University of Minnesota
Vogel H (2016) Travel industry economics: a guide for financial analysis. Springer, New York
Walton JK (1983) The English seaside resort: a social history 1750–1914. Leicester University Press, Leicester
Walton J (1997) The seaside resorts of England and Wales, 1900-1950. In: Shaw G, Williams A (eds) The rise and fall of the British coastal resorts: cultural and economic perspectives. Mansell, London, pp 21–48
Weiss T (2004) Tourism in America before world war II. J Econ Hist 64(2):289–327
Williams A, Shaw G (1997) Riding the big dipper: the rise and decline of the British seaside resorts in the twentieth century. In: Shaw G, Williams A (eds) The rise and fall of the British coastal resorts: cultural and economic perspectives. Mansell, London, pp 1–20
Wong KF, Song H, Chon KS (2006) Bayesian models for tourism demand forecasting. Tour Manag 27(5):773–780
World Tourism Organization (2017) UNWTO tourism highlights: 2017 edition https://www.e-unwto.org/doi/book/10.18111/9789284419029
World Tourism Organization (2018a) About the World Tourism Organization. http://www.world-tourism.org. Accessed 27 Jan 2018
World Tourism Organization (2018b) http://www.oecd.org/cfe/tourism/tourism-statistics.htm. Accessed 27 Jan 2018
Yazdi SK, Khanalizadeh B (2016) Tourism demand: A panel data approach. Curr Issue Tour 20(8):787–800

Part VIII

Technique and Measurement

Statistical Inference

Thomas Rahlf

Contents

Abstract

Statistical and, subsequently, econometric inferences have not undergone a cumulative, progressive process. We have seen instead the emergence of a number of different views, which have often been confused with each other in textbook literature on the subject. It therefore makes sense to approach the issue from a historical-scientific angle rather than a systematic one. We intend, using the extraordinarily complex development as a basis, to give a historical overview of the emergence of concepts that are of particular importance from the point of view of cliometrics. We shall start by describing the beginnings of modern probability theory, along with its connection with other statistical approaches.

T. Rahlf (✉)
German Research Foundation, Bonn, Germany
e-mail: thomas.rahlf@dfg.de

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_16

The following overview covers the basic principles of the current concepts of inference developed by R. A. Fisher on one hand and by J. Neyman and E. S. Pearson on the other. Neo-Bayesian approaches have meanwhile been developed in parallel, although they were not taken into account during the initial founding phase of econometrics. A "classic" approach was instead adopted in this respect, albeit with an additional difficulty: the taking into account of time. Cliometrics initially followed a Bayesian approach, but this did not finally prevail. Following on from econometrics, a correspondingly classic, inference-based position was adopted. This chapter concludes with a reference to a fundamental critique of the classic position by Rudolf Kalman, which we also find very promising as an inference-related concept for cliometrics. We often quote authors directly, in an effort to portray developments more vividly.

Keywords
Probability · Inference · Bayesianism · Frequentism · System theory

Introduction

Statistical inference possesses an ambivalence that is present in virtually no other field of science. Current doctrine is built up consistently on one hand (an impression furthermore reinforced an interdisciplinary examination of the relevant literature) across all disciplinary boundaries and along the same strictly schematic lines. The impression given is that it is a logically structured, self-contained edifice possessing universal validity. While the significance of individual methods can differ from subject to subject, their inherent statistical inference-related principles (with particular reference to the method of testing hypotheses and, more generally, assessment of the "evidence" supplied by statistical data) appear to be universally valid.

This was the objection expressed by Gerd Gigerenzer et al. (1989, p. 105f) regarding contradictory and illogical "hybridization":

> . . .scientific researchers in many fields learned to apply statistical tests in a quasi-mechanical way, without giving adequate attention to what questions these numerical procedures really answer.

A look "inside" the statistics gives a variegated impression. Certain quotations from the literature on statistics serve to illustrate the controversies within this area of science. Examples include R. A. Fisher (1956, p. 9), who stated, "The theory of inverse probability is founded upon an error, and must be wholly rejected." Von Mises (1951, p. 188) admitted, with reference to Fisher's "likelihood approach": "The many fine words that Fisher and his followers use to justify the likelihood theory are incomprehensible to me. The main argument [. . .] has nothing to say to me." A. Birnbaum, who brought up the likelihood concept in a widely read contribution to the likelihood principle as a fundamental basis of statistical inference, rejected the confidence principle developed by J. Neyman and Pearson on the grounds of its opposition to the likelihood principle (Birnbaum 1962; Neyman and

Pearson 1928a, b, 1933). He went on to reject the likelihood principle a few years later, however, precisely because of its opposition to the confidence principle.[1] Stegmüller (1973, p. 2) refers to Neyman, who had claimed that the test methods developed by Fisher "[. . .] were, *in a mathematically-definable sense*, 'worse than useless' [. . .]."[2] B. de Finetti (1981), one of the main representatives of a subjectivist theory of probability, was convinced that Fisher "[. . .] showed his feel for the necessity of a conclusion in Bayesian form (with the illusion of being able to express them with an indefinable 'fiducial probability'), with a desire to present the problem in a way that was opposed to the Bayesian approach (like Neyman, essentially)." L. J. Savage (1954), another important defender of a subjectivist approach, who wanted to incorporate into his influential book, *The Foundations of Statistics*, the conventional statistical inference methods developed by him as part of an axiomatic system of a subjectivist doctrine, wrote in the book's second edition: "Freud alone could explain how the rash and unfulfilled promise (made early in the first edition, to show how frequentist ideas can be justified by means of personalistic probabilities) went unamended through so many revisions of the manuscript."[3] O. Kempthorne (1971) finally characterized the various concepts of inference in a way that caused J. W. Pratt (1971 commentary, p. 496) to summarize Kempthorne's theses as follows: "Fiducial and structural methods are nonsense. Jeffrey's Bayesian and subjective Bayesian methods are nonsense. Likelihood methods are nonsense. He doesn't say directly that orthodox methods are nonsense, but he says it implicitly by his remarks [. . .]. In short, he says all methods are nonsense, therefore use orthodox methods." This list could easily be extended, but the impressions given should suffice.

Probability and Inference in Statistics

We would like to start by giving a broad-brush description of how the central concepts have developed.[4] The following milestones mark the most important steps along the way:

Historical milestones in the field of statistical inference

1700–1730	The first systematic definitions of the terms "probability" and "chance" (G. W. Leibniz, J. Bernoulli) and the attempt to arrive at statistical inference (as a conclusion) from probability theory (J. Bernoulli)
1750–1775	Inversion of the probability concept in connection with the error function of Laplace
	Inversion of the probability concept in connection with Bayesian binomial distribution

(*continued*)

[1] Cf. Birnbaum (1962, 1968, 1977).

[2] Original author's italics.

[3] Quoted from DuMouchel (1992, S. 527). The first edition was published in 1954. Cf. Savage (1954).

[4] A detailed treatment of the topic of this chapter can be found at Rahlf (1998) and Gigerenzer/Swijtink/Porter/Daston/Beatty/Krüger (1989).

Around 1810	Synthesis of error function and probability by P. S. Laplace and C. F. Gauss
1820–1840	The further development of statistical inference concepts (law of errors, law of large numbers) and their incorporation into the "social physics" of A. Quetelet
1870–1885	The incorporation of Quetelet's concepts into biology by F. Galton and the conceptual foundation of correlation and regression
1840–1870	Philosophical investigations into the concept of probability (as a parallel development)
1880–1895	The systematization and formalizing of statistical inference concepts by F. Y. Edgeworth and K. Pearson
1895–1900	The application of these systematized concepts of statistical inference to social science data and development into multiple regression by G. U. Yule
	Attempts by K. Pearson and G. U. Yule to clarify the concepts of correlation, spurious correlation, and causality
Around 1900	The concept of the significance test is developed by K. Pearson
1929/1930	The criteria of "good" valuations postulated by R. A. Fisher, a quantitative assessment of the quality of these valuations using the fiducial principle also developed by Fisher
1933	"Classic" test theory and confidence inference according to J. Neyman and E. S. Pearson
1926–1954	Subjectivist Bayesian approaches, such as those of F. P. Ramsey, B. de Finetti, H. Jeffreys, or L. J. Savage
1955	Objectivist Bayesian approaches, such as those of H. Robbins
1949/1962	The likelihood principle developed by Fisher and expanded by G. Barnard and A. Birnbaum

The theory of probability was regarded as something of a "brainteaser" until the middle of the seventeenth century, in the sense of pure combinatorics. The chance of rolling a certain dice number, of a tossed coin falling on one face or the other, or of drawing a particular card from the pack could be indicated without any profound philosophical consideration of the nature of probability. The probability of, for example, tossing a coin ten times and having it come up "heads" four times and "tails" six was to be determined by a combination of purely mathematical considerations, as a coin-tossing "experiment" could be based on a fully specified theoretical model: The events are mutually independent, thus making their sum binomial, the parameter being $\pi = 0.5$.

The questions that arose in a socioeconomic context at this time were, however, only apparently of the same nature. Even variables such as overall gender ratio, life expectancy, infant mortality rates, the proportion of the population available for military service, etc., were considered legitimate in this respect. But how was one to assess the *reliability* of the results obtained?

It was decisive that studies like those carried out by J. Graunt (1662 [1939]) of the register of deaths in London or by E. Halley (1693) of births and deaths in Breslau (present-day Wroclaw) attracted the attention of mathematicians such as G. W. Leibniz, J. Bernoulli, or A. de Moivre, thereby obliging those concerned to

consider the problem of inference. The proportion of people possessing a certain characteristic was unknown, as long as the characteristic concerned could not be computed for a given (sub)population. A possible entitlement to apply the binomial model existed, but there was definitely no theory capable of postulating a value for the parameter that was to be verified. Furthermore, this value could only be determined on the basis of the data and a measure indicated – by means of an interval – for the accuracy of the "estimate." An *inversion* of probability was therefore necessary, although neither Bernoulli nor de Moivre was able to complete this step. We follow S. Stigler at this point while assuming that the conceptual difficulties could be overcome only via the detour of the error function, ultimately by T. Bayes and P. S. Laplace. This "Copernican Revolution" in the development of theoretical statistics[5] was connected with the intention of Bernoulli. It is somewhat curious that this concept is nowadays associated with Bayes rather than Laplace. Bayes had a groundbreaking idea that was nevertheless developed at the same time and, presumably independently, by Laplace. However, Laplace had also constructed a systematic theory of probability that went on to form the basis for a number of applications over many years. The key statisticians (Gauss, Galton, and Edgeworth) subsequently followed a mainly Bayesian line of argument. The most probable parameter value for Gauss, for example, was the maximum of the likelihood function, since it emanated, as it also did for Laplace, from the principle of insufficient reason and thus from an a priori uniform distribution. K. Pearson meanwhile followed a (mostly) sampling-based approach however, and G. U. Yule worked within the same framework, albeit without attributing much importance, in general, to the question of inference.[6]

K. Pearson and G. U. Yule

The works of K. Pearson were of great significance for the further development of statistical inference. Pearson's first independent contribution to the field of statistics, which formed the basis of his subsequent fame, consisted of a system of frequency distributions included in two extensive papers published in the *Philosophical Transactions of the Royal Society* under the title *Contributions to the mathematical theory of evolution* (1894, 1895), which led to him being elected a fellow of the society. The question regarding the form of frequency distributions had been a fundamental issue since the end of the eighteenth century. There was a prevailing general belief that individual phenomena, which were homogeneous in the sense of many individually insignificant influencing factors, had to follow a normal distribution. Not everyone regarded normal distribution as being universally valid however, and collections of data that accumulated over the years implied a series of "skewed" distributions. Pearson above all regarded this fact as a challenge, and he eventually developed a

[5] Stigler (1986, p. 122).

[6] See, for example, Yule (1895, 1896a, b) and Pearson (1898).

"family" of curves, each based on four parameters, by which data could be assigned to different types of curve using their first four moments.

Pearson supplied not only the formulae but also a wealth of practical examples (distribution of air pressure, heights of schoolchildren, and sizes of crustaceans; statistics on poverty and divorce rates; etc.) and showed that these variables could be reconciled to a large extent by using his system. He went even further than Quetelet in this respect. It was not only data with a normal distribution that followed a uniform distribution law, without a need to isolate groups or major factors, but also many others whose distribution was in fact skewed but no less legitimate in this respect. If this were the case, the search for causative factors, as introduced by Galton as part of biology, was invalid:

> The law of frequency is based on the assumption of perfect ignorance of causes, but we rarely *are* perfectly ignorant, and where we have any knowledge it ought of course to be taken into account.[7]

The further application of Pearson to areas that are not necessarily closely subject to a law of constant distribution has been criticized[8]:

> [. . .] I see that there are many cases of 'skew' variation: but all cases which he has given, of variation with an unmistakably skew frequency, are taken from phenomena which are changing with a rapidity much greater than that of any organs in crabs, or such creatures. Pauperism, divorces, and the like, have only been invented, in their present form, for a short time, and as he himself shows, the maximum frequency changes its position at least in ten years.[9]

But the most important counterargument was that the numerous forms that could be adapted using Pearson's frequency curves lacked a theoretical foundation, as they were purely empirical constructs. If a frequency distribution did not lend itself to being represented by a normal distribution, the concept of causation based on a large number of random causes could not be effective. It is precisely this last point that was however, according to Stigler (1986, p. 339), likewise not the intention of Pearson, who was seen to represent a philosophy of science that had been guided by Kantian nominalism. On this basis, Pearson regarded frequency curves only as mental constructs that summarize empirical evidence, without providing any statements on possible causes. Pearson nevertheless also searched in this respect for a formal criterion for assessing deviation in the empirical distributions of his frequency curves and finally found one in the form of his chi-squared (χ^2) test, which he made public in 1900.

[7] F. Galton in a letter to K. Pearson of 18 Nov 1893, quoted by Stigler (1986, p. 336). Original author's italics.

[8] Despite criticism, Pearson's frequency curves soon became part of the standard repertoire of statistics.

[9] W. F. R. Weldon in a letter to F. Galton of 27 Jan 1895, quoted by Stigler (1986, p. 337).

Pearson made another important contribution to modern statistics in the field of correlation. He considered two variables with a normal bivariate distribution, deduced the correlation coefficient and a posteriori distribution[10] (on the basis of empirical standard deviations), and systematized the findings obtained to date. The theoretical derivation was followed by a series of applied examples, which he took from Galton. He did not admit any major possibilities regarding the application to social phenomena:

> Personally I ought to say that there is, in my own opinion, considerable danger in allying the methods of exact science to problems in descriptive science, whether they be problems of heredity or of political economy; the grace and logical accuracy of the mathematical process are apt to fascinate the descriptive scientist that he seeks for sociological hypotheses which fit his mathematical reasoning and this without first ascertaining whether the basis of his hypotheses is as broad as that human life to which the theory is to be applied.[11]

This move was finally made by Pearson's student G. U. Yule in a series of studies of Poor Law legislation. One important question in this respect was the extent to which the proportion of poor people in a given district was connected with its structure of care provision. Yule (1895, 1896b) found a "significant" link, which he nevertheless described as "suggestive," as the distributions of both variables were clearly shown to be skewed. In a subsequent step, he established a "regression line" between the two variables by minimizing the distances between this straight line and the data concerned. He perceived that this approach was easy to extend to higher dimensions, thereby leading to the "normal" system of equations that had been introduced by Gauss several decades earlier in the field of astronomy. From here, it was merely a technical matter, no longer requiring any conceptual step, to extend the approach to more than two variables.

Irrespective of the different views held by K. Pearson and Yule in this context regarding the concepts of correlation and causality, the general question surrounding all these considerations was the following: Did inference refer to a *population* or to *laws*? This was clear in Pearson's case and also, subsequently, in that of Fisher. The aim of studying biological data was to investigate conformity to natural laws. The situation was more difficult when it came to the investigations of socioeconomic data carried out by Yule or studies, such as those of Gosset, of the correlations between

[10]K. Pearson explicitly rejected the concept of inverse probability, although E. S. Pearson was of the view that he implicitly followed this approach on at least one occasion. Cf. Pearson (1898). "The basic of the approach used here is a little obscure and there seems to be implicit in it the classical concept of inverse probability" (Pearson 1967, p. 347), quoted by Dale (1991, p. 379). Pearson expressed himself most extensively on this issue in his paper *The fundamental problem of practical statistics* (1920), which has provoked different interpretations up to the present day. While Fisher (1922, p. 311), for example, believed he recognized a proof of Bayes' theorem in it, Dale (1991, p. 388) considered this as a "totally inaccurate observation." For further interpretations, cf. ibid., pp. 377-391. According to Stigler (1986, p. 345), Pearson worked on multiple occasions "[. . .] (implicitly) in a Bayesian framework."

[11]Pearson (1898, p. 1f), quoted by Stigler (1986, p. 304).

cancer rates and apple consumption, which included at least one exploratory element.[12] The interpretation of a correlation coefficient could only be hypothetical according to Yule, as it was normally possible to give a variety of alternative explanations whose distinction could not be provided by statistics. This problem would prove to be fundamental for statistical inference-based interpretations in the field of social science.

R. A. Fisher

The further development of statistical methodology in the field of biology has been characterized, since at least the time of Karl Pearson and R. A. Fisher, by the possibility of its application to the natural sciences. Fisher (1955, 1956, 1959) attempted to solve, by means of his *Design of Experiments*, the problems of inference-based conclusions in biology caused by their dependence on the conditions that prevail when taking samples.

Fisher's concept of inference was initially characterized by its explicit rejection, directed against Pearson in particular (in 1922), of inverse probability. This view was mainly due, in his opinion, to the confusing of theoretical parameters and estimates:

> It is this last confusion, in the writer's opinion, more than any other which has led to the survival of the present day of the fundamental paradox of inverse probability, which like an impenetrable jungle arrests progress towards precision of statistical concepts.[13]

He nevertheless developed a certain understanding at the same time:

> The criticisms (. . .) have done something towards banishing the method, at least from the elementary text-books of Algebra; but though we may agree wholly (. . .) that inverse probability is a mistake (perhaps the only mistake to which the mathematical world has so deeply committed itself), there yet remains the feeling that such a mistake would not have captivated the minds of Laplace and Poisson if there had been nothing in it but error.[14]

Although Fisher's concept of probability was frequentist, he vehemently rejected a definition of probability as a limit value applying to relative frequency in an unlimited number of repeated attempts (i.e., the von Mises definition subscribed to

[12]Cf. ibid., p. 373.

[13]Fisher (1922 [1992], p. 13), similar also to Fisher (1959, p. 34). There is in the case of Fisher (1956, p. 9) a (more or less) clear rejection of the Bayesian approach. He emphasized that he was "personally convinced" that "the theory of inverse probability is founded upon an error, and must be wholly rejected."

[14]Fisher (1922 [1992], p. 13). Ambiguities such as these are characteristic of Fisher's work. According to Geisser (1992, p. 4), Fisher subscribed – until at least 1912 – to approaches based on Bayesian logic. He then (p. 26f) explicitly rejected the validity of Bayes' theorem. Cf. Barnard (1988) regarding this question.

by most frequentists)[15]: "For Fisher, a probability is the fraction of a set, having no distinguishable subsets, that satisfies a given condition [. . .]."[16]

Fisher postulated that statistical inference should refer to theoretical, and thus fixed, parameters of hypothetically infinite populations, thereby determining the direction of research in the field of theoretical statistics for the following 50 years.[17] Otherwise, his concept of a statistical or "scientific" inference could not prevail. He used the term "inductive logic," not at least in order to set himself apart from the approach of his intellectual rival J. Neyman, who spoke of "inductive behavior."[18] It was possible, in cases where there was an indisputable a priori distribution, to speak of the probability of events, which were to be described as fiducial probabilities.[19] Intervals that express the uncertainty of an estimate were always to be construed as fiducial intervals.

The problem of the "significance test" is closely connected to the problem of using intervals to indicate the accuracy of an estimate. What we now understand as the logic of the significance test became increasingly important during the first two decades of the twentieth century.[20] It can largely be traced back to Fisher and has remained in force alongside the concept of the hypothesis test developed by Neyman and Pearson (see below). For Fisher, the level of significance of a test is a *measure of evidence*, which should neither be defined a priori nor regarded as unalterable, nor established as a guiding principle:

> A man who 'rejects' a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. This inequality statement can therefore be made. However the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all

[15]See supporting evidence in Savage (1976, p. 461). In Fisher (1959, p. 32), he emphasized, for example, that no probability of individual events could be established with such a definition.

[16]Savage (1976, p. 461) with corresponding supporting evidence. Savage observes in this respect: "Such a notion is hard to formulate mathematically, and indeed Fisher's concept of probability remained very unclear, which must have contributed to his isolation from many other statistical theorists" (p. 462).

[17]Cf. Geisser (1992). Partly ambiguous terms such as "mean," "standard deviation," or "correlation coefficient" have remained in use to this day to indicate, in various contexts, either theoretical variables or estimators for these theoretical variables.

[18]Cf. Savage (1976, S. 462) with supporting evidence.

[19]Ibid., p. 466: "Nobody knows just what they mean [. . .]. In a word, Fisher hopes by means of some process – the fiducial argument – to arrive at the equivalent of posterior distributions in a Bayesian argument without the introduction of prior distributions [. . .]." We would like to join in with this criticism. As observed by Menges (1972, p. 275): "The fiducial concept considers the results of an observation as indisputable fact in this respect, and as the basis on which to build inference. *It can thus do justice, in principle, to the historical character of social phenomena*" (original author's italics), although this also applies to Bayesian logic in our opinion.

[20]Such as Pearson's chi-squared goodness-of-fit test of 1900, Student's *t*-test, developed in 1908 and formalized by Fisher, or the *F*-test applied to the analysis of variance by Fisher.

> circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.[21]

This criticism was directed against a concept that had been propagated by J. Neyman and E. S. Pearson since the 1930s and which had quickly become the dominant view.

J. Neyman and E. S. Pearson

The works of J. Neyman and E. S. Pearson are likewise unanimously considered to be milestones in the history of theoretical statistics. While Fisher wished to allow, in relation to the testing of hypotheses, only the alternatives "rejection" and "no statement possible," Neyman and Pearson developed a closed test theory which introduced differentiated levels of rejection and acceptance, along with concepts such as the "power" of a test, Type I and Type II errors, and "uniformly most powerful test." Until the end of the nineteenth century, the testing of hypotheses was based on distributions of test statistics which were (1) best suited for use with large samples and (2) employed for intuitive reasons. The introduction of the *t*-distribution by W. S. Gosset (1908) and the contributions of R. A. Fisher, who differentiated the exact distributions of t, χ^2, F, and certain correlation coefficients in normal distributions, meant that at least problem (1) could be overcome. With this problem solved, the question then posed was that of a formally satisfactory test *theory*. E. S. Pearson stated in a review that the idea for this theory came to him via an observation made by Gosset:

> I had been trying to discover some principle beyond that of practical expediency which would justify the use of "Student's" ratio $z = (-m)/s$ in testing the hypothesis that the mean of the sample population was at m. Gosset's reply (to the letter in which Pearson [...] had raised the question) had a tremendous influence on the direction of my subsequent work, for the first paragraph contains the germ of that idea which has formed the basis of all the later joint researches of Neyman and myself. It is the simple suggestion that the only valid reason for rejecting a statistical hypothesis is that some alternative hypothesis explains the observed events with a greater degree of probability.[22]

Gosset argued in this letter that not even a probability value as small as 0.0001 could lead per se to rejection of a hypothesis for a random sample. Only comparison with an *alternative* hypothesis, "which will explain the occurrence of the sample with a more reasonable probability, say 0.05 (such as that it belongs to a different

[21] Fisher (1959, p. 41f). Fisher's failure to include tables of *p*-values in his famous textbook *Statistical Methods for Research Workers* (rather than the tables of significance values that he *did* include) arose from the fact that K. Pearson held the copyright to the former. Cf. Watson (1983, p. 714).

[22] Pearson in a paper from 1939 quoted from Lehmann's comments (1992, p. 68) on Neyman/Pearson (1933) (our italics).

population or that the sample wasn't random or whatever will do the trick) you will be very much more inclined to consider that the original hypothesis is not true."[23]

This idea was then jointly developed by Neyman and Pearson (1928a, b) in an extensive two-part paper, published in *Biometrika*, on the concept of the likelihood ratio test. While Pearson now saw in this the uniform method for which they had been seeking, Neyman was clearly still not satisfied:

> It seemed to him that the likelihood ratio principle itself was somewhat ad hoc and was lacking a fully logical basis. His search for a firmer foundation, which constitutes the third of the three steps, eventually led him to a new formulation: The most desirable test would be obtained by maximizing the power of the test, subject to the condition that under the hypothesis, the rejection probability has a preassigned value, the level of a test.[24]

The result was Neyman and Pearson (1933), which also includes the famous Neyman-Pearson lemma. This states that in the class of all tests with probability α, the criterion function of the likelihood ratio test dominates the criterion function of any other test (i.e., every other test has a greater probability of including a Type II error). Neyman and Pearson used a series of examples to demonstrate the application of this principle and thus laid the foundation for a widely recognized general test theory which today continues to be regarded as "classic," along with the "confidence interval" likewise formulated by Neyman (1937). The method based on Neyman-Pearson logic can be described, after Lehmann, in terms of four steps[25]:

1. Specification of a model using a parametric family of distributions which has produced the data
2. Specification of a hypothesis with regard to a parameter of interest, H_0: $\theta = \theta_0$, and one simple or one class of alternatives H_1, e.g., $\theta \leq \theta_0$
3. Specification of a level of significance α, indicating the maximum allowable probability of a Type I error
4. Selection of the optimum method for testing H_0 against H_1 by minimizing the ß-error[26]

Lehmann finally added a – quite fundamental – fifth item, but this is more of a prerequisite than a procedure:

1. All (four) steps must be completed "before any observations have been seen."

[23]Ibid.

[24]Lehmann (1992, p. 68). This highly important aspect of the Neyman-Pearson theory is often not taken into account. As Borovcnik (1992, p. 92) rightly points out, "[. . .] a frequency interpretation places too much emphasis on the α-error during testing, while the real trick with this method is to minimise the ß-error."

[25]According to Lehmann (1992, p. 69f).

[26]We do not intend to go into the corresponding techniques here but refer instead to textbook literature on the subject.

The approach postulated by Neyman and Pearson actually amounted only to a set of *guidelines*. The two authors expressed, as follows, the conviction that lay behind their theory:

> Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.[27]

Inference statements are therefore hypothetical-deductive and only possible *before* events occur. They therefore do not refer to specific hypotheses but to future *actions* in the long term. This approach was consequently extended by A. Wald (1950) to form a pure decision theory, with Neyman repeatedly emphasizing this behavior theory aspect in his later work.

There was however vehement criticism from no less a person than R. A. Fisher, who might have wished to recognize the Neyman-Pearson theory for situations where permanent decisions had to be taken but was in no way willing to accept statistical inference-based assessments in a *scientific* sense. A further argument concerned the claim of "repeated sampling *from the same population*." Fisher pointed out, following on from J. Venn, that a given sample could always have resulted from a variety of conceivable populations: "so [. . .] the phrase 'repeated sampling from the same population' does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination."[28]

These approaches were met with further reservations: On one hand, models would mostly be chosen *in practice* on the basis of data while often examining not just one but several hypotheses using the same data. In many situations, the eventual reduction of inference to a yes/no decision was not appropriate.

It has furthermore been demonstrated that optimum (i.e., uniformly most powerful) tests exist only for limited situations or are so complex (when maximizing their minimum power) that their application presents considerable problems. It should however be emphasized that these reservations are the exception and that an overwhelming majority has, particularly in the field of applied statistics, unconditionally accepted the Neyman-Pearson approach, which has become something of a paradigm, even though today's statisticians continue to argue about where the precise differences between this approach and Fisher's test concept lie.[29]

If we compare this approach to that of Fisher, point 5 (see above) becomes particularly decisive. The method according to Neyman and Pearson is therefore strictly deductive, while Fisher's approach is (also, at least) inductive, with assessment taking place only after obtaining evidence based on data and above all without

[27]Neyman/Pearson (1933 [1992], p. 74). Kyburg (1985, p. 119) sums up their intention in the observation: "That says nothing about the case before us, but it may make us feel better."

[28]Fisher (1955, S. 71).

[29]Cf. Lehmann (1993), for example.

considering an alternative hypothesis. Neyman and Pearson surely did not intend to promote a universal and constant level of significance but rather only in this sense: Even if they allow for different levels in different situations, these must be determined *before* the experiment and/or *before* obtaining any knowledge of the data evidence. The second fundamental difference lies in the *direction* of the inference. Fisher's test concept – and, in this respect, K. Pearson's logically equivalent significance test concept – applies to a state that exists or which, strictly speaking, may already have passed. The inference of Neyman and Pearson, on the other hand, applies to the future: If we act in one way or another in the future on the basis of the test, how often are we then likely to commit an error? The current practice is in fact to apply a blending of both concepts.[30]

Statistical inference was now reduced to the creation of guidelines for conduct in the long term. No contentious epistemological issues were settled using the Neyman-Pearson theory; it related only to a clear statement. Its success can perhaps also be explained by the fact that other positions (K. Pearson, Fisher) lacked such clarity.

The dominant approach since then has in any case been a supposedly objective, frequency-theory, and inference-based position, although a modern, Bayesian, statistical inference has continued to develop in parallel. It is remarkable that modern, subjectivist probability theory was not established by social scientists, who regarded as problematic its individual prerequisites or implications with regard to long-term experimental inference, but – without exception – by mathematicians (Ramsey, de Finetti, Savage) or geophysicists (Jeffreys) who saw problems of logic in the predominant frequency theory-based approaches.

This development took place in three stages: the reestablishing, by F. P. Ramsey, B. de Finetti, H. Jeffreys, and L. J. Savage, of Bayesian probability theories; the expanding, by G. A. Barnard and especially A. Birnbaum, of various likelihood-based approaches to form a likelihood *principle*; and finally the combining of these two components to create a modern Bayesian inference, which has come to exist in numerous forms. The following section considers at first the development of subjectivist probability theories.

Bayesian Probability

These are based on the following three basic assumptions, according to Howson (1995, p. 2):

[30]Johnstone (1986, p. 6) aptly describes the prevailing approach: "In general, tests of significance in practice follow Neyman formally, but Fisher philosophically. Formally, there is mention of 'alternative' hypotheses, errors 'of the second kind', and the 'power' of the test, which are terms due to Neyman (and his colleague Pearson). But philosophically, the result in a test, e.g. the result that the level of significance P equals 0.049, or that P is less than or equal to 5%, is interpreted as a measure of evidence, which is the interpretation following Fisher, and denied repeatedly by Neyman."

1. A hypothesis A is, in extreme cases, certainly true or certainly false. Intermediate degrees of belief in A are permitted.
2. These degrees of belief can be expressed numerically.
3. If they are rational and measured against the closed unit interval, they satisfy the finite additivity axioms.

The subjectivist Bayesian concepts of F. P. Ramsey, B. de Finetti, H. Jeffreys, and L. J. Savage were developed successively but independently of each other. We will now deal with them briefly in chronological order.

The first "modern" subjectivist probability theory was established by F. P. Ramsey in papers written in 1926 and 1928 but published posthumously in 1931.[31] As we have seen, the epistemological conception of probability from Bernoulli to Laplace was subjective as well as in the case of Gauss, Galton, and Edgeworth: "Probability" was interpreted by C. Huygens in terms of betting odds, with chance defined as ignorance. The principle of insufficient reason implied an a priori uniform distribution, which was linked, via Bayes' theorem, to the evidence from data in form of an a posteriori probability for a given parameter value.

Ramsey argued along similar lines, albeit combined with a critique of the logical and frequency theory-based interpretation. His starting point was John Maynard Keynes' *Treatise on Probability* (1921). For Keynes, probability meant a logical relationship between two different sets of propositions that are interconnected via a "degree of belief":

> Let our premises consist of any set of propositions h and our conclusion consist of any set of propositions a, then if a knowledge of h justifies a rational degree of belief in a of degree A, we may say that there is a probability-relation of degree A between a and h.[32]

Keynes did not however require all degrees of belief to be numerically measurable or comparable, thereby avoiding major difficulties. Ramsey postulated instead that probabilities should be expressed as betting odds, which must be rational (i.e., consistent and coherent). Ramsey's observations were of a purely philosophical nature and did not constitute a concept of *inference*. This was supplied in a famous paper by Bruno de Finetti (1937). It was totally clear to Finetti that the basis of all probability was subjective in nature.[33] Bayes' theorem was of central importance in this respect: Subjective assessments/probabilities must be revised constantly in the light of Bayes' theorem on the basis of data and knowledge obtained. This meant that subjectivist probabilities converge to relative frequencies as evidence accumulates. De Finetti did not criticize classical statistics for false results but for its false foundations:

[31] Ramsey (1931a, b).

[32] Keynes (1921, S. 4), quoted by Kyburg/Smokler (1964, p. 9).

[33] Cf. de Finetti (1937).

> The overwhelming majority of modern statistics are in practice completely normal, but their foundations are false. Intuition has however prevented statisticians from making mistakes. My thesis is that the Bayesian method justifies what they have always done, and that they are developing new methods which are missing in the orthodox approach.[34]

Harold Jeffreys (1939) argued along similar lines. He combined a probability theory with a theory of induction. Jeffreys stressed (like de Finetti) that a fundamental problem of science lay in learning from experience:

> Knowledge obtained in this way is partly merely description of what we have already observed, but partly consists of making inferences from past experience to predict future experience. This part may be called generalization or induction. It is the most important part; events that are merely described and have no apparent relation to others may as well be forgotten, and in fact usually are.[35]

It therefore follows that probability is not a frequency but a "reasonable degree of belief, which satisfies certain rules of consistency and can in consequence of these rules be formally expressed by numbers."[36] If an explanation is given for an observed event, a researcher might determine that it is "probably true." It is thus implied that he has a high degree of confidence in a hypothesis, which is in turn (1) quantifiable and (2) based on experience and information.[37] A rule now states how the cognitive process should operate: This is none other than Bayes' theorem. In every probability to which we assign a hypothesis, that hypothesis is conditioned by the information available to us. If this changes (increasingly), the probability associated with the hypothesis must be revised accordingly. This approach is what constitutes the basis of learning from experience, which is formalized using Bayes' theorem: A posteriori probabilities result from the evaluation of a priori probability with the data evidence, using the likelihood function.

L. J. Savage was another important forerunner of modern Bayesian probability theory. Savage, who was influenced mainly by Milton Friedman and John von Neumann, formulated his concept of probability in the late 1940s/early 1950s, on the basis of a utility theory. The year 1954 saw the publication of his seminal work *The Foundations of Statistics*, in which he tried to arrange within a unified framework the (in his view) rather loosely connected set of techniques developed by R. A. Fisher and J. Neyman/E. S. Pearson, intended to be based on a theory of decision-making under uncertainty. However, an examination of the details showed that the venture was doomed to failure. H. E. Robbins (1955) took a different path. He postulated probabilities that were "objective" and a priori rather than epistemic. He started with the question as to whether one could apply the Bayesian approach even if the a priori probability of a parameter is unknown but nevertheless "exists." This

[34] De Finetti (1981, p. 657).

[35] Jeffreys (1939, p. 8).

[36] Ibid., p. 401.

[37] Although the hypothesis can still be false in terms of rule 4.

supposition of an objectively existing a priori probability is not shared by most Bayesians however nor is it, in a positive sense, required.

Bayesian Inference

We have, in the case of the Bayesian works cited above, placed the issue of probability in the foreground. But there is a second Bayesian inference: the likelihood element. Approaches to likelihood initially emerged independently of Bayesian concepts. The likelihood ideas created by Fisher were further developed mainly by G. A. Barnard.[38] These ideas were given a basic theoretical foundation by the pioneering work of A. Birnbaum, who developed them into a likelihood *principle* (LP).[39] By this time, the field of statistics was already being dominated by the Neyman-Pearson approach and its decision theory-based further development by A. Wald (1950).

The likelihood principle had radical consequences. It stated that all the evidence from data was contained in the likelihood function. This made the sample space irrelevant, *after* the data had been obtained. It means that measures of evidence referring to the space of all possible data (i.e., the probability or parameter space), such as p-values or the confidence level, are irrelevant to inference *after* a given piece of data has been created. This was otherwise a rejection of the frequentist position, without having to resort to Bayesian arguments.

Let us now turn to the linking of a priori probabilities and likelihood inference to Bayesian inference. The Bayesian breakthrough eventually succeeded, in practical terms, with a paper by W. Edwards, H. Lindman, and L. J. Savage (1963), which finally made the corresponding approaches available to a wider public.[40]

Edwards, Lindman, and Savage dealt with the main reservations affecting the Bayesian approach, such as how scientific objectivity could be possible if different scientists held different a priori views, thereby creating different a priori probabilities (and probability distributions).[41] They did not bring in the argument proposed by Laplace and Edgeworth[42] (whereby an increase in the range of data causes the influence of a priori distribution to diminish progressively, before eventually disappearing altogether) but opted rather for the question as to whether an a priori distribution can be assumed to be uniform or whether the exact form of the a priori distribution is of no great importance to a posteriori distribution. They showed that "it suffices that your actual prior density change gently in the region favored by the

[38]Barnard (1947, 1949). For historical development, see Berger/Wolpert (1988, p. 22ff).

[39]Birnbaum (1962). Cf. also Bjornstad (1992) on the following. A "standard" work on the subject is that of Berger/Wolpert (1988).

[40]Edwards/Lindman/Savage (1963 [1992]). Our intention from here on is to deal only with certain ideas without going into technical detail.

[41]Ibid., pp. 534–540.

[42]For example, Laplace (1812) and Edgeworth (1884).

data and not itself too strongly favor some other region."[43] These vague indications were then given a mathematical form, thereby showing that such an approach is indeed justified under somewhat weak assumptions.[44]

The authors did however acknowledge, on the other hand, that there are also situations where the exact characteristics of a priori distribution are decisive.[45]

The following includes a section on "Bayesian hypothesis testing." If an alternative to the prevailing classical statistics was to be provided (and this was their claim), this would also have to include such a central aspect as the testing of scientific hypotheses.[46] They started by clarifying the terms "odds" and "likelihood ratios." Using the example of checking to see if a dice is "fair," the application of likelihood ratios in a Bayesian sense was then compared to the classic approach of Neyman/Pearson (see above). They paid particular attention to clarifying the problem whereby classical statistics favored a consideration of Type I and Type II errors on the basis of this test variable:

> The interesting point is made that a Bayesian hypothesis test can add extensive support to the null hypothesis whenever the likelihood ratio is large. The classical test can only reject hypotheses, and it is not clear just what sort of evidence classical statistics would regard as a strong confirmation of a null hypothesis.[47]

We would like to avoid going into the – mostly highly technical – details in this respect. Solutions have meanwhile been found for numerous individual problems and fundamental questions, such as the Bayesian interpretation of frequency theory-based points of view, purely empirical Bayesian approaches, or even a theory of Bayesian data analysis.

One important issue in this context is the assessment of significance tests and confidence intervals.[48] The use of significance tests in their frequency theory-based sense enjoys wide support from a number of Bayesians for use as a heuristic tool, while others reject this approach. If a priori information is lacking, the confidence intervals of classical statistics and the Bayesian probability intervals may be almost numerically identical. They should, however, be interpreted in totally different ways.[49] In the classic, frequentist interpretation, a confidence interval of 95% means that, with the indicated (identical) sample ranges n for $m \to \infty$ (where m is the number of samples), 95% of intervals cover the true, unknown, fixed parameter

[43]Ibid., p. 541. This is referred to as "stable estimation."

[44]DuMouchel (1992, p. 521) points out that this approach is closely related to the "reference priors" subsequently proposed by other Bayesians for use in situations where little a priori information is available, which are also acceptable to classical statisticians.

[45]Edwards/Lindman/Savage (1963 [1992], p. 546).

[46]Bayesian literature does not adopt a uniform position regarding the need for a test theory.

[47]DuMouchel (1992, p. 523). Cf. example no. 3 in appendix A3 and also example no. 2 in appendix A4.

[48]General reference is made to Hodges (1990) in this respect.

[49]The following according to Iversen (1984, p. 31).

and 5% do not. We do not know however (and can only hope) whether the specific interval concerned covers the parameter or not. A Bayesian analysis assumes, in contrast, that the unknown parameter has an (usually subjective) a priori distribution. There is still uncertainty after the data have been obtained, but less so than in the previous case. This uncertainty is still expressed in probabilities but with a wholly different interpretation: The parameter θ lies, with a probability of 95%, between the two values c_u and c_o. Such an interpretation is not possible in terms of classical statistical inference,[50] although misleading interpretations of this Bayesian epistemology can still be found to this day in classic literature on the subject.

The alternative definition of the concept of probability is fundamental, regardless of individual formulations. In order to highlight better the contrast with the classic approach, we should first turn to the classic concept of probability and its weaknesses.

W. Stegmüller counts eight objections, put forward in literature on the subject, to the frequency theory arising from von Mises' definition,[51] regarding at least the last of them as "deadly": He confuses practical certainty with logical necessity.[52] A particular weakness of this concept of probability was seen to lie in its rejection of individual probabilities. According to von Mises' definition, it was impossible, for example, to indicate the probability of a certain throw of a particular dice at a particular location.

K. R. Popper (1990), for example, one of the most vehement opponents of subjectivism, used this problem to develop his own concept of probability (mainly related to the problems of physics) which evolved over the years into a so-called propensity theory.

No agreement has been reached up to the present day (nor is such a clarification likely to be achieved in the near future) about the final definition of probability, as, for example, C. Howson established:

> It would be foolhardy to predict that philosophical probability has entered a final stable phase; surveys of the field tend to have useful lifetimes of a decade or so, at most two. It would also probably be incorrect to pretend that there is likely in the near future to be any settled consensus as to which interpretations of probability make viable and useful theories, and which are dead ends.[53]

Bayesian concepts of inference are however not limited to a subjective element that formalizes a priori probability but link it, by means of Bayes' theorem, with the "evidence of the data," which is in turn formalized in the likelihood function. The likelihood function already played an important role for Bernoulli, Laplace, and Gauss. Its importance as a central element of statistical inference was emphasized by

[50]Ibid: "This is the way many users of confidence intervals want to interpret a confidence interval, but in classical statistical inference such an interpretation is not possible."

[51]See above, p. 86f.

[52]Cf. Stegmüller (1973, p. 32ff, particularly p. 37).

[53]Howson (1995, p. 27).

A. Birnbaum in particular, who introduced the concept of the likelihood principle in this context.[54] The main difference between the likelihood principle and the frequency principle can be formulated as a question: Is it possible to obtain evidence about a parameter on the basis of a specific piece of data (i.e., a "sample")? Adherents of the frequency concept (particularly J. Neyman) emphasize that we can only assess the performance of a procedure if it is carried out repeatedly and measured on the basis of long-term averages.

However, if it is not possible to conduct experiments, and conclusions can only be drawn using existing, repeatable data that have not been scrutinized (e.g., as is the case in cliometrics), the relevance of such a concept must be seriously questioned. If repeatability is purely hypothetical, it should also be explicitly defined as a (subjective) conviction and not as an objective possibility. We therefore find it more reasonable, for such situations, to define probability as a degree of belief, which is then assigned to a parameter value. The evaluation and revision of this conviction with the evidence of existing, non-hypothetical data obtained by applying the likelihood function are also logically consistent in our opinion, especially as it does not depend on asymptotic generalizations. We would like to subscribe to the opinion of D. Lindley in this respect:

> The present position in statistical inference is historically interesting. The bulk of practitioners use well-established methods like least squares, analysis of variance, maximum likelihood and significance tests: all broadly within the Fisherian school and chosen for their proven usefulness rather than their logical coherence. If asked about their rigorous justification most of these people would refer to ideas of the NPW [Neyman-Pearson-Wald, T. R.] type; least-square estimates are best, linear unbiased; F-tests have high power and maximum likelihood values are asymptotically optimal. Yet these justifications are far from satisfactory: the only logically coherent system is the Bayesian one which disagrees with the NPW notions, largely because of their violation of the likelihood principle.[55]

Inference in Econometrics

Let us now turn to inference in econometrics. Two phases can be distinguished in economic statistics and econometrics: an initial phase, in which the description and exploration of economic series or processes predominated, and a second phase of inference and modeling.

The first phase can be characterized by its adoption of correlation concepts developed by Galton (1888) and Pearson. There was, however, a crucial difference: A body of theory did in fact exist in economics, but it was neither uniform nor

[54] See above, p. 99.

[55] Lindley (1991, p. 493).

sufficiently established to make it accessible for direct empirical application.[56] An explorative character was therefore dominant from the beginning in this respect. Phenomena such as "trade cycles" were not physical variables that only had to be measured, nor were they biological variables with distribution that could be determined with arbitrary precision and influencing factors that could be analyzed by experiment. On the contrary, the data were (1) passive in nature and not immediately suitable for reproducing, they had to be (2) precisely defined, and they were not (3) subject to universally stable distribution.

The use of the correlation calculation was theoretically based in the case of Galton. As the observed data came, for example, from a bivariate normal distribution, their relationship to each other could be expressed in a coefficient. But this theoretical reasoning was already abandoned by Yule upon its first application in the context of social science.[57] The functional relationships were considered linear for computational processing reasons, while the parameters were determined, on the same grounds, by means of the method of least squares. Yule's authority (he was one of the leading statisticians of his day) justified the application of biometric techniques, even though the theoretical justification for this approach was doubtful.

Two aspects are of particular significance in this context: Firstly, no in-depth statistical knowledge was needed in order to recognize that the structure of socio-economic phenomena was different to the structure used to determine the growth of plants or relationships between organism body sizes. Secondly, this was made all the more clear as attention turned to the analysis of data that represented *time series*.

The Time Dimension

The analysis of economic events in terms of their processuality did not find, in economic theory, any concrete statements regarding duration, form, or relationships of trade cycles to each other. The pioneers of empirical studies thus went their own ways, with H. L. Moore and W. S. Jevons seeking replacement in the field of astronomy. Not only astronomical phenomena, such as the periodically varying number of sunspots or the strictly periodic path of Venus (an 8-year cycle between the Sun and Earth), were used to provide explanations; the mechanics of astronomy, in the form of periodogram analysis, were also employed. A method such as this had the advantage of being able to make "hidden" periodicities visible. However, the initial euphoria created by the use of the periodogram analysis soon gave way to the sobering realization that the application lacked an important prerequisite: the stability of the object being examined. Trade cycles were not like the planets, with their constant movements of a duration that could be computed with fixed margins of

[56]Economic theories, from L. Walras to A. Marshall, started out from states of equilibrium, which were adapted, independently of historical context, by the same perpetual motives of human action. The economic laws contained in these theories were timeless.

[57]See above, p. 76 f.

error, but were instead phenomena whose length and intensity varied both with time and the intensity of their disturbance factors.

And even this was not enough, as economic data generally tended to be subject to trends. Their long-term development was therefore not distributed on the basis of stable averages. In these cases, there were no timeless states of equilibrium from which (at the most) transient deviations were possible. There was instead an irreversible development.

The solution to this problem did not however lie in using this irreversibility as an opportunity to adopt a fundamentally different view. Instead, two alternatives were taken up: one postulated, even for this long-term development, either a functional, measurement-error-conditioned context in the form of a polynomial or some other trend function (if the long-term curve had a reasonably smooth appearance). The method of least squares was used to determine this trend. This had already developed a life of its own, and its progress was barely stoppable. Either that or one could decide completely against a long-term development model and exclude it by observing the deviations from a moving average. In both cases however, the goal was not a comprehensive analysis of (historical) development, but rather an "exclusion" of whatever could not be incorporated into the scheme of identical timeless structures.[58]

It is to this extent obvious that a component-based concept dominated further research. Mutually independent explanatory factors therefore determined the long-, medium-, and short-term curves by which the trend component was found to be just as disruptive as its short-term "residual" counterpart. It was difficult in this context to respond to the question of correlation. The study of trends and cycles on one hand and of correlations on the other was not a separate epistemic interest but an interrelated factor. According to the statisticians, the trend first had to be excluded to allow the examination of correlations, while the goal of correlation analysis was to examine the conformity of medium-term (i.e., cyclic) curves.

On the other hand, one must however not overlook the fact that it was in this formulation phase that the issue of historical change in economic structures became highly problematic. If there was a long-term trend "component," why should the mutual links between economic variables not then also be made subject to long-term changes? The attempts by Hall (1925), Kuznets (1928a, b), Ezekiel (1928), or Frisch (1931) to extend existing concepts to include time-dependent models, or at least to point out the inadequacy of conventional formalizations, were therefore the obvious thing to do.

We can only speculate as to why this path was not pursued further. One possible explanation might be that the technical difficulties with regard to modeling were too great. However, as these papers were in any case barely implicated in statistical inference, another explanation seems plausible to us: the surprisingly great similarity between an economic index on the trade cycle and a series of computed random

[58]One of the few exceptions, who assigned independent significance to the trend, was S. Kuznets. See Kuznets (1930a, b) in particular.

variables, contained above all in a paper by Slutzky (1937) and presented shortly afterward to the English-speaking world by Kuznets (1929). Did this similarity mean that even trade cycles depended solely on random variables?

Research by Yule and Slutzky went on to form the conceptual basis of the modern theory of stochastic processes. Although both of them described different types of models – autoregressive processes in the case of Yule (1927) and so-called "moving average" processes in the case of Slutzky (1937) – their structures nevertheless had crucial factors in common. They regarded a time series as a realization of a stochastic differential equation. While Yule started with a trigonometric function that could be represented as a differential equation (albeit one in which the error term had a completely different effect to that of the functional form), Slutzky constructed various – at first glance rather arbitrary – sums of random variables. A deeper justification for the chosen type of model (e.g., regarding why a certain number of random variables was provided with different weightings and added up once or several times) was of less importance in this respect than the alarming fact that random variables could create cyclical phenomena.

It is highly surprising that there were apparently, in the case of this conceptualization, bigger problems regarding the acceptance of the idea of a random, yet legitimate, process than there were for cross-sectional regression analysis. Time series were therefore regarded either deterministically, in terms of their essential components (the component model), or as purely coincidental, with cycles which then had no significance. The key point was overlooked: It was not the random variables that were responsible for (pseudo-)cyclical character but the mechanism, i.e., the model.

This inner logic of these models remained hidden to Kuznets, just as it subsequently did to G. Tintner, J. Schumpeter, and John Maynard Keynes.[59] It is therefore not surprising that scientists with less of a mathematical background were no longer willing or able to follow the conceptual idea associated with such models.

R. Frisch (1933) on the other hand, an econometrician with physical background, had clearly recognized the inner logic of these models and had even included a corresponding economic justification of it in his famous article on propagation. In a dynamic model of an economy, certain parameter values not affected by disturbance factors could give rise to damped oscillations. The action of "shocks" could, on the other hand, produce the irregular cycles first referred to by Yule.[60]

[59]Even Tinbergen came to recognize that he "did not understand the role of the shocks as well as Frisch did" (Tinbergen in Magnus/Morgan (1987, p. 125)).

[60]The separation between the role of the mechanism and that of the shock was of great importance for the development of econometrics, even though Tinbergen regarded it critically in retrospect: "[. . .] I think that what interested economics most was not the shocks but the mechanism generating endogenous cycles, and it might very well be that we have overestimated the role of the mechanism. Maybe the shocks were really much more important. This problem was never solved, because the War came along and after the War we were not interested in business cycles anymore" (Tinbergen in Magnus/Morgan (1987, p. 125)).

With the reception of these models into economics, the ways divide. Kuznets' (1934) *Time Series* contribution to the *Encyclopedia of the Social Sciences* described only the component model, without any stochastic implications. No mention was made of models with variable parameters or of the fundamental significance of the models of Yule and Slutzky.[61] Papers by Schumpeter, and also by Burns and Mitchell (1946), took a similar line. Schumpeter did in fact write the opening article of the first issue of *Econometrica*, which was published in 1933, but played no further part in the development of econometrics.

It was of crucial importance to further development that the scientific orientation of econometrics was largely determined by individuals with an educational background in physics, such as Jan Tinbergen, Ragnar Frisch, Tjalling Koopmans, Charles Roos, or Harold T. Davis.[62] These thinkers possessed a different picture of economics to that of "traditional" empirical researchers. They brought a mechanistic, rigorously mathematical model of thinking to empirical research. One example of this development is an account by Koopmans of his career:

> Why did I leave physics at the end of 1933? In the depth of the worldwide economic depression, I felt that the physical sciences were far ahead of the social and economic sciences. What had held me back was the completely different, most verbal, and to me almost indigestible style of writing in the social sciences. Then I learned from a friend that there was a field called mathematical economics, and that Jan Tinbergen, a former student of Paul Ehrenfest, had left physics to devote himself to economics. Tinbergen received me cordially and guided me into the field in his own inimitable way. I moved to Amsterdam, which had a faculty of economics. The transition was not easy. I found that I benefited more from sitting in and listening to discussions of problems of economic policy than from reading the tomes. Also, because of my reading block, I chose problems that, by their nature, or because of the mathematical tools required, have similarity to physics.[63]

It was possible to have in this environment (1) modeling of the economic world in the form of differential equations and (2) a rigid stochastic process. It nevertheless appears strange, at first glance, that Koopmans should develop his approach using the theory of R. A. Fisher and did not see, as Frisch had, measurement errors, in physical analogy, as a justification for a stochastic approach but started out instead, in a biological analogy, from hypothetically infinite populations from which, with constant probabilities, the existing data would have stemmed. The basic stochastic concept was probably not of so much importance in this instance but rather the facts

[61] Cf. Kuznets (1934).

[62] See Epstein (1987, p. 75 note 39), Mirowski (1989, p. 234), and above all Boumans (1993). Even the statistician G. U. Yule, who was particularly involved in research in the field of time series analysis and its potential applications in economics, began his academic career in the study of electrical waves.

[63] Quoted from Mirowski (1991, p. 152). Frisch and Koopmans applied matrix calculus, which was being widely disseminated in physics in the mid-1920s, in the context of multiple regression analysis, to the field of econometrics, thereby making it more difficult for economists to comprehend the texts concerned. Cf. Mirowski (1989, p. 231).

that Fisher had developed a comprehensive statistical estimation theory and that he was regarded as a leading statistician.

Univariate time series analysis turned into a sideshow issue in this context, with thinking in terms of "complete" models coming to dominate instead.[64] These models did not however fully or consistently match, from the beginning, the theoretical economic models, although their consideration was the initial objective of econometrics. Tinbergen had already found himself forced into a series of compromises, as the existing economic theories of his day had not been specified to an extent that permitted direct empirical testing.

The uninhibited, iterative approach of Tinbergen infringed the rules of the stochastic concept of statistics that had just been adopted by Frisch and Koopmans. Some criticism of Keynes or Friedman was to this extent justified. The chosen way was nevertheless followed further and given a certain manifesto-like air by T. Haavelmo, a student of R. Frisch.

"Clarification": Trygve Haavelmo

Haavelmo's line of argument, which set the trend for further development, called – like Koopmans' – for a rigorously stochastic approach. Unlike Koopmans however, Haavelmo did not rely on Fisher's theory but on those of Neyman and Pearson. If we examine the foundations of this theory, its application to (macro)economic developments inevitably appears problematic.

We have seen that acceptance of the Neyman-Pearson approach brings with it a concept directed at rules of conduct. Even the application of Fisher's notion of hypothetically infinite populations, from which random samples are drawn, may appear strange. However, this is even more problematic for the Neyman-Pearson concept of "repeated sampling from the same population." When applying such a notion to macroeconomic time series, the question to ask is the following: "[. . .]how often is the question that an econometrician has to answer a decision problem in the context of repeated sampling?"[65]

Why did Haavelmo use precisely this approach as a basis?[66] One possible explanation could be that the rivalry of the early 1940s between the approaches of Fisher and Neyman/Pearson resulted in the latter emerging as the victor, thereby already representing a "paradigm" in the Kuhnian sense. There is also a personal reason: Haavelmo himself reported that he had for various months enjoyed the privilege of studying under the "world's famous statistician" J. Neyman. This may

[64]Research nevertheless still continued to take place in the "old" tradition, as econometrics began to develop. See, for example, Hotelling (1934), Schultz (1934), Greenstein (1935), and Regan (1936). Even the method of moving averages was still being recommended by Sasuly (1936) in this context.

[65]Keuzenkamp/Magnus (1995, p. 18).

[66]Heckman (1992, p. 881) also poses the question in this context, in criticism addressed to Morgan (1990): "Why was the Neyman-Pearson theory adopted as the paradigm of statistical inference in econometrics, and why were rival theories by Ronald Fisher and Harold Jeffreys less successful?".

have shown him, as someone who was then "young and naïve," "ways [. . .] to approach the problem of econometric methodology that were more promising than those that had previously resulted in so much difficulty and disappointment."[67]

Haavelmo certainly saw the problems that lay in a simple application of the Neyman-Pearson concept and therefore argued from an instrumentalist stance. His writings repeatedly contain remarks such as "it has been found fruitful" and similar. In addition, large parts of his explanations are based solely on "hopes":

> [. . .] we might hope to find elements of invariance in economic life, upon which to establish permanent laws [. . .]. Our hope in economic theory and research is that it may be possible to establish constant and relatively simple relations [. . .]. Our hope for simple laws in economics rests upon the assumption that we may proceed as if such natural limitations of the number of relevant factors exist.[68]

Is it justified, with a stance such as this, in starting out from objective inference? Even if we rule out the problematic underpinnings, there is a series of questions that the Neyman-Pearson approach fails to answer. As Heckman correctly notes, Haavelmo did not, for example, take into account the important aspect of model structure and selection:

> These claims have never been rigorously established, even for analyses conducted on large samples. There is no 'correct' way to pick an empirical model and the problems of induction, inference, and model selection are very much open. [. . .] The Neyman-Pearson theory espoused by Haavelmo and the Cowles group takes a narrow view of science. By its rules, hypotheses are constructed in advance of knowledge of the data and the role of empirical work is to test the hypotheses. This rigid separation of model construction and model verification was a cornerstone of classical statistics circa 1944. Even then, influential scholars, primarily Bayesians such as Harold Jeffreys quarreled with this view of empirical science. Since that time, the monopoly of classical statistics has broken.[69]

Haavelmo's application of the Neyman-Pearson paradigm nevertheless formed the basis in econometric research for several decades. Even Koopmans stopped citing Fisher and defended Haavelmo's approach with respect to R. Vining. The physical world view was thus cemented into place. Koopmans drew comparisons between the "complete" systems of structural equations and the explanatory power of Newton's theory of gravitation, while J. Marshak (1950), Chairman of the Cowles Commission, went so far as to regard the issue explicitly as "social engineering." But does this not ominously remind us of the "social physics" – vehemently rejected in its day – of Quetelet?

[67] Haavelmo (1994, p. 75).

[68] Haavelmo (1944, pp. 13, 22f, 24).

[69] Heckman (1992, p. 882). He gives reasons for Morgan's overestimation of Haavelmo's approach – rightly in our opinion – with the view, which can be traced back to the influence of Hendry, that these problems are generally solvable in the context of the Neyman-Pearson approach. This overestimation is also picked up by Malinvaud (1991, p. 635) and Zellner (1992, p. 220).

Alternatives

There have been increasing attempts, ever since the 1970s, to seek out alternative ways. C. Sims[70] proposed vector autoregressive time series models as a counter to traditional systems based on simultaneous equations. These models initially provided nothing more than a description of the delayed correlation structure present in existing time series. One could, in principle, regard vector autoregressive models as the ideal form for cliometrics. They are, however, associated with the same problem as univariate ARIMA models,[71] in that the "right" model must first be found on the basis of the data, which infringes in turn the assumptions of classical inference. It is moreover not possible, given the high degree of complexity of these models, to use the tools developed by Box and Jenkins for use in univariate time series analysis. Sims therefore proposed restricting the high number of parameters that result from such models, thereby ultimately advocating a Bayesian approach.

Bayesian approaches, which marked the beginnings of structural equation models in the econometrics of the 1960s, were still subject, in technical terms, to greater difficulties than classical statistical inference. These technical difficulties should not, however, obscure the fact that the Bayesian standpoint is considered by its representatives to be, from a conceptual point of view, a single approach:

> That there is a unified and operational approach to problems of inference in econometrics and other areas of science is a fundamental point that should be appreciated. Whether we analyze, for example, time series, regression, or 'simultaneous equation' models, the approach and principles will be the same. This stands in contrast to other approaches to inference that involve special techniques and principles for different problems.[72]

E. Leamer developed the most consistent Bayesian econometric methodology.[73] The main criticism of Leamer appears to us to be the part concerning modeling problems. Leamer rightly pointed out that the classical theory, in which the model is regarded as a given, required an almost "Orwellian" approach to econometrics:

> In such a fanciful world, personal uncertainties and public disagreements concerning how to interpret data would be completely resolved in advance. New data sets would not be distributed to humans at all, but instead would be delivered with elaborate security measures to a centralized warehouse where preprogrammed computers would pore over the numbers and pass the conclusions to the public. Once analyzed, the data would be entirely destroyed, to prevent the urge to try something else from becoming an unwanted reality.[74]

[70] See, for example, Sims (1980).

[71] They were also subject to the same statistical limitations, such as stationarity and linearity.

[72] Zellner (1971, p. 11).

[73] See references in Rahlf (1998).

[74] Leamer (1994, p. ix).

The nonexperimental nature of econometrics prohibits such a notion. Data relating to such factors as the development of a country's gross national product are available only once but are evaluated repeatedly. If there is uncertainty regarding the model and – with respect to the selection of relevant variables – (1) the data are not neutral and (2) the personal conviction of the scientist plays a role (e.g., selection of the determinants of criminality by conservative or liberal researchers, selection of the determinants of inflation by monetarists or Keynesians), then a Bayesian point of view is, in our opinion, the only one that can be justified. The indication of the effect of different assumptions and selected variables, or "sensitivity analysis," appears to offer a promising approach in this respect, although its future reliability would have to be underpinned by a larger number of applications.

D. Hendry (2001) has developed a third methodology. Hendry, unlike Leamer, is convinced that a model structure based on the intensive analysis of a data set can be justified by the methods of classical inference. One revealing example of his approach comes from the reanalysis of a selected model based on comprehensive research carried out by M. Friedman and A. Schwartz of monetary trends in Britain and the United States, although the individual steps of the modeling process involved remain partly obscure. The possibility of validation, using classical "testing" based on the theory of Neyman and Pearson, has therefore been questioned in literature on the subject.[75]

Milton Friedman, by his own account, put no trust in formal statistical criteria. He had already rightly pointed out, in his criticism of Tinbergen's consideration of economic theories, that the traditional testing of significance or of hypotheses becomes less meaningful when it is applied after the analysis of the same data. His own t-tests are therefore also more likely to be understood as pragmatic.

If we consider these methodologies and approaches as a whole, a natural science world view dominated econometric research. Most approaches are based above all on constant, time-invariant parameters. Although the consideration of parameter constancy is part of Hendry's testing batteries, seldom alternatives – other than dummy variables – are modeled. Friedman and Schwartz (1991) do in fact point out the significance of analyzing historically uniform periods, but they subject these periods, in turn, to rigid constraints. Complexity is as a rule reduced to a parameter matrix that reflects the time-invariant structure, regardless of whether it concerns short- or long-term relationships.

Inference for Cliometrics

The question regarding the importance of empirical research to economic history and economics was again picked up in 1949, by A. P. Usher. Usher offered numerous philosophical, psychological, and scientific approaches that should justify a modern

[75]Keuzenkamp (1995, p. 243) therefore uses, for Hendry's approach, the more apposite term "diagnostic *checks*" rather than "diagnostic *tests*."

take on empiricism while highlighting its relevance to economic history. However, his references regarding approaches to philosophical probability theory stand in isolation.[76] Seen as a whole, the discipline of economic history in the first half of the twentieth century was, even in the United States, geared more toward a qualitative approach, with a tendency to reject the quantitative.[77]

The Bayesian Origins of Cliometric Inference

What is the current position regarding the concept of inference in cliometrics? If we define cliometrics generally by (1) the application of explicitly theory-driven, neoclassically oriented economic history research along with (2) the intensive use of mass data and of formal methods for verifying the theories based on those data, the question immediately arises of what difference there is, if any, with respect to the intrinsic concept of econometrics. The headword entry for "cliometrics" in the *New Palgrave* defines the approach as an "amalgam of methods,[78] born of the marriage contracted between historical problems and advanced statistical analysis, with economic theory as bridesmaid and the computer as best man,"[79] while the *American Heritage Dictionary* lists it as "the study of history using economic models and advanced mathematical methods of data processing and analysis."[80]

If the predominant characteristic is therefore the use of certain methods,[81] it is consequently surprising that the criticism that cliometrics has attracted on the part of "traditional" economic history has not established methodological problems, in the strict sense of the term, as a subject for discussion. Discussions were centered on the question of whether the application of theoretical models and their verification was, if at all, the cognitive goal of economic history with respect to a specific time and place and whether historical data fulfilled the conditions for applying elaborate statistical methods. The methods themselves played no further role however.

One can say that cliometrics has followed, in terms of methodology, the "paradigm" of econometrics, thereby taking into account the problems described by this field.[82] If we start by assuming, as E. Heckscher (1939) did, that the purpose of

[76]Cf. Usher (1949, p. 148 and p. 155, note 29).

[77]Fogel (1995, S. 49): "The leading history journals, even in economic history, initially refused to accept articles with complex tables and even after such articles began to be accepted, equations were absolutely forbidden."

[78]Floud (1991, p. 452.

[79]Fogel/Elton (1983, S. 2), quoted by Floud (1991, p. 452.

[80]See above, p. 5.

[81]See also Fogel (1995, p. 52) on this subject: "By the early 1980s *cliometric methods* were so firmly established in certain fields of history that no scholar in these fields could afford to neglect them" (our italics).

[82]This is supported not least by the fact that cliometrics did emerge as an independent school of thought because an application for admission by a group of the founding fathers of cliometrics had been rejected by the *Econometric Society*. Cf. Hughes (1965).

economic history is not fundamentally different to that of economics (or econometrics), it becomes plain that the econometric tools available to it, which were already well developed and firmly established by the early 1960s, were accepted uncritically, because econometrics gave, at that time, the most complete impression of its entire history of development.

It is therefore surprising, against the background of this development, that the papers by A. Conrad and J. Meyer (1957, 1958), which are commonly regarded as the "starting pistol" of cliometrics, should go off in a completely different direction. The two economists presented, at a 1957 conference held jointly by the Economic History Association and the National Bureau of Economic Research, a paper entitled *The Economics of Slavery in the Antebellum South*, in which they expounded the thesis – based on statistical methods, data compiled from secondary literature, and a theoretical economic model – that the purchase of a slave in the period before the Civil War represented a profitable investment for a slave owner from the Southern United States. Their work, which was published the following year, raised a storm of protest – and not just because of their "econometric" approach.[83] Our intention here is not to follow this discussion however[84] but rather to examine their methodology. This was also pointed out by the authors in 1957, in a programmatic article on the relationship between economic theory, statistical inference, and economic history, which followed a surprisingly Bayesian line of argument.[85]

Conrad and Meyer set out here to emphasize the significance of a concept of causal orders, which should underpin every historical narrative. The denial of the possibility of causal explanations in history, which was put forward by a number of philosophers, is based mainly on the view that historical events are unique, complex, and unquantifiable.[86] They rightly pointed out that econometric modeling predetermines a causal order, which is only valid for the variables contained in the model[87]: "Causal order is an operational term, which does not require the involvement of any invisible forces or internal needs."[88] The claim that causal explanations are connected to the basic repeatability of an experiment, although historical events are unique, is likewise incorrect. Firstly, experiments are also essentially first-time events and, secondly, a science such as astronomy would no longer be capable of making causal statements, as it would then be dealing with non-repetitive phenomena.[89] This is where Bayesian reasoning came into play, as it is not based on the

[83]Conrad/Meyer (1958). The authors were at the time *assistant* professors of economics at Harvard. The expressions "starting gun" and "watershed" are therefore justified, since econometric methods were for the first time being applied to historical phenomena without any reference to the present.

[84]Cf. Conrad/Meyer (1964).

[85]Conrad/Meyer (1957).

[86]Cf. Conrad/Meyer (1957, p. 527).

[87]They refer here to an example given by Simon (1957) regarding the differing possible influences of the variables of weather, wheat harvest yield, and wheat price.

[88]Conrad/Meyer (1957, p. 147.

[89]They seek support, in this context, in the line of argument of H. Jeffreys.

repeatability of the events but is concerned rather with a subjective grasp of statements of probability:

> Explicitly, the formal tests attach an actual numerical probability to the correctness of the hypothesis in the light of the observed results. This introduces the question of relative plausibility into the empirical procedure and consequently helps the investigator to scale the degree of belief, an intrinsically ordinal concept at the very least, that should be placed in the hypothesis. There are, in sum, substantial advantages as well as disadvantages to the introduction of more formal procedures in the evaluation of historical hypotheses. The question therefore arises: Is there a satisfactory compromise that embodies maximum advantage with minimum disadvantage? Ideally, the best procedure would appear to be one in which the formal tests were adapted or altered to take account of a maximum of a priori information. This leads, admittedly, to an essentially Bayesian approach to statistical inference.[90]

They did indeed see it as problematic that the Bayesian approach was sinking into a "morass of subjectivism," in the immediate absence of a priori notions and probabilities. They were, however, confident that this could form a basis for creating guidelines and simplifying the communication of scientific results.

The discussion following the presentation of the paper, in which the economists present expressed their opposition to the application to historical data of econometric models and statistical tests, included (in the same manner as later papers on cliometrics) little evidence of this key difference with respect to the prevailing econometric approach.[91] Subsequent development tended rather to follow the path marked out by econometrics, albeit without influencing econometrics itself.

The cliometric (r)evolution, whose development had been swiftly gathering pace since the 1960s, then took over the methods of econometrics, along with its associated concepts. A way such as this was, for logical reasons, just as faintly compelling as the adoption by econometrics of "classic" statistical inference. The papers by Conrad and Meyer, which marked the beginnings of cliometrics, followed a Bayesian argument, although this was subsequently taken into account neither by cliometrics itself nor its critics. The course of econometrics was actually set by physicists in their role as "social engineers," harking back implicitly (or even explicitly) to Newton. We would like to conclude our overview by citing some criticism that is very revealing for statistical inference in the field of cliometrics: the critique of the mathematician Rudolf Kalman.

[90] Conrad/Meyer (1957, p. 544). Specific examples can be found in Conrad/Meyer (1964).

[91] Bayesian approaches were not to find fertile ground in the field of econometrics until several years later. It must however be emphasized that the line of argument maintained by Conrad and Meyer contained various terms and concepts (they speak of objective tests and significant differences, before returning to probabilities of hypotheses and a "morass of subjectivism") that cannot always be clearly differentiated from each other.

Fundamental Criticism: Rudolf Kalman

Rudolf Kalman took on a study, in the early 1980s, of the problem of model structure and inference in the field of econometrics and expressed fundamental criticism, in this context, from a system theory point of view.[92] In his opinion, econometrics mainly went along the following two paths:

1. Economic laws and relationships have been formulated as dynamic equations in terms of Newton's laws.
2. The coefficients of these equations have been determined quantitatively by the extraction, from real data, of statistically relevant information.

He establishes, with this development in mind, that the progress in knowledge subsequently achieved is, even in comparison to the 250 years that have elapsed since Newton, disappointingly low. He expounds the thesis (which requires, in his opinion, no discussion in terms of "hard" science):

> [. . .] that economics is not at all like physics and therefore that it is not accessible by a methodology that was successful for physics. Far from being governed by absolute, universal, and immutable laws, economic knowledge, unlike physical science, is strongly system (context) dependent; when economic insights are taken out of temporal, political, social, or geographical context, they become trivial statements with little information content. [. . .] Since economic 'laws' do not process the attributes of physical laws, writing down equations, in the style of physics, to translate economic statements into mathematics is not a productive enterprise. [. . .] System theory provides a simple but hard suggestion: Do not write equations expressing assumed relationships; deduce your equations from real data. [. . .] To put it differently, there will never be a Newton in economics; the path to be followed must be different.[93]

His opinion on the second step, the statistical determination of unique parameters, is even more negative. This only makes sense in his view if there are concrete, explicitly measurable parameters, as is the case, for example, with resistors in Ohm's law:

> Economists have often dreamed of imitating the simple situation characterized by Ohm's law just by hoping for the best, for example, by assuming that such a law (the Phillips curve) exists between inflation and unemployment. But unemployment and inflation, in any quantitative sense, are fuzzy and politically biased attempts to replace complex situations by (meaningless) numbers; consequently any hope that two such concepts can be tied to one another by a single coefficient is barbarously uninformed wishful thinking.[94]

[92] We rely mainly on Kalman (1982a, b) in this respect. We are therefore not concerned with the *application* of the so-called Kalman filter to econometrics.

[93] Kalman (1982a, p. 19f). Original author's italics.

[94] Ibid., p. 20.

Unique relationships such as these exist in astronomy, for example, where their parameters have a direct significance that is independent of any system, such as the determining of the position of an object as a function of moment and angle. It is not surprising, against this background, that Kalman is especially critical of Haavelmo's approach: "The aspiration of Haavelmo to give a solid foundation to econometrics by dogmatic application of probability theory has not been fulfilled (in the writer's opinion), no doubt because probability theory has nothing to say about the underlying system-theoretic problems."[95] He calls instead for a rigorous application of system theory. System theory does not set out from a directly measurable relationship between input and output: "instead of determining a single parameter, such as a resistance, system theory is concerned with the much more general question of determining a system."[96] Parameters contained in systems have, according to Kalman, a completely different significance to that hitherto assumed by econometricians; they are therefore to be defined only *locally*. It is by no means self-evident, for Kalman, that the cognitive goal of statistical analysis should lie in the obtaining of constant figures, such as with the application of maximum likelihood estimates or the method of least squares: "[. . .]common sense should tell us, that such a miracle is possible only if additional assumptions (*deus ex machina*) are imposed on the data which somehow succeed in neutralizing the intrinsic uncertainty."[97] The method of least squares is thus so popular in these terms because it delivers a clear ("unique") response. However, the assumptions associated with such an approach cannot normally be justified.

The common approach of using data that show variance to determine a specific value that reveals maximum likelihood or minimizes deviations (thereby making it preferable to all others) is, for him, "fundamentally wrong and extremely harmful to scientific progress."[98] Such an approach implies the following suppositions (or "prejudices"):

1. The data have been generated using a probabilistic mechanism.
2. This probabilistic mechanism is very simple; it is constant in terms of time, and a distribution function explains everything.
3. There is a "true" value, which can be regarded as the "particularly striking feature" of the hypothetical distribution function, such as the expected value, median, or modal value.
4. A single figure constitutes the response of a deductive process based on self-evident postulates.

[95]This sentence, which was supposed to appear in Kalman (1982c), was deleted at editorial request and included instead in Kalman (1982b, p. 194).

[96]Kalman (1982a, p. 23). Linearity and finiteness might be reasonable assumptions for such a system.

[97]Kalman (1982b, p. 162). Original author's italics.

[98]Ibid., p. 171.

The assumption of exact conformity to natural laws in probabilistic phenomena analogous to Newtonian physics, which is what such an approach is supposed to aspire to, has nevertheless long since proved to be an illusion. Apart from "mathematical artifacts," such as the law of large numbers, there have not been any universal laws of random phenomena – even in physics – but rather ones that depend on the very system that surrounds them.[99] A view such as this has profound implications:

> The implications of this situation for econometric strategy are devastating. Since the problem is to identify a system and since systems cannot be described in general by globally definable parameters, the whole idea of a parameter loses its (uncritically assumed) significance. [. . .] The Jugendtraum of econometrics, determining economically meaningful parameters from real data via dynamical equations supplied from economic theory, turns out to have been a delusion.[100]

This criticism of Kalman has not as yet – as far as we can see – had any impact on econometrics. Even if one does not wish to follow the path to its ultimate consequences, the fundamental reasonableness of applying physics-based approaches to economic developments should still be examined. It is surely a highly promising basis for inference statements in the field of cliometrics.

References

Barnard G (1947) The meaning of a significance level. Biometrika 34:179–182
Barnard G (1949) Statistical inference (with discussion). J R Stat Soc B 11:115–149
Barnard G (1988) R. A. Fisher – a true Bayesian? Int Stat Rev 55:183–189
Berger J, Wolpert R (1988) The likelihood principle. , vol 6, 2nd edn, Lecture notes – monograph series. Institute of Mathematical Statistics, Hayward
Birnbaum A (1962) On the foundations of statistical inference (with discussion). J Am Stat Assoc 57:269–306
Birnbaum A (1968) Likelihood. In: Sills D (ed) International encyclopedia of the social sciences. Macmillan, New York, pp 299–301
Birnbaum A (1977) The Neyman-Pearson theory as decision theory and as inference theory: with a criticism of the Lindley-Savage argument for Bayesian theory. Synthese 36:19–49
Bjornstad J (1992) Introduction to Birnbaum (1962) on the foundations of statistical inference. In: Kotz S, Johnson N (eds) Breakthroughs in statistics. Bd. I. Foundations and basic theory. Springer, New York, pp 461–477
Borovcnik M (1992) Stochastik im Wechselspiel von Intuitionen und Mathematik. Spektrum Akademischer Verlag, Mannheim
Boumans M (1993) Paul Ehrenfest and Jan Tinbergen: a case of limited physics transfer. In: de Marchi N (ed) Non-natural social sciences: reflecting on the enterprise of 'More heat than light',

[99]Cf. ibid., p. 172.

[100]Kalman (1982a, pp. 26, 27). He describes the calculation of a constant parameter (e.g., in the context of the Phillips curve) as a "conceptual absurdity" (ibid.). Kalman consequently also rejects any causal interpretation. Cf. Kalman (1982b, p. 177), for example.

vol 25, Supplement to history of political economy. Duke University Press, Durham/London, pp 131–156
Burns AF, Mitchell WC (1946) Measuring business cycles. National Bureau of Economic Research, New York
Conrad A, Meyer J (1957) Economic theory, statistical inference, and economic history. J Econ Hist 17:524–544
Conrad A, Meyer J (1958) The economics of slavery in the antebellum south. J Polit Econ 66:95–130
Conrad A, Meyer J (eds) (1964) The economics of slavery. Studies in econometric history. Aldine, Chicago
Dale A (1991) A history of inverse probability. From Thomas Bayes to Karl Pearson, vol 16, Studies in the history of mathematics and physical sciences. Springer, New York
de Finetti B (1937) La prévision: Ses lois logiques, ses sources subjectives. Ann V Institut Henri Poincaré 1:1–68
de Finetti B (1981) Wahrscheinlichkeitstheorie. Einführende Synthese mit kritischem Anhang. Oldenbourg, Wien/München
DuMouchel W (1992) Introduction to Edwards, Lindman, Savage (1963) Bayesian statistical inference for psychological research. In: Kotz S, Johnson N (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory. Springer, New York, pp 519–530
Edgeworth FY (1884) The philosophy of chance. Mind 9(34):223–235
Edwards W, Lindman H, Savage L (1963) Bayesian statistical inference for psychological research. Psychol Rev 70:193–242 [Reprinted in: Kotz S, Johnson N (1992) (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory, New York]
Epstein RJ (1987) A history of econometrics. North Holland, Amsterdam
Ezekiel M (1928) Statistical analysis and the law' of price. Q J Econ 42:199–227
Fisher R (1922 [1992]) On the mathematical foundations of theoretical statistics. Philos Trans R Soc Lond A 222:309–368 [Reprinted in: Kotz S, Johnson N (1992) (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory, New York]
Fisher R (1955) Statistical methods and scientific induction. J R Stat Soc B 17:69–78
Fisher R (1956) Statistische Methoden für die Wissenschaft, 12 Aufl. Oliver and Boyd, Edinburg
Fisher R (1959) Statistical methods and scientific inference, 2nd edn. Oliver and Boyd, London
Floud R (1991) Cliometrics. In: Eatwell J, Milgate M, Newman P (eds) The new Palgrave. A dictionary of economics. Bd. 1, 2nd edn. Macmillan, London/New York/Tokyo, pp 452–454
Fogel R (1995) History with numbers: the American experience. In: Etemad B, Batou J, David T (eds) Pour une histoire économique et sociale internationale. Ed. Passé Présent. Genf, Genéve, pp 47–56
Fogel R, Elton G (1983) Which road to the past? Two views of history. Yale University Press, New Haven/London
Friedman M, Schwartz A (1991) Alternatives approaches to analyzing economic data. Am Econ Rev 81(1):39–49
Frisch R (1931) A method of decomposing an empirical series into its cyclical and progressive components. J Am Stat Assoc (Suppl) 26:73–78
Frisch R (1933) Propagation problems and impulse problems in dynamic economics. In: Essays in honour of Gustav Cassel. Allen & Unwin, London
Galton F (1888) Co-relations and their measurement. Proc R Soc Lond Ser 45:135–145
Geisser S (1992) Introduction to Fisher (1922) on the mathematical foundations of theoretical statistics. In: Kotz S, Johnson N (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory. Springer, New York, pp 1–10
Gigerenzer G, Swijtink T, Porter T, Daston L, Beatty J, Krüger L (1989) The Empire of chance: how probability changed science and everyday life. Cambridge University Press, Cambridge/New York
Gosset WS (1908) The probable error of a mean. Biometrika 6:1–25

Graunt J (1662 [1939]) Natural and political observations made upon the bills of mortality. Edited with an introduction by Willcox WF John Hopkins University Press, Baltimore
Greenstein B (1935) Periodogram analysis with special application to business failure in the U.S. 1867–1932. Econometrica 3:170–198
Haavelmo T (1944) The probability approach in econometrics. Econometrica 12(Suppl):1–115
Haavelmo T (1994) Ökonometrie und Wohlfahrtsstaat. Nobel-Lesung vom 7. Dezember 1989. In: Grüske K-D (ed) Die Nobelpreisträger der ökonomischen Wissenschaft. Bd. 3. 1989–1993. Wirtschaft und Finanzen, Düsseldorf, pp 71–80
Hall L (1925) A moving secular trend and moving integration. J Am Stat Assoc 20:13–24
Halley E (1693) An estimate of the degrees of mortality of mankind drawn from curious tables of the births and funerals at the city of Breslau; with an attempt to ascertain the price of annuities upon lives. Philos Trans R Soc 17:596–610. Electronic reprint: http://www.pierre-marteau.com/editions/1693-mortality.html
Heckman J (1992) Haavelmo and the birth of modern econometrics: a review of the history of econometric ideas by Mary Morgan. J Econ Lit 30:876–886
Heckscher E (1939) Quantitative measurement in economic history. Q J Econ 53:167–193
Hendry DF (2001) Econometrics: alchemy or science? 2nd edn. Oxford University Press, Oxford
Hodges J (1990) Can/may Bayesians do pure tests of significance? In: Geisser S, Hodges J, Press S, Zellner A (eds) Bayesian and likelihood methods in statistics and econometrics. Essays in honor of George A. Barnard, vol 7, Studies in Bayesian econometrics and statistics. North Holland Publishing, New York, pp 75–90
Hotelling H (1934) Analysis and correlation of time series. Econometrica 2:211
Howson C (1995) Theories of probability. Br J Philos Sci 46:1–32
Hughes J (1965) A note in defense of Clio. Explor Entrep Hist 3:154
Iversen G (1984) Bayesian statistical inference. Sage, Newbury Park
Jeffreys H (1939) Theory of probability. The Clarendon Press, London/New York
Johnstone D (1986) Tests of significance in theory and practice (with discussion). Statistician 35:491–504
Kalman R (1982a) Dynamic econometric models: a system-theoretic critique. In: Szegö G (ed) New quantitative techniques for economic analysis. Academic, New York, pp 19–28
Kalman R (1982b) Identification from real data. In: Hazewinkel M, Rinnooy Kan A (eds) Current developments in the interface: economics, econometrics and mathematics. Reidel, Dordrecht, pp 161–196
Kalman R (1982c) Identifiability and problems of model selection in econometrics. In: Hildenbrand W (ed) Advances in econometrics. Cambridge University Press, Cambridge
Kempthorne O (1971) Comment on 'Applications of statistical inference to physics'. In: Godambe V, Sprott D (eds) Foundations of statistical inference. Holt, Rinehart and Winston of Canada, Toronto, pp 286–287
Keuzenkamp H (1995) The econometrics of the Holy Grail – a review of Econometrics: alchemy or science? Essays in econometric methodology. J Econ Surv 9:233–248
Keuzenkamp H, Magnus J (1995) On tests and significance in econometrics. J Econ 67:5–24
Keynes J (1921) A treatise on probability. Macmillan, London
Koopmans T (1941) The logic of econometric business-cycle research. J Polit Econ 49:157–181
Kuznets S (1928a) On moving correlation of time sequences. J Am Stat Assoc 23:121–136
Kuznets S (1928b) On the analysis of time series. J Am Stat Assoc 23:398–410
Kuznets S (1929) Random events and cyclical oscillations. J Am Stat Assoc 24:258–275
Kuznets S (1930a) Secular movements in production and prices. Houghton Mifflin, Boston/New York
Kuznets S (1930b) Wesen und Bedeutung des Trends. Zur Theorie der säkularen Bewegung, Veröffentlichungen der Frankfurter Gesellschaft für Konjunkturforschung. Schroeder, Bonn
Kuznets S (1934) Time series. In: Seligman E, Johnson A (eds) Encyclopedia of the social sciences. Bd. 13. Macmillan, New York, pp 629–636

Kyburg H (1985) Logic of statistical reasoning. In: Kotz S, Johnson N (eds) Encyclopedia of statistical sciences. Bd. 5. Wiley, New York, pp 117–122
Kyburg H, Smokler H (eds) (1964) Studies in subjective probability. Wiley, New York
Laplace P-S (1812) Théorie analytique des probabilités. Courcier, Paris. https://archive.org/details/thorieanalytiqu01laplgoog
Leamer EE (1994) Introduction. In: Leamer EE (ed) Sturdy econometrics. Elgar, Aldershot, pp ix–xvi
Lehmann E (1992) Introduction to Neyman and Pearson (1933) on the problem of the most efficient tests of statistical hypotheses. In: Kotz S, Johnson N (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory. Springer, New York, pp 67–72
Lehmann EL (1993) The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two. J Am Stat Assoc 88:1242–1249
Lindley D (1991) Statistical inference. In: Eatwell J, Milgate M, Newman P (eds) The new Palgrave. A dictionary of economics, vol 4, 2 Aufl. Macmillan, London/New York/Tokyo, pp 490–493
Malinvaud E (1991) Review of Morgan, Morgan M (1990) the history of econometric ideas. Econ J 101:634–636
Magnus J, Morgan M (1987) The ET interview: Professor J. Tinbergen. Econ Theory 3:117–142
Marshak J (1950) Statistical inference in economics. In: Koopmans T (ed) Statistical inference in dynamic economic models. Wiley, New York
Menges G (1972) Grundriß der Statistik. 1. Theorie, 2nd edn. Westdeutscher Verlag, Opladen
Mirowski P (1989) The probabilistic counter revolution, or how stochastic concepts came to neoclassical economic theory. Oxf Econ Pap 41:217–235
Mirowski P (1991) The when, the how and the why of mathematical expression in the history of economic analysis. J Econ Perspect 5:145–157
Morgan M (1990) The history of econometric ideas. Cambridge University Press, Cambridge
Neyman J (1937) Outline of a theory of statistical estimation based on the classical theory of probability. Philos Trans R Soc Lond Ser A Math Phys Sci 236(767):333–380
Neyman J, Pearson ES (1928a) On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. Biometrika 20A:175–240
Neyman J, Pearson ES (1928b) On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. Biometrika 20A:263–294
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond Ser A, containing papers of a mathematical or physical character 231:289–337 [Reprinted in: Kotz S, Johnson N (1992) (eds) Breakthroughs in statistics. Bd. 1. Foundations and basic theory, New York]
Pearson K (1894) Contributions to the mathematical theory of evolution. Philos Trans R Soc Lond 85:71–110
Pearson K (1895) Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philos Trans R Soc Lond 186:343–414
Pearson K (1898) Mathematical contributions to the theory of evolution: on the law of ancestral heredity. Proc R Soc Lond 62:386–412
Pearson K (1920) The fundamental problem of practical statistics. Biometrika 13(1):1–16
Pearson ES (1967) Some reflections on continuity in the development of mathematical statistics, 1885–1920. Biometrica 52:3–18
Popper K (1990) A world of propensities. Thoemmes, Bristol
Pratt J (1971) Comment on: 'probability, statistics and knowledge business' by O. Kempthorne. In: Godambe V, Sprott D (eds) Foundations of statistical inference. Holt, Rinehart and Winston, Toronto
Rahlf T (1998) Deskription und Inferenz Methodologische Konzepte in der Statistik und Ökonometrie, vol 9, Historical social research supplement. Zentrum für Historische Sozialforschung, Köln

Ramsey F (1931a) Truth and probability (1926). In: Braithwaite R (ed) The foundations of mathematics and other logical essays by Frank Plumpton Ramsey. International Library of Psychology, Philosophy and Scientific Method, London [Reprinted in Kyburg, Smokler (1964)]
Ramsey F (1931b) Further considerations (1928). In: Braithwaite R (ed) The foundations of mathematics and other logical essays by Frank Plumpton Ramsey. International Library of Psychology, Philosophy and Scientific Method, London, pp 199–211
Regan F (1936) The admissibility of time series. Econometrica 4:189
Robbins H (1955) An empirical Bayes approach to statistics. In: Neyman J (ed) Proceedings of the 3rd Berkeley symposium on mathematical and statistical probability, University of California. Statistical Laboratory: University of California Press, vol 1, pp 157–163 [Reprinted in Kotz/Johnson (1992)]
Sasuly M (1936) A method of smoothing economic time series by moving averages. Econometrica 4:206
Savage L (1954) The foundations of statistics. Wiley, New York
Savage L (1976) On rereading R. A. Fisher (with discussion). Ann Stat 4:441–500
Schultz H (1934) Discussion of the question 'Is the theory of harmonic oscillations useful in the study of business cycles?'. Econometrica 2:189
Sims C (1980) Macroeconomics and reality. Econometrica 48:1–48
Simon H (1957) Models of man. Wiley, New York
Slutzky E (1937) The summation of random causes as the source of cyclic processes. Econometrica 5:105–146 [originally published in Russian 1927]
Stegmüller W (1973) Personelle und Statistische Wahrscheinlichkeit. Erster Halbband: Personelle Wahrscheinlichkeit und Rationale Entscheidung. Zweiter Halbband. Statistisches Schließen, Statistische Begründung, Statistische Analyse. Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie IV. Springer, Berlin/Heidelberg/New York
Stigler S (1986) The history of statistics: the measurement of uncertainty before 1900. Belknap Press of Harvard University Press, Cambridge, MA
Usher A (1949) The significance of modern empiricism for history and economics. J Econ Hist 9:131–155
von Mises R (1951) Wahrscheinlichkeit, Statistik und Wahrheit. Springer, Wien
Wald A (1950) Statistical decision functions. Wiley, New York
Watson G (1983) Hypothesis testing. In: Kotz S, Johnson N (eds) Encyclopedia of statistical sciences. Bd. 3. Wiley, New York, pp 712–722
Yule GY (1895) On the correlation of total pauperism with proportion of out-relief, I: all ages. Econ J 5:603–611
Yule GU (1896a) Notes on the history of pauperism in England and Wales from 1850, treated by the method of frequency-curves; with an introduction on the method. J R Stat Soc 59(2):318–357
Yule GY (1896b) On the correlation of total pauperism with proportion of out-relief, II: males over sixty-five. Econ J 6:613–623
Yule GY (1927) On a method of investigating the periodicities of disturbed series, with special reference to Wolfer's sunspot numbers. Philos Trans R Soc A 226(1927):267–298
Zellner A (1971) An introduction to Bayesian statistics in econometrics. Wiley, New York
Zellner A (1992) Review of Morgan, Morgan M (1990) the history of econometric ideas. J Polit Econ 100:218–222

Recommended Reading

The best starting point is still Gigerenzer et al. (1989). See Cited Literature. Other helpful overviews are:

Cohen IB (2005) The triumph of numbers: how counting shaped modern life. W. W. Norton, New York

Kotz S, Johnson NL (eds) (1992) Breakthroughs in statistics, 1. Foundations and basic theory. 2. Methodology and distribution, Springer series in statistics. Springer, New York

Lenhard J (2006) Models and statistical inference: the controversy between Fisher and Neyman-Pearson. Br J Philos Sci 57:69–91

Salsburg D (2001) The lady tasting tea: how statistics revolutionized science in the twentieth century. Freeman, New York

Sprenger J (2014) Bayesianism vs frequentism in statistical inference. In: Hájek A, Hitchcock C (eds) Handbook of the philosophy of probability. Oxford University Press, Oxford

Sprenger J, Hartmann S (2001) Mathematics and statistics in the social sciences. In: Jarvie IC, Bonilla JZ (eds) The SAGE handbook of the philosophy of social sciences. Sage, London, pp 594–612

Stigler SM (1999) Statistics on the table: the history of statistical concepts and methods. Harvard University Press, Cambridge, MA

Trends, Cycles, and Structural Breaks in Cliometrics

Terence C. Mills

Contents

Abstract

The calculation of trends and their growth rates, along with the related calculation of cycles, is an important area of cliometrics. The methods traditionally employed to estimate trend were either the estimation of regressions containing simple functions of time, typically in conjunction with a method to deal with regime shifts or structural breaks, or simple unweighted moving averages. In both cases the cycle was determined by residual and, because the trend was, possibly locally, deterministic, the cyclical component took up most of the fluctuations in the observed series. The last 25 years or so, however, have seen major developments in both macroeconomics and time series econometrics and statistics on the modelling of trends and cycles that allow all components to be stochastic and

T. C. Mills (✉)
School of Business and Economics, Loughborough University, Loughborough, UK
e-mail: t.c.mills@lboro.ac.uk

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_21

perhaps determined by the statistical properties of the observed time series. This chapter provides a survey of these developments.

Keywords
Cycles · Filters · Segmented trends · Structural models

Introduction

The calculation of trends and their growth rates, along with the related calculation of cycles, is an important area of cliometrics. The methods traditionally employed to estimate trend were either the estimation of regressions containing simple functions of time, typically in conjunction with a method to deal with regime shifts or structural breaks, or simple unweighted moving averages. In both cases the cycle was determined by residual and, because the trend was, possibly locally, deterministic, the cyclical component took up most of the fluctuations in the observed series.

The last 25 years or so, however, have seen major developments in both macroeconomics and time series econometrics and statistics on the modelling of trends and cycles, with perhaps the first cliometric paper to use these new techniques being Crafts et al. (1989a). Since then Crafts and Mills (1994a, b, 1996, 1997, 2004) and Mills and Crafts (1996a, b, 2000, 2004) have provided a variety of extensions to the range of techniques and cliometric applications. This chapter outlines these developments and may also be regarded as an update of previous surveys of this area by Mills (1992, 1996, 2000).

Sections "History of Modelling Trends and Cycles in Economics" and "Modelling Trends and Cycles in Economic History" contain a brief history of the modelling of trends and cycles and of their traditional application in economic history, respectively. Section "Segmented Trend Models" introduces segmented and breaking trend models, while section "Filters for Extracting Trends and Cycles" considers the modern filter approach for extracting trends and cycles. The link between these filters and structural time series models is developed in section "Filters and Structural Models" and their link with ARIMA models in section "Model-Based Filters." Section "Structural Trends and Cycles" introduces the latest generalizations of structural time series models, while section "Models with Correlated Components" considers the implications of relaxing the identifying constraint of all these models that the components are uncorrelated. Section "Multivariate Extensions of Structural Models" looks at multivariate extensions of structural models and section "Estimation of Structural Models" briefly considers their estimation via a state space framework using the Kalman filter. The possibility of common breaks across a set of series, the phenomenon of co-breaking, is the topic of section "Structural Breaks Across Series," while section "Concluding Remarks" offers some concluding remarks on the nature of trends.

Several examples illustrate the methods introduced in this chapter and these use the British per capita GDP series recently provided by Broadberry et al. (2011), all calculations being done with the commercial software *Econometric Views 8* and *STAMP 8*.

History of Modelling Trends and Cycles in Economics

The analysis of cycles in economic time series began in earnest in the 1870s with the sunspot and Venus theories of William Stanley Jevons and Henry Ludwell Moore and the rather more conventional credit cycle theory of Clément Jugler (see Morgan 1990, Chap. 1). Secular, or trend, movements were first studied somewhat later, with the term "trend" only being coined in 1901 by Reginald Hooker when analyzing British import and export data (Hooker 1901). The early attempts to take into account trend movements, typically by detrending using simple moving averages or graphical interpolation, are analyzed by Klein (1997), while the next generation of weighted moving averages, often based on actuarial graduation formulae using local polynomials, are surveyed in Mills (2011, Chap. 10).

The first half of the twentieth century saw much progress, both descriptive and theoretical, on the modelling of trends and cycles, as briefly recounted in Mills (2009a), but it took a further decade for techniques to be developed that would, in due course, lead to a revolution in the way trends and cycles were modelled and extracted. The seeds of this revolution were sown in 1961 – a year termed by Mills (2009a) as the "annus mirabilis" of trend and cycle modelling – when four very different papers, by Klein and Kosobud (1961), Cox (1961), Leser (1961), and Kalman and Bucy (1961), were published. The influence of Klein and Kosobud for modelling trends in macroeconomic time series – the "great ratios" of macroeconomics – is discussed in detail in Mills (2009b) and that of Cox in Mills (2009a). It is the last two papers that are of prime interest here. As is discussed in section "Filters for Extracting Trends and Cycles," Leser's paper, in which he considered trend extraction from an observed series using a weighted moving average with the weights derived using the principle of penalized least squares, paved the way for one of the most popular trend-extraction methods in use today, the Hodrick-Prescott (H-P) filter. Kalman and Bucy, along with its companion paper, Kalman (1960), set out the details of the Kalman filter algorithm, an essential computational component of many trend and cycle extraction techniques (see section "Estimation of Structural Models"; Young 2011 may be consulted for both historical perspective and a modern synthesis of the algorithm with recursive estimation techniques).

Modelling Trends and Cycles in Economic History

The difficulties in separating out cyclical fluctuations from the longer-term, secular, movements of economic time series were certainly well appreciated by economic historians such as Aldcroft and Fearon (1972) and Ford (1981). However, the methods of trend and cycle decomposition used by them were essentially ad hoc, designed primarily for ease of computation without real regard for the statistical properties of the time series (or set of series) being analyzed: for statements supporting this position, see Aldcroft and Fearon (1972, p. 7) and Matthews et al. (1982, p. 556).

The underlying model in such analyses is that of an additive decomposition of the series x_t, observed over the period $t = 1, 2, \ldots, T$, into a trend, μ_t, and a cycle, ψ_t, typically assumed to be independent of each other, i.e.,

$$x_t = \mu_t + \psi_t \qquad E(\mu_t \psi_s) = 0 \quad \text{for all } t \text{ and } s \tag{1}$$

The observed series x_t is often the logarithm of the series under consideration, while the data are usually observed annually.

The trend and cycle components are, of course, unobservable and hence need to be estimated. The methods of estimation traditionally employed by economic historians are termed ad hoc above because they do not arise from any formal statistical analysis of x_t or its components. Perhaps the simplest model for μ_t that might be considered is the linear time trend $\mu_t = \alpha + \beta t$, which, if x_t is indeed the logarithm of the series, assumes constant growth. Estimation of the regression model

$$x_t = \alpha + \beta t + u_t \tag{2}$$

by ordinary least squares (OLS) then provides asymptotically efficient estimates of α and β. Given such estimates $\hat{\alpha}$ and $\hat{\beta}$, the trend component is then $\hat{\mu}_t = \hat{\alpha} + \hat{\beta} t$ and the cyclical component is obtained by residual as $\hat{\psi}_t = x_t - \hat{\mu}_t$.

The trend component will only be efficiently estimated in small samples, an important proviso given the often limited number of observations available on historical time series, if the cyclical component is, *inter alia*, serially uncorrelated. This is unlikely to be the case if cycles, often defined as "recurring alternations of expansion and contraction" (Aldcroft and Fearon 1972, p. 4), are in fact present in the data, in which case either generalized least squares (GLS) or an equivalent estimation technique should be used or Newey and West (1987)-type consistent variances should be employed with the OLS estimates.

Although the linear model (2) has been used on occasions, most notably by Frickey (1947) and Hoffmann (1955), economic historians have typically rejected the view that trend growth is constant through time, preferring models that allow for variable trend growth rates. The linear trend model can readily be adapted to allow trend growth to vary across cycles, or *growth phases* as they are sometimes referred to, the terminal years of which are chosen through a priori considerations. Thus, if T_1 and $T_2 = T_1 + k$ are the terminal years of two successive cycles, trend growth across the cycle spanning the years T_1 and T_2 is given by the OLS estimate of β_k in the regression

$$x_t = \alpha_k + \beta_k t + u_{kt} \qquad t = T_1, T_1 + 1, \ldots, T_2 \tag{3}$$

A variant of this approach was used by Feinstein et al. (1982), who preferred to estimate β_k by connecting the actual values of the series in the chosen terminal years. This estimate is approximately given by $k^{-1}(x_{T_2} - x_{T_1})$, but Crafts et al. (1989b) show that it is never a more efficient estimator than the OLS estimator $\hat{\beta}_k$.

The common feature of linear trend models of this type is that trend growth across cycles is regarded as being deterministic, so that *all* fluctuations in x_t must be attributed to the cyclical component. Furthermore, any fluctuation from trend can only be temporary: since the cyclical component is estimated by the regression residual, it must have zero mean and be stationary, so that shocks to x_t that force it away from its trend path must dissipate through time. A further drawback of models such as Eq. 3 is that the selection of the terminal years of the cycles could be subjectively biased.

Because of these shortcomings, many economic historians have favored an alternative method of trend estimation, that of using a *moving average*. The typical moving average used to isolate trend in annual macroeconomic time series is one of nine years (as used, e.g., by both Aldcroft and Fearon 1972 and Ford 1969, 1981). Formally, a trend component estimated by a $(2h + 1)$ year moving average of x_t can be defined using the lag operator B as

$$\hat{\mu}_t = M(B)x_t = \frac{1}{2h+1}\left(\sum_{j=-h}^{h} x_{t-j}\right) = \frac{1}{2h+1}\left(\sum_{j=-h}^{h} B^j\right)x_t \tag{4}$$

where $B^j x_t \equiv x_{t\,-\,j}$, so that setting $h = 4$ gives the 9-year moving average referred to above. An advantage of using a moving average to estimate the trend component, apart from the obvious one of computational simplicity, is that the trend now becomes stochastic and, although "smooth," is influenced by the local behavior of x_t: fluctuations in x_t are therefore not entirely allocated to the cyclical component.

A property of moving averages is that a $(2h + 1)$ year moving average will smooth out a $2h + 1$ year cycle in the data. Since many economic historians believe business cycles are between 7 and 11 years in duration, the setting $h = 4$ thus has a rational basis, at least in terms of prior beliefs.

One obvious disadvantage of moving averages is that $2h$ trend observations, equally allocated at the beginning and end of the sample period, are necessarily lost. As Aldcroft and Fearon (1972) note, this can cause major difficulties when the available number of observations is limited. An important illustration of this is the estimation of trends during the interwar years, when less than 20 annual observations are available: Aldcroft and Fearon have to resort to linear trends in their analysis of this period. A less well known disadvantage of using moving averages of the form (4) is that, although they eliminate a linear trend, which is certainly what is required, they also tend to produce too smooth a trend and thus a potentially distorted cyclical component.

Segmented Trend Models

A natural generalization of Eq. 3 is to incorporate the models for individual cycles into a single composite model, where it is assumed that the end points of the cycles are at $T_1, T_2, \ldots, T_{m+1} = T$:

$$x_t = \alpha + \beta t + \sum_{i=1}^{m+1} \gamma_i d_{it} + \sum_{i=1}^{m+1} \delta_i t d_{it} + u_t \tag{5}$$

where the d_{it}, $i = 1, 2, \ldots, m + 1$, are 0–1 dummies taking the value 1 in the ith cycle and zero elsewhere. As trend growth in the ith cycle is given by $\beta + \delta_i$, the hypothesis of a constant trend growth rate across the entire sample period is thus $\delta_1 = \delta_2 = \ldots = \delta_{m+1} = 0$, while the further hypothesis $\gamma_1 = \gamma_2 = \ldots = \gamma_{m+1} = 0$ restricts x_t to having a *single* trend path. However, if this second hypothesis is rejected, then the presence of nonzero γ_is will result in horizontal shifts in the trend, so that models of the form (5) are referred to as *breaking trends*. If it is thought that the trend function should be smooth, continuity can be imposed by considering the class of *segmented trend* models. A segmented linear trend can be written as

$$x_t = \alpha + \beta t + \sum_{i=1}^{m} \theta_i D_{it} + u_t \tag{6}$$

where

$$D_{it} = \begin{cases} t - T_i & t > T_i \\ 0 & \text{otherwise} \end{cases}$$

Thus, trend growth in the ith segment is given by $\beta + \theta_1 + \ldots + \theta_i$. Extensions to higher-order trend polynomials are straightforward: for example, Mills and Crafts (1996b) fit a segmented quadratic trend with three breaks to (the logarithm of) British industrial production for 1700–1913:

$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \sum_{i=1}^{3} \theta_i D_{it}^{(2)} + u_t \tag{7}$$

where

$$D_{it}^{(2)} = \begin{cases} (t - T_i)^2 & t > T_i \\ 0 & \text{otherwise} \end{cases}$$

and with the breaks selected to be at 1776, 1834, and 1874.

As an example of a segmented trend model, Fig. 1 shows the logarithms of UK per capita GDP for 1855–2010 upon which a segmented linear trend with two breaks at 1918 and 1920 has been superimposed. The fitted model is

$$x_t = \underset{(0.0184)}{0.7016} + \underset{(0.0006)}{0.0096}\, t - \underset{(0.0239)}{0.0975}\, D_{1t} + \underset{(0.0243)}{0.1068}\, D_{2t} + \hat{u}_t \qquad R^2 = 0.9927 \tag{8}$$

Figures in parentheses are heteroskedasticity and autocorrelation corrected standard errors, an important proviso here as the estimated cycle $\hat{u}_t$ is undoubtedly serially correlated. The growth rate pre-1919 is estimated to be 0.96% per annum,

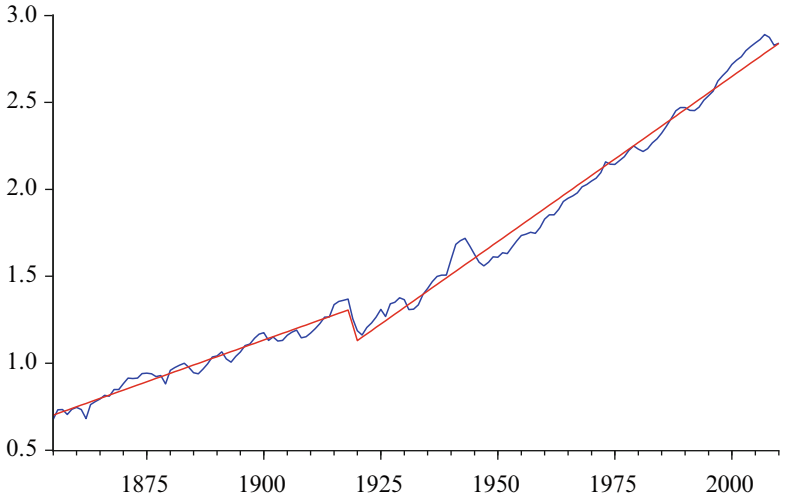


Fig. 1 Logarithms of UK per capita GDP for 1855–2010 with a fitted segmented linear trend having breaks at 1918 and 1920

while post-1920 it is estimated to be $0.0096 - 0.0975 + 0.1068 = 1.89\%$ per annum, with the decline in trend GDP in 1919 and 1920 being of the order of 15%.

On fitting a model of the form Eq. 6, say, various hypotheses may be tested. After the final break, the model becomes

$$x_t = \alpha + \beta t + \sum_{i=2}^{m} \theta_i (t - T_i) + u_t \tag{9}$$

which can be rewritten as

$$x_t = \alpha + \beta t + \left(\sum_{i=2}^{m} \theta_i \right) t - \sum_{i=2}^{m} \theta_i T_i + u_t \tag{10}$$

If both $\sum \theta_i = 0$ and $\sum \theta_i T_i = 0$, the time path of x_t after the final break at T_m will be the same as the path extrapolated from T_1. Crafts and Mills (1996) refer to this as the *Janossy hypothesis* after Janossy (1969), who argued that when the shocks brought about by the world wars and subsequent reconstruction had worked themselves out, growth would return to a historically normal path. If only the first of these restrictions holds, then growth returns to its original (i.e., pre-$T_1 + 1$) rate, so that the path of x_t after T_m will be *parallel* to the path extrapolated from T_1 (Crafts and Mills call this the modified Janossy hypothesis). The hypothesis $\theta_1 + \theta_2 = 0$ is clearly rejected in Eq. 8 so that neither of the Janossy hypotheses holds for the UK, as is clear from Fig. 1.

Of course, the break points in Eq. 8 have been selected a priori and so may be subjectively biased and, furthermore, there may be more than two breaks in trend.

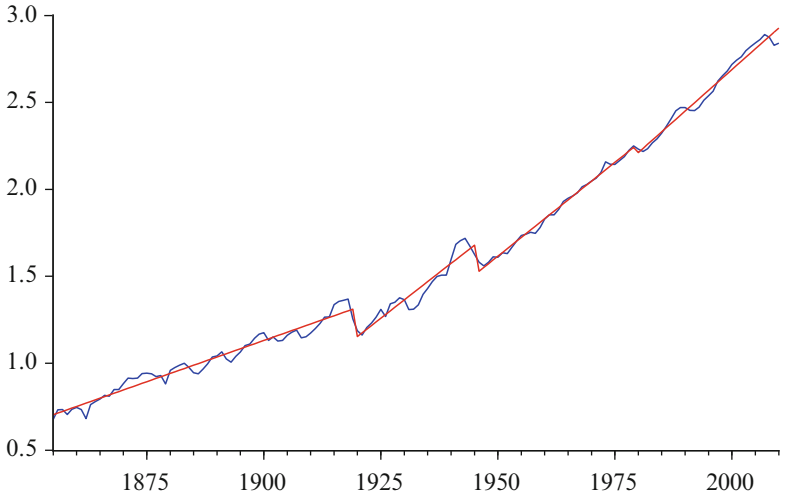


Fig. 2 Logarithms of UK per capita GDP for 1855–2010 with a fitted breaking linear trend having breaks at 1918, 1946, and 1979

Mills and Crafts (1996b) selected their three break dates in Eq. 7 "endogenously" by choosing ranges within which the breaks were most likely to fall and then using a goodness-of-fit criterion based on the fitting of regressions for alternative combinations of break dates. Since then the selection of breaks in such models has been the subject of much interest and research and there are now a variety of procedures with which to determine the number and dating of breaks in trend (key references are Bai 1997; Bai and Perron 1998, 2003a, b; Perron 2006, for a detailed survey of the area).

Figure 2 shows a breaking linear trend fitted to UK per capita GDP containing three breaks at 1919, 1945, and 1979, the number and dating being determined endogenously by several of the extant procedures. The fitted model is

$$\begin{aligned} x_t = {} & \underset{(0.0095)}{0.7039} + \underset{(0.0006)}{0.0095}\, t - \underset{(0.1311)}{0.2112}\, d_{1t} - \underset{(0.0538)}{0.4353}\, d_{2t} - \underset{(0.1158)}{0.7703}\, d_{3t} \\ & + \underset{(0.0017)}{0.0210}\, td_{1t} + \underset{(0.0005)}{0.0216}\, td_{2t} + \underset{(0.0008)}{0.0239}\, td_{3t} + \hat{u}_t \qquad R^2 = 0.9967 \end{aligned}$$

The trend growth rates are estimated to be 0.95% before 1920, 2.10% between 1920 and 1945, 2.16% between 1946 and 1979, and 2.39% from 1980. The Janossy hypotheses are here $\gamma_3 = \delta_3 = 0$ and $\delta_3 = 0$, both of which are clearly rejected. However, a test of the hypothesis $\delta_1 = \delta_2$ is insignificant (marginal level 0.74) so that the hypothesis that trend growth was constant throughout the period 1920 to 1979 may be accepted.

Figure 3 shows the "cycle," obtained by residual, from the segmented trend model. An interesting feature of this component is that, although it may be adequately modelled as a second-order autoregression, being $\psi_t = 1.060\psi_{t-1}$ $_t = 1.060\psi_{t-1} - 0.220\psi_{t-2} + e_t$, the roots of this autoregression are 0.78 and

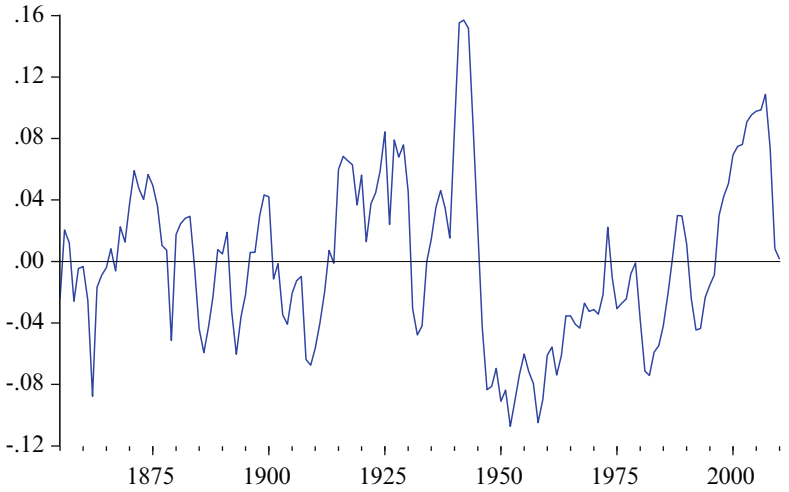


Fig. 3 Cyclical component of UK per capita GDP for 1855–2010 obtained by residual from the segmented trend model

0.28 and hence are both real, so that true "periodic" behavior, in the sense that an average period may be calculated, is ruled out.

All breaking and segmented trend models assume that the cyclical component ψ_t, given by the regression error u_t in Eqs. 5, 6, and 7, is stationary and hence does not contain a unit root. Thus, before these models can be entertained, tests of a unit root in the presence of a breaking trend need to be performed. This area of time series econometrics has attracted much attention recently, with Kim and Perron (2009) and Harris et al. (2009) being notable contributors, but it is not a subject that can be considered in this chapter.

Filters for Extracting Trends and Cycles

Dissatisfaction with the use of unweighted moving averages with preselected spans to extract trends led eventually to the consideration of more flexible moving averages. Their development may be traced back to Leser (1961), although earlier prototypes were proposed in the 1920s (see Mills 2011, Chap. 10).

Leser (1961) implicitly considered the additive decomposition Eq. 1 and invoked the penalized least squares principle, which minimizes, with respect to μ_t, $t = 1, 2, \ldots, T$, the criterion

$$\sum_{t=1}^{T} (x_t - \mu_t)^2 + \lambda \sum_{t=3}^{T} \left(\Delta^2 \mu_t\right)^2 \tag{11}$$

The first term measures the goodness of fit of the trend, and the second penalizes the departure from zero of the variance of the second differences of the trend, so that it is a measure of smoothness: λ is thus referred to as the smoothness parameter. Successive partial differentiation of Eq. 11 with respect to the sequence μ_t leads to the first-order conditions

$$\Delta^2\mu_{t+2} - 2\Delta^2\mu_{t+1} + \Delta^2\mu_t = (\lambda - 1)(x_t - \mu_t)$$

Given T and λ, μ_t will then be a moving average of x_t with time-varying weights, so that no observations are lost at the sample extremes. Leser developed a method of deriving these weights and provided a number of examples in which the solutions were obtained in, it has to be said, laborious and excruciating detail, which must certainly have lessened the impact of the paper at the time!

Some two decades later, Hodrick and Prescott (1997) approached the solution of Eq. 11 rather differently. By recasting Eq. 11 in matrix form as $(\mathbf{x} - \boldsymbol{\mu})'(\mathbf{x} - \boldsymbol{\mu}) + \lambda\boldsymbol{\mu}'\mathbf{D}^{2\prime}\mathbf{D}^2\boldsymbol{\mu}$, where $\mathbf{x} = (x_1, \ldots, x_T)'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_T)'$, and $\mathbf{D}$ is the $T \times T$ "first difference" matrix with elements $d_{t,t} = 1$ and $d_{t-1,t} = -1$ and zero elsewhere, so that $\mathbf{D}\boldsymbol{\mu} = (\mu_2 - \mu_1, \ldots, \mu_T - \mu_{T-1})'$, then differentiating with respect to $\boldsymbol{\mu}$ allows the first-order conditions to be written as

$$\mu = \left(\mathbf{I} + \lambda\mathbf{D}^{2\prime}\mathbf{D}^2\right)^{-1}\mathbf{x} \tag{12}$$

with the rows of the inverse matrix containing the H-P filter weights for estimating the trend μ_t at each t. Setting $\lambda = 100$ is often suggested when extracting a trend from an annual series and other choices are discussed in, for example, Ravn and Uhlig (2002) and Maravall and del Rio (2007).

In filtering terminology the H-P filter Eq. 12 is a *low-pass filter*. To understand this terminology, some basic concepts in filtering theory are useful. Define a *linear filter* of the observed series x_t to be the two-sided weighted moving average

$$y_t = \sum_{j=-n}^{n} a_j x_{t-j} = \left(a_{-n}B^{-n} + a_{-n+1}B^{-n+1} + \ldots + a_0 + \ldots + a_nB^n\right)x_t = a(B)x_t$$

Two conditions are typically imposed upon the filter $a(B)$: (i) that the filter weights either (a) sum to zero, $a(1) = 0$, or (b) sum to unity, $a(1) = 1$, and (ii) that these weights are symmetric, $a_j = a_{-j}$. If condition (ia) holds, then $a(B)$ is a "trend-elimination" filter, whereas if (ib) holds it will be a "trend-extraction" filter. If the former holds then $b(B) = 1 - a(B)$ will be the corresponding trend-extraction filter, having the same, but oppositely signed, weights as the trend-elimination filter $a(B)$ except for the central value, $b_0 = 1 - a_0$, thus ensuring that $b(B) = 1$.

The *frequency response function* of the filter is defined as $a(\omega) = \sum_j e^{-i\omega j}$ for a frequency $0 \leq \omega \leq 2\pi$. The *power transfer function* is then defined as

$$|a(\omega)|^2 = \left(\sum_j a_j \cos \omega j\right)^2 + \left(\sum_j a_j \sin \omega j\right)^2$$

and the *gain* is defined as $|a(\omega)|$, measuring the extent to which the amplitude of the ω–frequency component of x_t is altered through the filtering operation. In general, $a(\omega) = |a(\omega)|e^{-i\theta(\omega)}$, where

$$\theta(\omega) = \tan^{-1} \frac{\sum_j a_j \sin \omega j}{\sum_j a_j \cos \omega j}$$

is the *phase shift*, indicating the extent to which the ω–frequency component of x_t is displaced in time. If the filter is indeed symmetric, then $a(\omega) = a(-\omega)$, so that $a(\omega) = |a(\omega)|$ and $\theta(\omega) = 0$, known as phase neutrality.

With these concepts, an "ideal" low-pass filter has the frequency response function

$$a_L(\omega) = \begin{cases} 1 & \text{if } \omega < \omega_c \\ 0 & \text{if } \omega > \omega_c \end{cases} \tag{13}$$

Thus, $a_L(\omega)$ passes only frequencies lower than the cutoff frequency ω_c, so that just slow-moving, low-frequency components of x_t are retained. Low-pass filters should also be phase-neutral, so that temporal shifts are not induced by filtering. The ideal low-pass filter will take the form

$$a_L(B) = \frac{\omega_c}{\pi} + \sum_{j=1}^{\infty} \frac{\sin \omega_c j}{\pi j}\left(B^j + B^{-j}\right)$$

In practice, low-pass filters will not have the perfect "jump" in $a_L(\omega)$ implied by Eq. 13. The H-P trend-extraction filter, i.e., the one that provides an estimate of the trend component $\hat{\mu}_t = a_{\mathrm{H-P}}(B)x_t$, where the weights are given by Eq. 12, has the frequency response function

$$a_{\mathrm{H-P}}(\omega) = \frac{1}{1 + 4\lambda(1 - \cos \omega)^2} \tag{14}$$

while the H-P trend-elimination filter, which provides the cycle estimate $\hat{\psi}_t = b_{\mathrm{H-P}}(B)x_t = (1 - a_{\mathrm{H-P}}(B))x_t$, has the frequency response function

$$b_{\mathrm{H-P}}(\omega) = 1 - a_{\mathrm{H-P}}(\omega) = \frac{4\lambda(1 - \cos \omega)^2}{1 + 4\lambda(1 - \cos \omega)^2}$$

Rather than setting the smoothing parameter at an a priori value such as $\lambda = 100$, it could also be set at the value that equates the gain to 0.5, i.e., at the value that separates frequencies between those mostly associated with the trend and those mostly associated with the cycle. Since the H-P weights are indeed symmetric, the gain is given by Eq. 14, so equating this to 0.5 yields $\lambda = 1/4(1 - \cos \omega_{0.5})^2$, where $\omega_{0.5}$ is the frequency at which the gain is 0.5 (for more on this idea, see Kaiser and Maravall 2005).

The ideal low-pass filter removes high-frequency components while retaining low-frequency components. A high-pass filter does the reverse, so that the complementary high-pass filter to Eq. 13 has $a_H(\omega) = 0$ if $\omega < \omega_c$ and $a_H(\omega) = 1$ if $\omega \geq \omega_c$. The ideal band-pass filter passes only frequencies in the range $\omega_{c,1} \leq \omega \leq \omega_{c,2}$, so that it can be constructed as the difference between two low-pass filters with cutoff frequencies $\omega_{c,1}$ and $\omega_{c,2}$ and it will have the frequency response function $a_B(\omega) = a_{c,2}(\omega) - a_{c,1}(\omega)$, where $a_{c,2}(\omega)$ and $a_{c,1}(\omega)$ are the frequency response functions of the two low-pass filters, since this will give a frequency response of unity in the band $\omega_{c,1} \leq \omega \leq \omega_{c,2}$ and zero elsewhere. The weights of the band-pass filter will thus be given by $a_{c,2,j} - a_{c,1,j}$, where $a_{c,2,j}$ and $a_{c,1,j}$ are the weights of the two low-pass filters, so that

$$a_B(B) = \frac{\omega_{c,2} - \omega_{c,1}}{\pi} + \sum_{j=1}^{\infty} \frac{\sin \omega_{c,2} j - \sin \omega_{c,1} j}{\pi j}\left(B^j + B^{-j}\right) \tag{15}$$

A conventional definition of the business cycle emphasizes fluctuations of between 1½ and 8 years (see Baxter and King 1999), which leads to $\omega_{c,1} = 2\pi/8 = \pi/4$ and $\omega_{c,2} = 2\pi/1.5 = 4\pi/3$. Thus, a band-pass filter that passes only frequencies corresponding to these periods is defined as $y_t = a_{B,n}(B)x_t$ with weights

$$a_{B,0} = a_{c,2,0} - a_{c,1,0} = \frac{4}{3} - \frac{1}{4} - \left(\zeta_{c,2,n} - \zeta_{c,1,n}\right) \tag{16}$$

$$a_{B,j} = a_{c,2,j} - a_{c,1,j} = \frac{1}{\pi j}\left(\sin \frac{4\pi j}{3} - \sin \frac{\pi j}{4}\right) - \left(\zeta_{c,2,n} - \zeta_{c,1,n}\right) \qquad j = 1, \ldots, n$$

where

$$\zeta_{c,i,n} = -\frac{\sum_{j=-n}^{n} a_{c,i,n}}{2n+1} \qquad i = 1,2$$

The infinite length filter in Eq. 15 has been truncated to have only n leads and lags and the appearance of the $\zeta_{c,i,n}$ terms ensures that the filter weights sum to zero, so that $a_{B,n}(B)$ is a trend-elimination (i.e., cycle) filter. The filter in Eq. 16 is known as the Baxter-King (B-K) filter, with further extensions being provided by Christiano and Fitzgerald (2003).

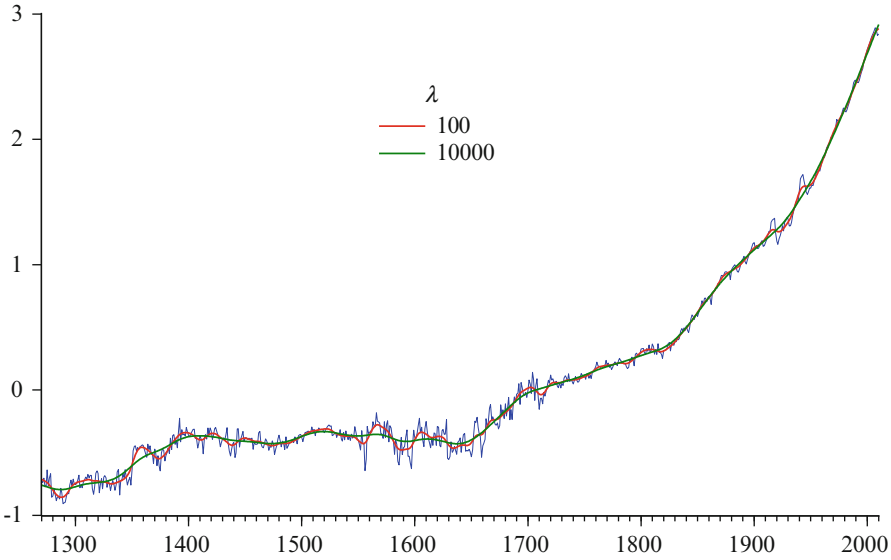


Fig. 4 Logarithms of British per capita GDP for 1270–2010 with H-P trends for $\lambda = 100$ and 10, 000

Figure 4 shows the logarithms of British GDP per capita from 1270 to 2010. Superimposed on this series are H-P trends for $\lambda = 100$ and 10, 000. Figure 5 shows the trend growth rates of these two H-P variants. The larger the value of the smoothing parameter λ, the smoother is the trend, and this is particularly noticeable in the growth rates, where the larger value produces much more stable growth rates than the "conventional" choice of 100 and may thus be more appropriate for examining trend growth over long historical time spans.

Filters and Structural Models

Several of the filters in common use can be shown to be optimal for the following class of *structural unobserved component* (UC) models:

$$x_t = \mu_t + \psi_t$$

$$\Delta^m \mu_t = (1+B)^r \xi_t \qquad \xi_t \sim WN\left(0, \sigma_\xi^2\right)$$

$$\psi_t \sim WN\left(0, \lambda\sigma_\xi^2\right) \qquad E\left(\xi_t \psi_{t-j}\right) = 0 \quad \text{for all } j$$

Here the notation $y_t \sim WN(0, \sigma_y^2)$ is to be read as stating that the variable y_t is white noise (i.e., identically and independently distributed) with zero mean and variance σ_y^2. For (doubly) infinite samples, the minimum mean square error

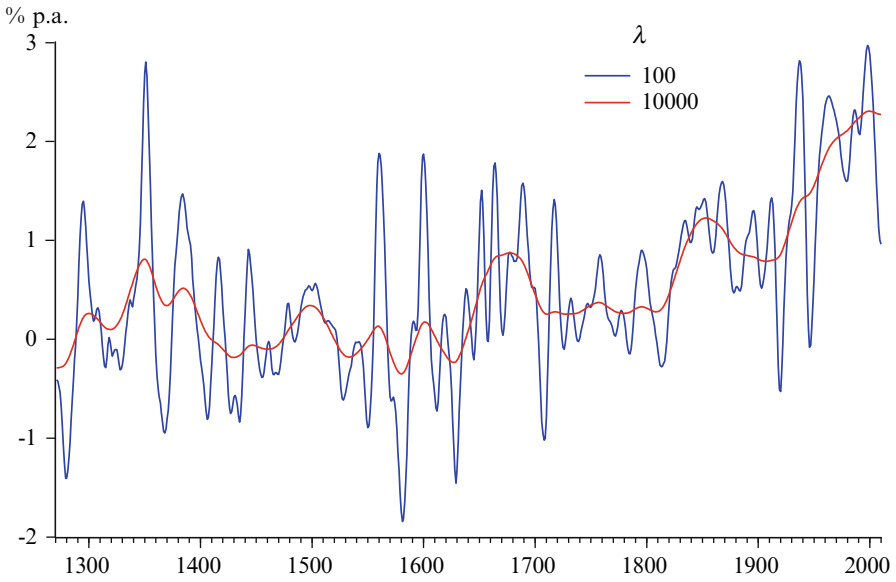


Fig. 5 Trend growth rates of British per capita GDP for $\lambda = 100$ and 10, 000

(MMSE) estimates of the components are $\hat{\mu}_t = a_\mu(B)x_t$ and $\hat{\psi}_t = x_t - \hat{\mu}_t = \big(1 - a_\mu(B)\big)x_t = a_\psi(B)x_t$, where

$$a_\mu(B) = \frac{(1+B)^r}{(1+B)^r + (1-B)^m}$$

and

$$\big|a_\mu(B)\big| = \frac{|1+B|^{2r}}{|1+B|^{2r} + \lambda|1-B|^{2m}} \tag{17}$$

and the notation $|\alpha(B)| = \alpha(B)\alpha(B^{-1})$ is used. This result uses Wiener-Kolmogorov filtering theory and its derivation may be found in, for example, Proietti (2009a). This filter is therefore defined by the order of integration of the trend, m, which regulates its flexibility, by the parameter r (which technically is the number of unit poles at the Nyquist frequency and which thus regulates the smoothness of $\Delta^m\mu_t$) and by λ, which measures the relative variance of the noise component.

The H-P filter is obtained for $m = 2$ and $r = 0$, so that $\Delta^2\mu_t = \xi_t$. If $m = 1$ and $r = 0$, $\Delta\mu_t = \xi_t$ and the filter corresponds to a two-sided exponentially weighted moving average with smoothing parameter $\big((1+2\lambda) + \sqrt{1+4\lambda}\big)/2\lambda$ (recall Cox 1961). If $r = 0$ then any setting of m defines a *Butterworth filter*, as does setting $m = r$, which is known as the Butterworth *square-wave* filter (see Gómez 2001).

Setting $m = r = 1$ and $\lambda = 1$ produces the multi-resolution Haar scaling and wavelet filters (see Percival and Walden 1999).

Using Eq. 17 and the idea that the cutoff frequency can be chosen to be that frequency at which the gain is 0.5 (as above) enables the smoothing parameter to be determined as

$$\lambda = 2^{r-m} \frac{(1 + \cos \omega_{0.5})^r}{(1 - \cos \omega_{0.5})^m}$$

Detailed development of the models and techniques discussed in this and the preceding section may be found in Pollock (2009) and Proietti (2009a).

Model-Based Filters

The setup of the previous section is very assumption laden and implies, among other things, that the observed series is generated as

$$\Delta^m x_t = (1 + B)^r \zeta_t + (1 - B)^m \psi_t = \theta_q(B) a_t$$

i.e., as a heavily restricted ARIMA $(0, m, q)$ process, where $\theta_q(B) = 1 - \theta_1 B - {}_q(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$ with $q = \max(r, m)$ (the subscript q denoting the order of the polynomial will be dropped when appropriate to simplify notation). A less restrictive approach is to begin by assuming that the observed series has an ARIMA (p, d, q) representation

$$\phi_p(B)\left(\Delta^d x_t - c\right) = \theta_q(B) a_t \qquad a_t \sim WN\left(0, \sigma_a^2\right)$$

where $\phi_p(B)$ has p stationary roots and $\theta_q(B)$ is invertible and to derive filters with the desired properties from this representation. This is done by exploiting the idea that a_t can be decomposed into two orthogonal stationary processes (see, e.g., Proietti 2009a, b, for technical details):

$$a_t = \frac{(1 + B)^r \zeta_t + (1 - B)^m \kappa_t}{\phi_{q^*}(B)} \tag{18}$$

where $q^* = \max(r, m)$, $\zeta_t \sim WN(0, \sigma_a^2)$, $\kappa_t \sim WN(0, \lambda \sigma_a^2)$, and

$$\left|\phi_{q^*}(B)\right|^2 = |1 + B|^{2r} + \lambda |1 - B|^{2m} \tag{19}$$

Given Eqs. 18 and 19, the following orthogonal trend-cycle decomposition $x_t = \mu_t + \psi_t$ can be defined:

$$\phi(B)\varphi(B)\left(\Delta^d \mu_t - c\right) = (1+B)^r \theta(B)\zeta_t \tag{20}$$

$$\phi(B)\varphi(B)\psi_t = \Delta^{m-d}\theta(B)\kappa_t$$

The trend, or low-pass component, has the same order of integration as x_t, regardless of m, whereas the cycle, or high-pass component, is stationary provided that $m \geq d$. The MMSE estimators of the trend and cycle are again given by Eq. 17 and its "complement." Band-pass filters may be constructed by decomposing the low-pass component in Eq. 20: see Proietti (2009a).

The H-P and B-K filters are often referred to as being ad hoc, in the sense here that they are invariant to the process actually generating x_t. This has the potential danger that such filters could produce a cyclical component, say, that might display cyclical features that are absent from the observed series, something that is known as the Slutsky-Yule effect. For example, it has been well documented that when the H-P filter is applied to a random walk, which obviously cannot contain any cyclical patterns, the detrended series can nevertheless display spurious cyclical behavior. The (ARIMA) model-based filters are designed to overcome these limitations.

Structural Trends and Cycles

An alternative approach to modelling trends and cycles is to take the UC decomposition $x_t = \mu_t + \psi_t$ and *assume* particular models for the components. The most general approach to *structural model* building is that set out by Harvey and Trimbur (2003), Trimbur (2006), and Harvey et al. (2007), who consider the UC decomposition

$$x_t = \mu_{m,t} + \psi_{n,t}$$

where the components are assumed to be mutually uncorrelated. The trend component is defined as the mth order stochastic trend

$$\mu_{1,t} = \mu_{1,t-1} + \zeta_t \qquad \zeta_t \sim WN\left(0,\sigma_\zeta^2\right)$$

$$\mu_{i,t} = \mu_{i,t-1} + \mu_{i-1,t} \qquad i = 2,\ldots,m$$

Note that repeated substitution yields $\Delta^m \mu_{m,t} = \zeta_t$. The random walk trend is thus obtained for $m = 1$ and the integrated random walk, or "smooth trend," with slope $\mu_{1,t}$, for $m = 2$.

The component $\psi_{n,t}$ is an nth order stochastic cycle, for $n > 0$, if

$$\begin{bmatrix} \psi_{1,t} \\ \psi_{1,t}^* \end{bmatrix} = \rho \begin{bmatrix} \cos\varpi & \sin\varpi \\ -\sin\varpi & \cos\varpi \end{bmatrix} \begin{bmatrix} \psi_{1,t-1} \\ \psi_{1,t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ 0 \end{bmatrix} \qquad \kappa_t \sim WN\left(0,\sigma_\kappa^2\right) \tag{21}$$

$$\begin{bmatrix} \psi_{i,t} \\ \psi^*_{i,t} \end{bmatrix} = \rho \begin{bmatrix} \cos\varpi & \sin\varpi \\ -\sin\varpi & \cos\varpi \end{bmatrix} \begin{bmatrix} \psi_{i,t-1} \\ \psi^*_{i,t-1} \end{bmatrix} + \begin{bmatrix} \psi_{i-1,t} \\ 0 \end{bmatrix} \qquad i = 2,\ldots,n$$

Here $0 \leq \varpi \leq \pi$ is the frequency of the cycle and $0 < \rho \leq 1$ is the damping factor. The reduced form representation of the cycle is

$$\left(1 - 2\rho\cos\varpi B + \rho^2 B^2\right)^n \psi_{n,t} = (1 - \rho\cos\varpi B)^n \kappa_t$$

and Harvey and Trimbur (2003) show that, as m and n increase, the optimal estimates of the trend and cycle approach the ideal low-pass and band-pass filters, respectively. Defining the "signal to noise" variance ratios $q_\zeta = \sigma^2_\zeta/\sigma^2_\varepsilon$ and $q_\kappa = \sigma^2_\kappa/\sigma^2_\varepsilon$, the low-pass filter (of order m, n) is

$$\hat{\mu}_t(m,n) = \frac{q_\zeta/|1-B|^{2m}}{q_\zeta/|1-B|^{2m} + q_k|c(B)|^n + 1}$$

where $c(B) = (1 - \rho\cos\varpi B)/(1 - 2\rho\cos\varpi B + \rho^2 B^2)$. The corresponding band-pass filter is

$$\hat{\psi}_t(m,n) = \frac{q_k|c(B)|^n}{q_\zeta/|1-B|^{2m} + q_k|c(B)|^n + 1}$$

Harvey and Trimbur (2003) discuss many of the properties of this general model. They note that applying a band-pass filter of order n to a series that has been detrended by a low-pass filter of order m will not give the same result as applying a generalized filter of order (m, n), as a jointly specified model enables trends and cycles to be extracted by filters that are mutually consistent. Using higher-order trends with a fixed order band-pass filter has the effect of removing lower frequencies from the cycle. However, setting m greater than 2 will produce trends that are more responsive to short-term movements than is perhaps desirable, and this might be felt to be a particular drawback in historical applications.

Replacing the zero component in the right hand side of Eq. 21 by a white noise uncorrelated with κ_t produces a *balanced cycle*, the statistical properties of which are derived in Trimbur (2006). For example, for $n = 2$ the variance of the cycle is given by

$$\sigma^2_\psi = \frac{1+\rho^2}{(1-\rho^2)^3}\sigma^2_\kappa$$

as opposed to $\sigma^2_\kappa/(1-\rho^2)$ for the first-order case, while its autocorrelation function is

$$\rho_2(\tau) = \rho^\tau \cos \varpi\tau\left(1 + \frac{1-\rho^2}{1+\rho^2}\tau\right), \qquad \tau = 0,\ 1,\ 2, \ldots$$

compared to $\rho_1(\tau) = \rho^\tau \cos \varpi\tau$. Harvey and Trimbur prefer the balanced form as it seems to give better fits in empirical applications and offers computational advantages over Eq. 21.

Trimbur (2006) shows that an nth order stochastic cycle admits an ARMA $(2n, 2n - 1)$ representation in which the AR polynomial has n pairs of roots, each given by the complex conjugate pair $\rho^{-1} \exp(\pm i\varpi)$.

If the cyclical component is not characterized by such "cyclical" behavior, the specification Eq. 21 may be replaced by a simple AR(1) or AR(2) process. Such a model with $m = 2$ and an AR(1) cycle is found to be the best structural model with which to characterize British per capita GDP for 1270–2010:

$$x_t = \mu_{2,t} + \psi_t$$

$$\mu_{2,t} = \mu_{2.t-1} + \mu_{1,t} = \mu_{2.t-1} + \mu_{1,t-1} + \zeta_t \qquad \hat{\sigma}^2_\zeta = 0.00058$$

$$\psi_t = 0.396\psi_{t-1} + e_t \qquad \hat{\sigma}^2_e = 0.00241$$

The trend growth obtained from this model is shown in Fig. 6 along with trend growth computed from the H-P filter with $\lambda = 10,000$. The latter is seen to be a smoothed version of the former, which might be thought to be too volatile to be considered as a viable estimate for "long-run" trend growth.

Models with Correlated Components

A feature of all the models introduced so far has been the identifying assumption that all component innovations are mutually uncorrelated, so that the components are orthogonal. Such an assumption can be relaxed: for example, Morley et al. (2003) consider the UC model $x_t = \mu_t + \psi_t$ with contemporaneously correlated innovations:

$$\mu_t = \mu_{t-1} + c + \zeta_t \qquad \zeta_t \sim WN\left(0, \sigma^2_\zeta\right) \tag{22}$$

$$\psi_t = \phi_1\psi_{t-1} + \phi_2\psi_{t-2} + \kappa_t \qquad \kappa_t \sim WN\left(0, \sigma^2_\kappa\right)$$

with $\sigma_{\zeta\kappa} = E(\zeta_t\kappa_t) = r\sigma_\zeta\sigma_\kappa$, so that r is the contemporary correlation between the innovations. The reduced form of Eq. 22 is the ARIMA (2, 1, 2) process

$$\left(1 - \phi_1 B - \phi_2 B^2\right)(\Delta x_t - c) = \left(1 - \theta_1 B - \theta_2 B^2\right)a_t \tag{23}$$

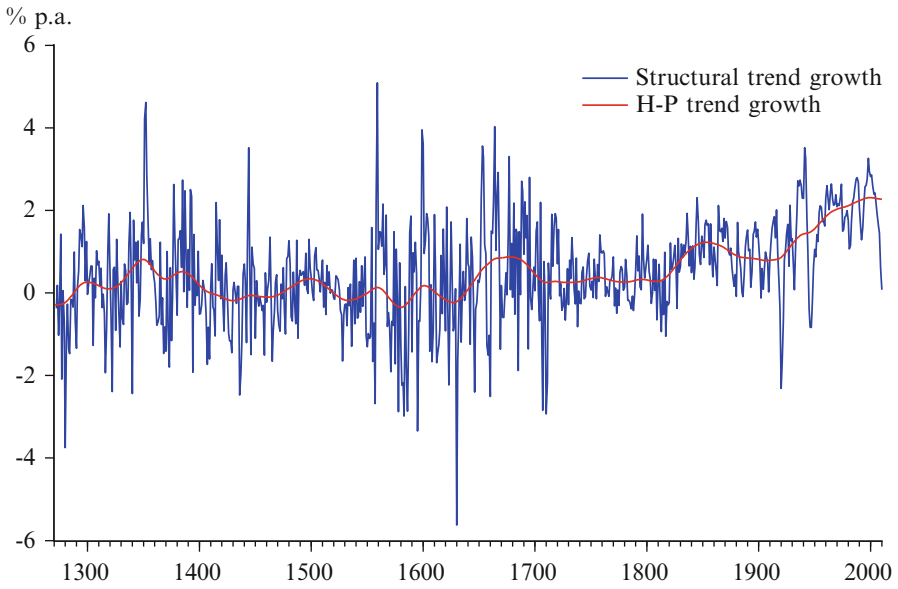


Fig. 6 Trend growth for British per capita GDP, 1270–2010, computed from a structural model, compared to H-P trend growth

Morley et al. (2003) show that the structural form is exactly identified, so that the correlation between the innovations can be estimated and the orthogonality assumption $\sigma_{\zeta\kappa} = 0$ tested.

A related model decomposes an ARIMA(p, 1, q) process $\phi(B)(\Delta x_t - c) = \theta(B)\, a_t$ into a random walk trend

$$\mu_t = \mu_{t-1} + c + \frac{\theta(1)}{\phi(1)} a_t = \frac{\theta(1)}{\phi(1)} \frac{\phi(B)}{\theta(B)} x_t$$

and a cyclical (or transitory) component

$$\psi_t = \frac{\phi(1)\theta(B) - \theta(1)\phi(B)}{\phi(1)\phi(B)\Delta} a_t = \frac{\phi(1)\theta(B) - \theta(1)\phi(B)}{\phi(1)\phi(B)} x_t$$

which has a stationary ARMA (p, max(p, q) − 1) representation. Thus, for the ARIMA (2, 1, 2) process (Eq. 23), the random walk trend will be

$$\mu_t = \mu_{t-1} + c + \left(\frac{1 - \theta_1 - \theta_2}{1 - \phi_1 - \phi_2} \right) a_t$$

while the ARMA (2, 1) cycle will be

$$\left(1-\phi_1 B-\phi_2 B^2\right)\psi_t=(1+\vartheta B)\left(\frac{\theta_1+\theta_2-(\phi_1+\phi_2)}{1-\phi_1-\phi_2}\right)a_t$$

$$\vartheta=\frac{\phi_2(1-\theta_1-\theta_2)+\theta_2(1-\phi_1-\phi_2)}{\theta_1+\theta_2-(\phi_1+\phi_2)}$$

The two components are seen to be driven by the *same* innovation, a_t, and hence are perfectly correlated. Whether this correlation is plus or minus one depends upon the *persistence* $\theta(1)/\phi(1)$: if this is less (greater) than one, the correlation is + 1 (−1). This decomposition is familiarly known as the Beveridge-Nelson (B-N) decomposition (Beveridge and Nelson 1981).

The following ARIMA (1, 1, 2) model was found to adequately fit the British per capita GDP data:

$$\Delta x_t=\underset{(0.114)}{0.486}+\frac{\left(1-\underset{(0.037)}{0.277}\,B^2\right)}{\left(1+\underset{(0.037)}{0.303}\,B\right)}a_t$$

Since $p=1$ the structural form (22) is no longer identified, but the B-N decomposition may be obtained: with $\phi_1=-0.303$, $\theta_2=0.277$, and $\phi_2=\theta_1=0$, the B-N random walk trend is

$$\mu_t=\mu_{t-1}+0.486+0.554a_t$$

and the ARMA(1,1) cycle is

$$(1+0.303B)\psi_t=(1+0.622B)(0.446a_t)$$

Proietti and Harvey (2000) define the B-N "smoother" as

$$\mu_t^{\mathrm{B-N}}=\left[\frac{\theta(1)}{\phi(1)}\right]^2\frac{\phi(B)\phi\left(B^{-1}\right)}{\theta(B)\theta\left(B^{-1}\right)}x_t$$

which will be a symmetric two-sided filter with weights summing to unity. For the model here

$$\mu_t^{\mathrm{B-N}}=0.307\frac{(1+0.303B)\left(1+0.303B^{-1}\right)}{\left(1-0.277B^2\right)\left(1-0.277B^{-2}\right)}x_t$$

Multivariate Extensions of Structural Models

Since co-movement between macroeconomic series is a key aspect of business cycles, many of the filters and structural models have been extended to multivariate setups (see, e.g., Kozicki 1999). Multivariate structural models have been introduced

by Carvalho and Harvey (2005) and Carvalho et al. (2007). Suppose there are N time series gathered together in the vector $\mathbf{x}_t = (x_{1t}, \ldots, x_{Nt})'$, which may be decomposed into trend, $\boldsymbol{\mu}_t$; cycle, $\boldsymbol{\psi}_t$; and irregular, $\boldsymbol{\varepsilon}_t$, vectors such that

$$\mathbf{x}_t = \boldsymbol{\mu}_t + \boldsymbol{\psi}_t + \boldsymbol{\varepsilon}_t \qquad \boldsymbol{\varepsilon}_t \sim MWN(\mathbf{0}, \Sigma_{\boldsymbol{\varepsilon}})$$

where $MWN(\mathbf{0}, \Sigma_{\boldsymbol{\varepsilon}})$ denotes zero mean multivariate white noise with $N \times N$ positive semi-definite covariance matrix $\Sigma_{\boldsymbol{\varepsilon}}$. The trend is defined as

$$\begin{aligned} \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t-1} + \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim MWN\left(\mathbf{0}, \Sigma_{\boldsymbol{\eta}}\right) \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\zeta}_t \qquad \boldsymbol{\zeta}_t \sim MWN\left(\mathbf{0}, \Sigma_{\boldsymbol{\zeta}}\right) \end{aligned} \tag{24}$$

With $\Sigma_{\boldsymbol{\zeta}} = \mathbf{0}$ and $\Sigma_{\boldsymbol{\eta}}$ positive definite, each trend is a random walk with drift. If, on the other hand, $\Sigma_{\boldsymbol{\eta}} = \mathbf{0}$ and $\Sigma_{\boldsymbol{\zeta}}$ is positive definite, the trends are integrated random walks and will typically be much smoother than drifting random walks.

The *similar cycle* model is

$$\begin{bmatrix} \boldsymbol{\psi}_t \\ \boldsymbol{\psi}_t^* \end{bmatrix} = \left[\rho \begin{pmatrix} \cos\varpi & \sin\varpi \\ -\sin\varpi & \cos\varpi \end{pmatrix} \otimes \mathbf{I}_N \right] \begin{bmatrix} \boldsymbol{\psi}_{t-1} \\ \boldsymbol{\psi}_{t-1}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\kappa}_t \\ \boldsymbol{\kappa}_t^* \end{bmatrix}$$

where $\boldsymbol{\psi}_t$ and $\boldsymbol{\psi}_t^*$ are N-vectors and $\boldsymbol{\kappa}_t$ and $\boldsymbol{\kappa}_t^*$ are N-vectors of mutually uncorrelated zero mean vector multivariate white noise with the same covariance matrix $\Sigma_{\boldsymbol{\kappa}}$. As the damping factor ρ and cyclical frequency ϖ are the same for all series, the individual cycles have similar properties, being centered around the same period as well as being contemporaneously correlated, on noting that the covariance matrix of $\boldsymbol{\psi}_t^*$ is

$$\Sigma_{\boldsymbol{\psi}} = \left(1 - \rho^2\right)^{-1} \Sigma_{\boldsymbol{\kappa}}$$

Suppose that $\Sigma_{\boldsymbol{\zeta}} = \mathbf{0}$ in Eq. 24. The model will have common trends if $\Sigma_{\boldsymbol{\eta}}$ is less than full rank. If the rank of $\Sigma_{\boldsymbol{\eta}}$ is one, then there will be a single common trend and

$$\mathbf{x}_t = \boldsymbol{\theta}\mu_t + \boldsymbol{\alpha} + \boldsymbol{\psi}_t + \boldsymbol{\varepsilon}_t \tag{25}$$

where the common trend is

$$\mu_t = \mu_{t-1} + \beta + \eta_t \qquad \mu_0 = 0 \qquad \eta_t \sim WN\left(0, \sigma_\eta^2\right)$$

and $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are N-vectors of constants. If $\Sigma_{\boldsymbol{\eta}} = \mathbf{0}$ the existence of common trends depends on the rank of $\Sigma_{\boldsymbol{\zeta}}$. When this rank is less than N, some linear combinations of the series will be stationary. A rank of one again leads to the model (25) but with

$$\mu_t = \mu_{t-1} + \beta_{t-1} \qquad \beta_t = \beta_{t-1} + \zeta_t \qquad \zeta_t \sim WN\left(0, \sigma_\zeta^2\right)$$

When $\boldsymbol{\theta} = \mathbf{i}$, where $\mathbf{i}$ is an N-vector of ones, there is *balanced growth*, and the difference between any pair of series in $\mathbf{x}_t$ is stationary.

A mechanism for capturing convergence to a common growth path can be incorporated by specifying the decomposition

$$\mathbf{x}_t = \boldsymbol{\alpha} + \boldsymbol{\mu}_t + \boldsymbol{\psi}_t + \boldsymbol{\varepsilon}_t$$

with

$$\boldsymbol{\mu}_t = \boldsymbol{\Phi}\boldsymbol{\mu}_{t-1} + \boldsymbol{\beta}_{t-1} \qquad \boldsymbol{\beta}_t = \boldsymbol{\Phi}\boldsymbol{\beta}_{t-1} + \boldsymbol{\zeta}_t$$

where $\boldsymbol{\Phi} = \phi\mathbf{I} + (1-\phi)\mathbf{i}\overline{\boldsymbol{\phi}}$ and $\overline{\boldsymbol{\phi}}$ is a vector of weights. With this setup, a convergence mechanism can be defined to operate on both the gap between an individual series and the common trend and on the gap in the growth rates of the individual series and the common trend. When ϕ is less than but close to unity, the convergence components tend to be quite smooth and there is a clear separation of long-run movements and cycles. The forecasts for each series converge to a common growth path, although they may exhibit temporary divergences. If $\phi = 1$ there will be no convergence.

Extensions to incorporate multivariate mth order trends and nth order cycles may also be contemplated, as indeed may multivariate low-pass and band-pass filters.

Estimation of Structural Models

All the structural models introduced here may be estimated by recasting them in state space form, whence they can be estimated using the Kalman filter algorithm. This will produce a MMSE estimator of the state vector, along with its mean square error matrix, conditional on past information. This is then used to build the one-step-ahead predictor of $\mathbf{x}_t$ and its mean square error matrix. The likelihood of the model can be evaluated via the prediction error decomposition and both filtered (real time) and smoothed (full sample) estimates of the components may then be obtained using a set of recursive equations. Harvey and De Rossi (2006) and Proietti (2009a) are convenient references for technical details, while comprehensive software for the estimation and analysis of structural models is provided by the STAMP package (see Koopman et al 2009).

Structural Breaks Across Series

While section "Segmented Trend Models" considered breaks in a single series, the focus of this section is on the recently developed modelling of breaks across a set of time series, known as *co-breaking*, as synthesized by Hendry and Massmann (2007). Their basic definition of co-breaking again focuses on the vector $\mathbf{x}_t = (x_{1t}, \ldots, x_{Nt})'$, which is now assumed to have an unconditional expectation around an initial parameterization of $E(\mathbf{x}_0) = \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ depends only on deterministic variables

whose parameters do not change: for example, $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{c,0} + \boldsymbol{\beta}_{t,0}t$. A *location shift* in $\mathbf{x}_t$ is then said to occur if, for any t, $E(\mathbf{x}_t - \boldsymbol{\beta}_0) = \boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_t \neq \boldsymbol{\beta}_{t-1}$, i.e., if the expected value of $\mathbf{x}_t$ around its initial parameterization in one time period deviates from that in the previous time period. (Contemporaneous mean) Co-breaking is then defined as the cancelation of location shifts across linear combinations of variables and may be characterized by there being an $n \times r$ matrix $\boldsymbol{\Omega}$, of rank $r < n$, such that $\boldsymbol{\Omega}'\boldsymbol{\beta}_t = \mathbf{0}$. It then follows that $\boldsymbol{\Omega}'E(\mathbf{x}_t - \boldsymbol{\beta}_0) = \boldsymbol{\Omega}'\boldsymbol{\beta}_t = \mathbf{0}$, so that the parameterization of the r co-breaking relationships $\boldsymbol{\Omega}'\mathbf{x}_t$ is independent of the location shifts.

Various extensions of contemporaneous mean co-breaking may be considered, such as variance co-breaking and intertemporal mean co-breaking, defined as the cancelation of deterministic shifts across both variables and time periods. Co-breaking may also be related to cointegration: the "common trends" incorporated in a VECM can be shown to be equilibrium-mean co-breaking, while the cointegrating vector itself is drift co-breaking.

To formalize the co-breaking regression approach, consider the following regression model for $\mathbf{x}_t$:

$$\mathbf{x}_t = \boldsymbol{\pi}_0 + \boldsymbol{\kappa}\mathbf{d}_t + \boldsymbol{\delta}\mathbf{w}_t + \boldsymbol{\varepsilon}_t \tag{26}$$

where $\mathbf{w}_t$ is a vector of exogenous variables and $\mathbf{d}_t$ is a set of $k > n$ deterministic shift variables. Assuming that the rank of $\boldsymbol{\kappa}$ is $n - 1$ allows it to be decomposed as $\boldsymbol{\kappa} = \boldsymbol{\xi}\boldsymbol{\eta}'$, where $\boldsymbol{\xi}$ is $n \times (n-1)$, $\boldsymbol{\eta}$ is $k \times (n-1)$, and both $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are of full rank $n - 1$. There will then exist an $n \times 1$ vector $\boldsymbol{\xi}_\perp$ such that $\boldsymbol{\xi}'_\perp\boldsymbol{\xi} = \mathbf{0}$, which then implies that the linear combination $\boldsymbol{\xi}'_\perp\mathbf{x}_t = \boldsymbol{\xi}'_\perp\boldsymbol{\pi}_0 + \boldsymbol{\xi}'_\perp\boldsymbol{\delta}\mathbf{w}_t + \boldsymbol{\xi}'_\perp\boldsymbol{\varepsilon}_t$ will not contain the shift variables $\mathbf{d}_t$. Partitioning $\mathbf{x}_t$ as $(y_t \ \vdots \ \mathbf{z}_t)$ and partitioning and normalizing $\boldsymbol{\xi}_\perp$ as $(1 \ \vdots \ -\boldsymbol{\xi}'_{\perp,1})$ define the structural break-free co-breaking regression

$$y_t = \boldsymbol{\xi}'_{\perp,1}\mathbf{z}_t + \tilde{\boldsymbol{\pi}}_0 + \tilde{\boldsymbol{\delta}}\boldsymbol{w}_t + \tilde{\boldsymbol{\varepsilon}}_t \tag{27}$$

where $\tilde{\boldsymbol{\pi}}_0 = \boldsymbol{\xi}'_\perp\boldsymbol{\pi}_0$, etc. This co-breaking regression procedure may be implemented in two steps. First, test whether the k shifts $\mathbf{d}_t$ are actually present in each of the n components of $\mathbf{x}_t$ by estimating Eq. 26 and testing for the significance of $\boldsymbol{\kappa}$: second, augment Eq. 27 by $\mathbf{d}_t$ and test whether the shifts are now insignificant, with the co-breaking vector either estimated or imposed. Various extensions of this basic approach are discussed by Hendry and Massmann (2007), who also relax the assumption that the number of co-breaking relationships is known (assumed to be one above), so that the rank of $\boldsymbol{\kappa}$ is $n - r$, where r is to be estimated. Although such procedures are still in their infancy, they represent an important advance in the co-breaking framework, in which linear combinations of the form $\boldsymbol{\Omega}'\mathbf{x}_t$ depend on fewer breaks in their deterministic components than does $\mathbf{x}_t$ on its own.

Concluding Remarks

While cycles have been shown to be relatively straightforward to define, there is much less consensus on what actually constitutes a trend and trying to pin this down has attracted some attention recently for, as Phillips (2005) has memorably

remarked, "no one understands trends, but everyone sees them in the data" and that to "capture the random forces of change that drive a trending process, we need sound theory, appropriate methods, and relevant data. In practice, we have to manage under shortcomings in all of them." The variety of trend estimation methods discussed in this chapter would seem to bear Phillips out.

White and Granger (2011) have set out "working definitions" of various kinds of trends, and this taxonomy may prove to be useful in developing further models of trending processes, in which they place great emphasis on "attempting to relate apparent trends to appropriate underlying phenomena, whether economic, demographic, political, legal, technological, or physical." This would surely require taking account of possible co-breaking phenomena of the type discussed in the previous section as well, so producing a richer class of multivariate models for trending and breaking processes. It is thus clear that the modelling of trends and cycles will continue to be a key area of research in time series econometrics for some time to come and that any new developments should become part of the cliometrician's tool kit for analyzing historical time series.

References

Aldcroft DH, Fearon P (1972) Introduction. In: Aldcroft DH, Fearon P (eds) British economic fluctuations, 1790–1939. Macmillan, London, pp 1–73

Bai J (1997) Estimating multiple breaks one at a time. Econom Theory 13:315–352

Bai J, Perron P (1998) Estimating and testing linear models with multiple structural changes. Econometrica 66:47–78

Bai J, Perron P (2003a) Computation and analysis of multiple structural change models. J Appl Econom 18:1–22

Bai J, Perron P (2003b) Critical values for multiple structural change tests. Econom J 6:72–78

Baxter M, King RG (1999) Measuring business cycles: approximate band-pass filters for economic time series. Rev Econ Stat 81:575–593

Beveridge S, Nelson CR (1981) A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the "business cycle". J Monet Econ 7:151–174

Broadberry S, Campbell B, Klein A, Overton M, van Leeuwn B (2011) British economic growth, 1270–1870: an output based approach. LSE, London

Carvalho V, Harvey AC (2005) Growth, cycles and convergence in US regional time series. Int J Forecast 21:667–686

Carvalho V, Harvey AC, Trimbur TM (2007) A note on common cycles, common trends and convergence. J Bus Econ Stat 25:12–20

Christiano L, Fitzgerald T (2003) The band pass filter. Int Econ Rev 44:435–465

Cox DR (1961) Prediction by exponentially weighted moving averages and related methods. J R Stat Soc Ser B 23:414–422

Crafts NFR, Mills TC (1994a) The industrial revolution as a macroeconomic epoch: an alternative view. Econ Hist Rev 47:769–775

Crafts NFR, Mills TC (1994b) Trends in real wages in Britain, 1750–1913. Explor Econ Hist 31:176–194

Crafts NFR, Mills TC (1996) Europe's golden age: an econometric investigation of changing trend rates of growth. In: van Ark B, Crafts NFR (eds) Quantitative aspects of Europe's postwar growth. Cambridge University Press, Cambridge, pp 415–431
Crafts NFR, Mills TC (1997) Endogenous innovation, trend growth and the British industrial revolution. J Econ Hist 57:950–956
Crafts NFR, Mills TC (2004) After the industrial revolution: the climacteric revisited. Explor Econ Hist 41:156–171
Crafts NFR, Leybourne SJ, Mills TC (1989a) Trends and cycles in U.K. industrial production: 1700–1913. J R Stat Soc Ser A 152:43–60
Crafts NFR, Leybourne SJ, Mills TC (1989b) The climacteric in late victorian Britain and France: a reappraisal of the evidence. J Appl Econom 4:103–117
Feinstein CH, Matthews RCO, Odling-Smee JC (1982) The timing of the climacteric and its sectoral incidence in the UK. In: Kindleberger, CP, di Tella, G (eds) Economics of the Long View, volume 2, part 1, Clarendon Press, Oxford, pp 168–185
Ford AG (1969) British economic fluctuations, 1870–1914. Manch Sch 37:99–129
Ford AG (1981) The trade cycle in Britain 1860–1914. In: Floud RC, McCloskey DN (eds) The economic history of Britain since 1700. Cambridge University Press, Cambridge, pp 27–49
Frickey E (1947) Production in the USA, 1860–1914. Harvard University Press, Cambridge, MA
Gómez V (2001) The use of Butterworth filters for trend and cycle estimation in economic time series. J Bus Econ Stat 19:365–373
Harris D, Harvey DI, Leybourne SJ, Taylor AMR (2009) Testing for a unit root in the presence of a possible break in trend. Econom Theory 25:1545–1588
Harvey AC, De Rossi P (2006) Signal extraction. In: Mills TC, Patterson K (eds) Palgrave handbook of econometrics: volume 1, econometric theory, 970–1000, Palgrave Macmillan, Basingstoke, pp 970–1000
Harvey AC, Trimbur TM (2003) General model-based filters for extracting cycles and trends in economic time series. Rev Econ Stat 85:244–255
Harvey AC, Trimbur TM, van Dijk HK (2007) Trends and cycles in economic time series: a Bayesian approach. J Econom 140:618–649
Hendry DF, Massmann M (2007) Co-breaking: recent advances and a synopsis of the literature. J Bus Econ Stat 25:33–51
Hodrick RJ, Prescott EC (1997) Postwar U.S. business cycles: an empirical investigation. J Money Credit Bank 29:1–16
Hoffman WG (1955) British industry, 1700–1950. Blackwell, Oxford
Hooker RH (1901) Correlation of the marriage rate with trade. J R Stat Soc 64:485–492
Janossy F (1969) The end of the economic miracle. IASP, White Plains
Kaiser R, Maravall A (2005) Combining filter design with model-based filtering (with an application to business cycle estimation). Int J Forecast 21:691–710
Kalman RE (1960) A new approach to linear filtering and prediction theory. J Basic Eng Trans ASME Ser D 82:35–45
Kalman RE, Bucy RE (1961) New results in linear filtering and prediction theory. J Basic Eng Trans ASME Ser D 83:95–108
Kim D, Perron P (2009) Unit root tests allowing for a break in the trend function at an unknown time under both the null and alternative hypotheses. J Econom 148:1–13
Klein JL (1997) Statistical visions in time. A history of time series analysis, 1662–1938. Cambridge University Press, Cambridge
Klein LR, Kosobud RF (1961) Some econometrics of growth: great ratios in economics. Quart J Econ 75:173–198
Koopman SJ, Harvey AC, Doornik JA, Shephard N (2009) STAMP™ 8: structural time series analysis and predictor. Timberlake Consultants, London
Kozicki S (1999) Multivariate detrending under common trend restrictions: implications for business cycle research. J Econ Dyn Control 23:997–1028

Leser CEV (1961) A simple method of trend construction. J R Stat Soc Ser B 23:91–107
Maravall A, del Rio A (2007) Temporal aggregation, systematic sampling, and the Hodrick-Prescott filter. Comput Stat Data Anal 52:975–998
Matthews RCO, Feinstein CH, Odling-Smee JC (1982) British economic growth, 1856–1973. Stanford University Press, Stanford
Mills TC (1992) An economic historians' introduction to modern time series techniques in econometrics. In: Crafts NFR, Broadberry SN (eds) Britain in the international economy 1870–1939. Cambridge University Press, Cambridge, pp 28–46
Mills TC (1996) Unit roots, shocks and VARs and their place in history: an introductory guide. In: Bayoumi T, Eichengreen B, Taylor MP (eds) Modern perspectives on the gold standard. Cambridge University Press, Cambridge, pp 17–51
Mills TC (2000) Recent developments in modelling trends and cycles in economic time series and their relevance to quantitative economic history. In: Wrigley C (ed) The first world war and the international economy. Edward Elgar, Cheltenham, pp 34–51
Mills TC (2009a) Modelling trends and cycles in economic time series: historical perspective and future developments. Cliometrica 3:221–244
Mills TC (2009b) Klein and Kosobud's great ratios revisited. Quant Qual Anal Soc Sci 3:12–42
Mills TC (2011) The foundations of modern time series analysis. Palgrave Macmillan, Basingstoke
Mills TC, Crafts NFR (1996a) Modelling trends and cycles in economic history. Statistician (J Roy Stat Soc Ser D) 45:153–159
Mills TC, Crafts NFR (1996b) Trend growth in British industrial output, 1700–1913: a reappraisal. Explor Econ Hist 33(277–295):1996
Mills TC, Crafts NFR (2000) After the golden age: a long run perspective on growth rates that speeded up, slowed down and still differ. Manch Sch 68:68–91
Mills TC, Crafts NFR (2004) Sectoral output trends and cycles in Victorian Britain. Econ Model 21:217–232
Morgan MS (1990) The history of econometric ideas. Cambridge University Press, Cambridge
Morley JC, Nelson CR, Zivot E (2003) Why are Beveridge-Nelson and unobserved-component decompositions of GDP so different? Rev Econ Stat 85:235–243
Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55:703–708
Percival DB, Walden AT (1999) Wavelet methods for time series analysis. Cambridge University Press, Cambridge
Perron P (2006) Dealing with structural breaks. In: Mills TC, Patterson K (eds) Palgrave handbook of econometrics: volume 1, econometric theory, vol 1. Palgrave Macmillan, Basingstoke, pp 278–352
Phillips PCB (2005) Challenges of trending time series econometrics. Math Comput Simul 68:401–416
Pollock DSG (2009) Investigating economic trends and cycles. In: Mills TC, Patterson K (eds) Palgrave handbook of econometrics: volume 2, applied econometrics. Palgrave Macmillan, Basingstoke, pp 243–307
Proietti T (2009a) Structural time series models for business cycle analysis. In: Mills TC, Patterson K (eds) Palgrave handbook of econometrics: volume 2, applied econometrics. Macmillan Palgrave, Basingstoke, pp 385–433
Proietti T (2009b) On the model based interpretation of filters and the reliability of trend-cycle filters. Econom Rev 28:186–208
Proietti T, Harvey AC (2000) A Beveridge-Nelson smoother. Econ Letts 67:139–146
Ravn MO, Uhlig H (2002) On adjusting the Hodrick-Prescott filter for the frequency of observation. Rev Econ Stat 84:371–376
Trimbur TM (2006) Properties of higher order stochastic cycles. J Time Ser Anal 27:1–17
White H, Granger CWJ (2011) Consideration of trends in time series. J Time Ser Econom 3(Article 2):1–40
Young PC (2011) Gauss, Kalman and advances in recursive parameter estimation. J Forecast 30:104–146

Path Dependence

Douglas J. Puffert

Contents

Abstract

Path dependence is the dependence of economic outcomes on the sequence of previous outcomes, rather than simply on current exogenous causal elements. For a path-dependent process, economic explanation requires attention to the interplay of forward-looking optimizing behavior with the legacy of past conditions,

D. J. Puffert (✉)
LCC International University, Klaipeda, Lithuania
e-mail: dpuffert@lcc.lt

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_78

events, and choices. Sources of path dependence may include technical interrelatedness, irreversible investments, coordination costs, and various sorts of increasing returns. Influential or disputed cases of path dependence include the QWERTY keyboard, various information technologies, and railway track gauges. Path-dependent outcomes may be inefficient by some criteria but not by others.

Keywords

Path dependence · lock-in · Increasing returns · Technical interrelatedness · Non-ergodic processes · QWERTY

The term path dependence refers to the dependence of economic outcomes on the sequence of previous outcomes, rather than simply on current exogenous conditions. It implies that an allocation – a pattern of production and distribution – cannot be explained simply on the basis of current fundamental causal elements, typically listed as technological knowledge, resource endowments, preferences, and institutions. Rather, economic explanation requires the investigation of past conditions and events. In a shorthand expression, history matters.

The concept of path dependence has been invoked in historical explanations of numerous technical standards, including the QWERTY typing keyboard, various regional standard railway track widths (gauges), and numerous systems in information technology. Path dependence has also been applied to interpreting the evolution of institutions and other practices, economic geography, patterns of economic development, macroeconomic outcomes, and various concerns of political scientists and other social scientists. This chapter will give greatest attention to applications in microeconomics.

The Meaning and Significance of Path Dependence

Formally, a path-dependent process of economic allocation is a non-ergodic process (David 1999, 2001, 2007). In the physical sciences, a non-ergodic process is one in which a system becomes incapable of reaching all of the states that are consistent with the system's energy. In mathematical representations of such processes, the limiting distribution of potential states is itself a function of the evolution of the process. The process has branching paths, and at least some of the potential paths are nonreversible.

Likewise, in economics, early conditions, events, and choices direct a process of allocation along one potential path rather than another, limiting the range of later potential outcomes. This matters when the selected path of later outcomes exhibits, for example, an inferior technology or an otherwise lower pattern of payoffs than some alternative path would have yielded. The selected path also matters when alternative paths would bring different distributions of payoffs, with different

winners and losers. (Path dependence may also apply to non-consequential curiosities, such as driving on the left or right side of the road.)[1]

An economic non-ergodic process differs markedly from a physical process, of course, in that the selection among alternative paths involves, in large part, intentional agents who make choices on the basis of incentives and some degree of foresight. Properly speaking, if the early selection of a particular path is based "teleologically" on the foreknown payoffs of final outcomes, then the allocation process is not path dependent. (The earlier choice of a subsequent path exhibits what David (1997) called "moderate to mild" history, as opposed to "strong history" where a preceding path limits later choices.) An allocation process can be path dependent only if the early considerations that select among alternative paths are to some extent "orthogonal" to (i.e., less than fully aligned with) the system-level issues at stake in later outcomes. These early considerations are not random, in the sense of being economically inexplicable, although they might in some cases be partly the result of "historical accidents," as when the Liverpool and Manchester Railway company hired one engineer rather than another to build its track and choose its gauge (Puffert 2004, 2009). The key feature of the process is that the economic considerations that led to decisive early choices are not those that mattered later. The builder of the L&M Railway had reasons for his choice of track gauge, but most engineers soon had reasons to favor broader widths. In this case, the chosen gauge – the chosen path – was in part the result of imperfect foresight into the later implications of the choice.

The term path dependence has been applied not only to cases where "small" early considerations had "large" later system-level effects, but also to cases where the early considerations were themselves "large" and initially well aligned with system-level considerations, but where later conditions made early choices obsolete. In the case of Britain's small coal wagons, discussed below, the coal-wagon technology and related institutions established in the mid-nineteenth century served well for around half a century before becoming obsolete. Such a case may stretch the strict meaning of a non-ergodic process, as the small-coal-wagon path was perhaps the only one available at the earlier date, but it fits into the larger issue of whether the legacies of history are flexible for adaptation to evolving conditions.

The case also fits into a larger distinction between selection processes that are historically contingent and those that are teleological. Contingency is the normal mode of historical explanation. It explains event E as the result of earlier event D, which in turn is explicable (and even necessary) in terms of event C, and so on back to some event A (David 2001, referencing paleontologist Stephen J. Gould 1989). Teleology, on the other hand, is the normal mode of economic explanation. It explains an outcome on the basis of choices that are directed toward a present or future goal or end (Greek *telos*). A key issue in path dependence, then, is whether

[1]Even in that case, however, the emergence of different practices in different locations has later entailed costs.

the legacies of history are flexible for adaptation to evolving economic ends in the context of evolving exogenous or a priori conditions.

To say that an outcome is inflexible in the face of evolving conditions is equivalent to saying that the previously selected path is nonreversible or locally stable. The path of outcomes may then be called "locked-in." A standard typewriter keyboard or railway track gauge can be called locked-in, for example, if it continues to be adopted by new users and/or is retained by established users even when an alternative keyboard or gauge becomes inherently superior under evolving exogenous or a priori conditions. The causal elements that may give rise to this form of lock-in – increasing returns to scale of adoption, sunk physical or human capital giving rise to conversion costs, technical interrelatedness that requires the replacement of a whole system of obsolete components rather than their piecemeal replacement, coordination costs or transaction costs, and imperfect information – are discussed in some detail below. Also discussed below, skeptics of the importance of path dependence offer a different definition of lock-in that fits within their differing frame of reference.

The nature and significance of path dependence have indeed become matters of dispute. At the heart of the dispute are two questions: first, the extent to which later outcomes are objects of choice at earlier times when different potential paths branch apart and, second, the extent to which intentional actions can reverse an inferior path that was previously selected. Further questions include whether market failure (or institutional failure) may play a role in either the initial selection or the later nonreversibility of an inferior path, whether particular policies or sets of institutions may promote superior outcomes, and what efficiency criteria and counterfactual scenarios are appropriate for the evaluation of outcomes.

In presenting the literature on path dependence, this chapter will seek to clarify the points where interpretations differ as well as points where differing interpretations might be harmonized. This will require more attention to analytical issues than is usual in economic history. Indeed, it will also require careful attention to different usages of terms (beginning with "efficiency"), because participants in the dispute have frequently talked past one another rather than engage one another's arguments in the others' own terms. The analysis here will be grounded in historical case studies.

The Proposed Sources and Settings of Path Dependence

The concept of path dependence emerged through the confluence of two lines of economic analysis: the first focused on the concept of technical interrelatedness and the second focused on certain forms of increasing returns.

Technical Interrelatedness: The Analysis of Paul David

The analysis of technical interrelatedness traces back to Thorstein Veblen's (1915, 125–28) discussion of the small coal wagons used for contemporary rail traffic in

Britain but not elsewhere. Veblen attributed the continuation of an obsolete practice to the "facilities and all the ways and means of handling freight on this oldest and most complete of railway systems, . . . all adapted to the bobtailed car." Thus, while "the community at large" would profit from "junking" the small wagons and related facilities, individual firms would not find it profitable to convert only their own equipment while the complementary equipment remained unchanged. Marvin Frankel (1955) elaborated on the role of the interrelatedness of technical system components in preserving past practice, while Charles Kindleberger (1964) gave added attention to the role of institutions, particularly the separate ownership of different system components. This institutional pattern added coordination costs to the physical conversion costs of a potential change in practice.

Paul David (1975) made technical interrelatedness a theme in his historical studies of technical choice. He later incorporated the theme into his work on path dependence, beginning with his analysis of the emergence of the QWERTY keyboard as a technical standard (David 1985, 1986). David reflected the previous literature in specifying the first two of three conditions that, he said, may work together to make an allocation process path dependent: technical interrelatedness of system components, quasi-irreversibility of investment (or switching costs), and positive externalities or increasing returns to scale. The third condition had also arisen in the previous literature, but its implications were developed much further by an emerging line of analysis in the 1980s.

Increasing Returns: The Analysis of W. Brian Arthur

W. Brian Arthur (1989, 1990, 1994) and other economists in the 1980s explored the implications of certain sorts of increasing returns, where increased adoption of a product, practice, or technology led to rising marginal net benefits. They found, generally, that increasing returns can give rise to a multiplicity of potential equilibrium outcomes, and these equilibria might have differing patterns of payoffs. Arthur's unique contribution came through applying mathematical models, including models of stochastic processes, to treat the selection of a particular equilibrium as the result of sequential choices over time. His models exhibited path dependence.

Arthur emphasized two sources of "increasing returns to adoption": learning effects (learning by doing or by using), which increase the payoffs for a technology or product that has had greater cumulative adoption, and positive network externalities, which increase the value of a technology or product for each user as the total number of other users increases. For example, each user of a communication network (or today an Internet social platform) creates benefits not only for herself but also for other users with whom she exchanges messages.

In Arthur's main illustrative heuristic model, a technology or platform that gains a lead in market share even as a result of nonsystematic or stochastic "small events" (modeled in terms of the random order of arrival of heterogeneous adopters) becomes more valuable to later adopters, so that positive feedbacks raise its market share still further. If increasing returns are sufficiently strong, then even those new

adopters who might have a basic, stand-alone preference for a minority technology will adopt the majority one, so that it then sweeps the market and becomes "locked-in" as a de facto standard. Which technology or platform wins the market is not predictable on the basis of systematic causal elements; it depends rather on early nonsystematic events or stochastic fluctuations in choices. An ultimately inferior technology or platform can win if early temporary advantages or nonsystematic events had given it a sufficient early lead in adoptions over a technology or platform that ultimately proves superior. Such outcomes stand in contrast to the unique, predictable outcomes of standard economic models, where such "decreasing returns" features as increasing marginal costs and decreasing marginal utility (or consumer desire for variety) lead to an efficient, flexible sharing of markets among competing suppliers, so that outcomes are predictable on the basis of their superior payoffs or efficiency. The market dynamics of decreasing-returns models lead allocation processes to "forget" their history, while the market dynamics of increasing-returns processes lead them to remember their history.

In presenting his main illustrative heuristic model, Arthur (1989) noted that this model's results depend on two assumptions: first, that competing technologies are non-sponsored – not promoted by suppliers with stakes in outcomes – and, second, that the source of increasing returns is learning effects, not network effects, so that payoffs for each adopter depend only on cumulative past adoptions, not expected future adoptions. These assumptions effectively ruled out any role for forward-looking behavior in the evolution of outcomes.

Arthur did not incorporate forward-looking behavior into his modeling, nor did his economic papers usually address opportunities for such behavior when he discussed how his models might apply to specific cases. Nonetheless, he regarded the dynamic properties of his models as illustrative of a quite general pattern for allocation processes under conditions of increasing returns. Nobel economics laureate Kenneth Arrow wrote in his foreword to Arthur's (1994) collected papers that Arthur's models apply to contexts where foresight is imperfect, or "expectations are based on limited information," but Arthur himself did not interpret his models that way.

Writing later in the *Harvard Business Review*, Arthur (1996) did consider forward-looking behavior by the promoters of competing technologies. Addressing contemporary information technology markets, he asserted that successful firms base their competitive strategies on harnessing the underlying path-dependent dynamics that arise both in consumer choices and in the actions of the suppliers of complementary goods and services. His article offered extensive advice on how to understand and win such markets, and it has reportedly had great influence on the strategies of firms in Silicon Valley (Tetzeli 2016).

Other Analyses of Increasing Returns

Other economists have developed non-path-dependent models where increasing returns give rise to multiple potential equilibria, either in the initial selection of

one product or technology as a de facto standard (e.g., Katz and Shapiro 1985) or else in a potential switch from an established standard to a possibly superior alternative (Farrell and Saloner 1985). The solution concept used in these models is typically that of fulfilled rational expectations, with no role for time or sequential choices. Farrell and Saloner addressed lock-in in the context of a potential change in standards, not in Arthur's context of initial adoption. They also addressed how imperfect information might lead to excessive inertia in maintaining an established standard.

David (1993) explored models of path-dependent coordination among agents whose local interactions are represented by one- or two-dimensional network graphs. This approach enabled him to apply results from the mathematical theory of interacting particle systems or Markov random fields. In these models, agents sequentially adjust their "states" to the states of neighboring agents. Path-dependent fluctuations lead, in some configurations of models, to the adoption of a common state over the whole network, which might be interpreted as standardization on a common technology.

David also supervised (with further input from Arthur) the doctoral research of Douglas Puffert (1991, 2009), who developed a model of technical choice within a spatial network and applied it to the path-dependent selection of regional standards for railway track gauge. Puffert's model combined three innovative features: first, consideration within a single model of both initial adoptions of a technology (gauge) and subsequent conversions; second, costs for conversion; and third, the use of multiple numerical simulations by computer to explore the nature and range of potential outcomes. He used variations in parameters to simulate variations in the conditions that affected the gauge selection process in different historical-geographic contexts.

The Proposed Reasons for Skepticism About Path Dependence

The Analysis of Liebowitz and Margolis

The most prominent skeptics of the importance of path dependence, S.J. Liebowitz and Stephen E. Margolis (1994, 1995), argued that forward-looking optimizing behavior is likely to override Arthur's mechanisms for path dependence in any context where outcomes truly matter. In their analysis, adopters who choose among alternative technologies would typically have foresight into future payoffs of their choices, not just current payoffs. Furthermore, adopters and other agents would have opportunities to coordinate adopters' choices through communication, various sorts of market transactions, and the ownership and profit-seeking promotion of alternative competing products or technologies – in short, actions that internalize the mutual externalities of adopters' choices. Thus, they argued, purposeful, incentivized behavior can frequently override nonsystematic or random elements in a selection process, and path dependence can only affect aspects of the economy that no economic agent has an opportunity or incentive to change.

Liebowitz and Margolis allowed that path dependence might affect (1) aspects of the economy with no implications for efficiency or else (2) outcomes whose efficiency consequences had previously been unforeseeable and thus not subject to rational economic behavior. In the second case, agents might later express naïve regret that an outcome with higher payoffs had not been chosen, but Liebowitz and Margolis argued that it would not be meaningful to call the chosen outcome inefficient. Efficiency, they urged, should be defined in relation to what can be pursued on the basis of available information and feasible purposeful actions. In contrast to the first two cases, Liebowitz and Margolis expressed skepticism that path dependence would affect outcomes in a further case, where (3) an inferior outcome is locked-in despite the existence, at some point in time, of both the foresight and the means to direct a selection process to a superior outcome. In that case, they argued, path dependence would be the result of irrational errors, or inefficiencies that are (or had been) profitably remediable but nonetheless remain unremedied. They urged that economists cease looking for causal significance in small events or historical accidents, except perhaps in analyzing error-prone government actions, and focus instead on "the neoclassical model of relentlessly rational behavior leading to efficient, and therefore predictable, outcomes" (Liebowitz and Margolis 1995, 207).

Liebowitz and Margolis (1995) suggested, furthermore, that Arthur's illustrative model should be read as implying a claim to the unlikely "third-degree form" of path dependence, because forward-looking behavior would lead to a different outcome than the one that Arthur highlighted. The critics did not note how Arthur had explicitly stated assumptions that ruled out their proposed mechanisms for forward-looking behavior.[2] Liebowitz and Margolis also argued that proposed empirical cases of path-dependent lock-in, notably the QWERTY keyboard and the VHS videorecording system, likewise involve implicit claims for the third-degree form. Those two cases are addressed below.

The critics also urged that "lock-in" be defined solely on the basis of the third-degree criterion. That is, lock-in should refer to the unlikely case that an inferior technology predominates despite profitable opportunities to achieve a superior alternative.

Responses to the Skeptics

Arthur had ceased writing economic papers on path dependence by the time of the critique, and he did not respond extensively to his critics. In various interviews and one invited response paper (Arthur 2013), he reiterated his analysis that increasing-returns processes do not necessarily lead to the most "efficient" (i.e., highest payoff) potential outcomes in the way that decreasing-returns processes do.

[2]Those assumptions may indeed have limited applicability, but they do make it clear that Arthur was deliberately avoiding a remediable situation. He thus made no third-degree claim.

David (1999, 2001, 2007) responded more vigorously. His main contention was that Liebowitz and Margolis had not come to terms with, or ever properly addressed, the dynamic, non-ergodic nature of a path-dependent process, and thus they had misconstrued the issues. Opportunities for forward-looking choices within a path-dependent process would not be reducible to the static context of a single moment when agents might have the opportunity to select an optimal final outcome. Thus, optimizing behavior would typically achieve "path-constrained melioration" rather than unconstrained optimization. Moreover, a path-dependent process is properly defined by the dynamic nature of the process itself, not by the varying efficiency characteristics of outcomes. David therefore found the critics' redefinition of path dependence, in terms of "forms" or "degrees" based on a static criterion, to be tendentious and obscurantist rather than analytically meaningful and useful.

Likewise, David objected to the redefinition of lock-in as unlikely perverse error, when in the proper dynamic context the term can only refer to the local stability and irreversibility of an equilibrium path of allocation. This irreversibility, he claimed, is typically rooted in technical interrelatedness and in coordination costs that may be quite high, particularly under conditions of incomplete information.

Thus, David did not dispute the critics' analysis that inferior outcomes might be due to incomplete information. Indeed, he noted, the major policy implication that he had repeatedly drawn from path dependence is that government should seek to delay lock-in until the payoffs of alternative paths are better known.

More broadly, David argued that it is nonsensical to attribute significance, interest, and "efficiency" only to what remediation and optimizing market behavior can achieve. To do so, he quipped, is "tantamount to saying that market failure cannot happen, because if it did happen, markets would work to correct it" David (1999). Furthermore, David noted, Liebowitz and Margolis had not addressed how a path-dependent process might select among alternative outcomes that are each Pareto optimal, but with different sets of winners and losers.

Puffert (2004) responded to Liebowitz and Margolis by seeking to integrate their proposed means of forward-looking, profit-seeking behavior into an analytical framework that supports path dependence. In his interpretation, forward-looking agents can frequently influence but not fully control a dynamic allocation process, due to both imperfect foresight and imperfect opportunities to internalize externalities. The critics' taxonomy of three forms of path dependence leaves insufficient space (if any) between situations where there is no significant role at all for intentional action and situations where such action can fully supersede the legacy of history. Moreover, Puffert argued, the strategies of forward-looking firms frequently demonstrate an awareness and intentional usage of path-dependent dynamics by seeking to influence the early user choices that have a disproportionate effect on later outcomes. Intentional action, in Puffert's view, is fully part of path dependence.

Historical case studies illustrate the allocation processes that path dependence seeks to explain. They also offer tests of the views of both affirmers and skeptics of path dependence. Here we consider the most influential case studies, the most disputed ones, and a few studies that offer further lessons.

The Disputed Case of QWERTY

David (1985, 1986) introduced the theory of path dependence with his interpretation of how the QWERTY typewriter keyboard became established as a de facto standard. The skeptical critique of path dependence began with a response by Liebowitz and Margolis (1990).

David's Analysis

In David's telling, the QWERTY arrangement of keys was designed not for fast typing but rather for minimizing the jamming of clashing typebars on the first commercially successful typewriter, invented by Christopher Latham Sholes and brought to market by Remington in the 1870s. The 1880s saw a "rapid proliferation of competitive designs, manufacturing companies, and keyboard arrangements rivalling the Sholes-Remington QWERTY," and by the mid-1890s "it had become evident that any micro-technological rationale for QWERTY's dominance was being removed by the progress of typewriter engineering" (David 1985, 334) A few years after that, a new standard typewriter design arose, with front-stroke keys and a visible typing surface, which then fully displaced the Remington machine design for which the QWERTY keyboard had been designed.

Nonetheless, according to David (1985, 334), during "the fateful interval of the 1890s" the QWERTY keyboard became the industry-wide standard, adopted not only by Remington but also by rival manufacturers with differing machines featuring different typebar arrangements. QWERTY soon even displaced the superior Ideal keyboard, which had been designed for fast typing rather than for avoiding typebar clashes. (David noted that the Ideal keyboard was used on machines that eliminated typebars altogether by placing the type on a rotating cylindrical sleeve. Its home row DHIATENSOR could supposedly form 70% of English words.)

The reason for QWERTY's emerging dominance, according to David, was the happenstance that third-party instruction in eight-finger "touch" typing was promoted first for QWERTY. It was only then, in the late 1880s, that path dependence began to affect market competition. The best-trained typists used QWERTY, so office managers hired them and bought QWERTY machines to match. This, in turn, gave budding typists, typing schools, the writers of typing manuals, and typewriter manufacturers a further incentive to focus on QWERTY, to the exclusion of alternative systems. Positive feedbacks reinforced QWERTY's early lead until it gained virtually the whole market. (The role of typing schools and typing manuals is presented more fully in the longer 1986 version of David's paper.)

Critical to the emergence of QWERTY, wrote David, was that the "larger system of production," comprising typists, employers, manufacturers, and typing instructors, "was nobody's design"; it was characterized by "decentralized decision-making." David (1985, 334, 336) concluded, "competition in the absence of perfect

futures markets drove the industry prematurely into standardization on the wrong system."[3]

David's story ended in "the fateful interval of the 1890s" with QWERTY's "premature" establishment as a standard shortly before the rise of the new machine design that made QWERTY obsolete. In David's telling, technical interrelatedness, switching costs, increasing returns, and the absence of opportunity for centralized decision-making led to an inferior outcome.

David did not examine the later history of QWERTY, but he succinctly attributed the persistence of QWERTY to "decentralized decision-making." To establish that the QWERTY standard continues to matter, he cited claims that the Dvorak Simplified Keyboard, introduced in the 1930s, offers substantially faster typing speeds.

Liebowitz's and Margolis's Analysis

In their responding article, Liebowitz and Margolis (1990) gave attention to only part of David's account of the emergence of QWERTY as a standard. They gave more attention to popular accounts, not advocated by David, that QWERTY had been designed to slow typists down and that QWERTY won its market due to the publicity brought to QWERTY by a victory in a single typing speed contest. Liebowitz and Margolis showed that other contemporary contests were won by other typists using other keyboards, and so they claimed that they had refuted the "received history" that QWERTY had won its market due to the path-dependent consequences of an insignificant early event. They did not address David's argument that instruction in touch-typing was the starting point for path dependence in the rise of QWERTY.[4]

Instead, they noted that early typewriter manufacturers competed vigorously on features of their machines, and they inferred that QWERTY succeeded due to a market test of its relative fitness. They asserted that hypothetical suppliers of a superior alternative keyboard would have found a profit opportunity in providing training to offices where they sold their machines, and they offered evidence that the large typewriter companies of 1923 did provide such training.

Liebowitz and Margolis devoted most of their article to a topic that David had addressed only in passing, the persistence of the QWERTY standard in the face of competition from the Dvorak keyboard. They offered evidence that Dvorak offers typical typists no more than a few percentage points of increased typing speed. They

[3]In context, the absence of perfect futures markets seems to mean that there was no means by which adopters of the post-1900 generation of typewriters could have negotiated to prevent the widening adoption of QWERTY by users and manufacturers during the 1890s.

[4]In their original article, Liebowitz and Margolis (1990) cited the popular accounts of QWERTY's origin while not making clear how these accounts differed from David's interpretation. In later articles (e.g., Liebowitz and Margolis 1994), the critics presented only the popular accounts, not given by David, while citing only David's work as representing a single received history of QWERTY.

offered particularly strong evidence against claims, cited by David, that Dvorak's superiority over QWERTY was dramatic. If Dvorak really were so superior, they argued, then it would have offered substantial profit opportunities to offices that adopted it and retrained their staffs to use it.[5] Companies did not seize such opportunities, so they must not really have existed.

Liebowitz and Margolis concluded that QWERTY both won and kept its status as a standard by offering the superior realizable outcome.

Kay's Analysis

Recently Neil M. Kay (2013a) revived the literature on QWERTY with a study of the fitness of the keyboard in its original technological setting. Comparing details of the Remington machine's up-strike typebar configuration with statistical details of successive letters in typical English-language texts, he found that QWERTY was extremely well adapted for reducing typebar clashes and jamming. He inferred that QWERTY's inventor, Christopher Latham Sholes, had engaged in careful optimization in ordering the letters, going beyond "trial-and-error rearrangement" as described by David (1985, 1986). His statistical analysis also supported David's speculation that the letters spelling the brand name "TYPE WRITER" were placed deliberately on the same row; that arrangement would not likely have emerged by chance.

In Kay's analysis, the Remington typewriter's leading market share in the late nineteenth century made compatibility between "device" (machine) and "format" (keyboard arrangement) more important than compatibility between format and "user" (touch typist), and thus QWERTY was optimal for the time. Later, by contrast, the Underwood front-strike typewriter models of 1897 and 1901 made typing immediately visible to the user and thus improved compatibility between device and user, but its different typebar configuration ruined the compatibility of the QWERTY format with the device (Kay 2013b). Jamming problems then became frequent. The rise of the Underwood machine's general design, quickly adopted throughout the industry, allowed in principle a new freedom to redesign keyboard format for compatibility with the touch-typing user rather than the Remington device, but by that time, QWERTY had become a well-established standard.

Kay devised what he considered to be a test of David's argument that the typewriter industry standardized "prematurely … on the wrong system." He interpreted this as referring to the Dvorak keyboard as the "right" system. Thus, his test compared the simulated frequencies of typebar clashes, on the

[5]Liebowitz and Margolis (1995, 214) argued that David's attribution of QWERTY's persistence to "decentralized decision-making" amounted to an implicit claim for what they called third-degree path dependence. "This attribution … suggests that alternative, presumably centralized, decision mechanisms would correct the error."

It is difficult to follow the critics' reasoning. David had argued that there was no agent who could exercise centralized control over the larger system of production in which QWERTY was situated. David made no implicit claim that the relative inefficiencies of QWERTY were remediable.

Remington machine, for both the QWERTY keyboard and the Dvorak keyboard. As QWERTY led to far fewer clashes, Kay concluded that QWERTY was far fitter for the conditions of the 1870s–1890s, and thus that QWERTY would have won any hypothetical (albeit anachronistic) early competition between the systems.

Kay's account took the early leading market share of Remington's device with the QWERTY format as given, and he did not address David's argument that the contingencies of early touch-typing instruction were reinforcing this lead during the 1890s. Rather, Kay contrasted the demonstrated role of optimizing behavior in QWERTY's design with the supposed role of historical accident in the popular received history of QWERTY, where a single early typing contest had supposedly proved crucial. But this received history was not part of David's own interpretation, and David did affirm (albeit with less detail than Kay) that QWERTY was designed for fitness with the Remington device. The question that David had posed was: How a keyboard that was fit for the Remington device also became the standard for non-Remington devices? Kay affirmed David's interpretation of why a standard had to emerge eventually, and his ultimate disagreement with David may be in taking it as given that QWERTY became the standard as early as it did, before the introduction of the Underwood device.

In an invited response, Margolis (2013) regarded Kay's analysis as a vindication of the role of optimizing behavior rather than early historical accident. Arthur (2013) regarded Kay as supporting David's account of the origin of QWERTY.

Britain's Coal Cars

As noted above, the small coal cars long used in British rail traffic have been cited for a century as the paradigmatic example of technical interrelatedness. D.N. McCloskey (1973), however, called the example doubtful, because small cars were still widely used at his writing, 20 years after Britain had nationalized both the coal mines and the railways, bringing them into common ownership. Va Nee L. Van Vleck (1997) argued further that small coal cars in the early twentieth century were well adapted to the larger system of distribution at the time. They offered flexible delivery for small users of coal while economizing on the cost of road haulage that would have been necessary for small deliveries if railway coal wagons were larger.

Peter Scott (2001) argued in response that few coal users benefited from small, car-size deliveries. Rather, he wrote, the cars' small size, widely dispersed ownership (by collieries), antiquated braking and lubrication systems, and generally poor physical condition made them quite inefficient indeed. Replacing these cars and associated infrastructure with modern, larger wagons owned and controlled by the railways would have offered savings in railway operating costs of about 56%, yielding a social rate of return of 24% on the costs of new cars and the conversion of facilities. They were not replaced because regulations forced the railways to accept colliery cars at set rates or else buy out the cars at high prices and offer additional compensation. This transformed the sunk costs of obsolete cars into real

costs for the railways, and it reduced the realizable return for system-wide rationalization to around 10%. Furthermore, due to technical interrelatedness, the railways would not have saved much in operating costs until virtually all the antiquated cars were replaced, further increasing the transaction costs entailed in any transition.

In discussing how the persistence of small wagons was path dependent, Scott wrote that the technology embodied in the small cars and the institutions that supported fragmented ownership had long outlasted the earlier conditions to which they were a rational response. Ownership of cars by the collieries had been advantageous to both railways and collieries in the mid-nineteenth century, and government regulation had assigned rights in a way that protected car owners from opportunistic behavior by the railways. By the early twentieth century, these regulatory institutions were obsolete, as were the cars and physical facilities.

To situate Scott's analysis within the larger literature on path dependence, Scott did not argue that some alternative path of allocations featuring large wagons would have been just as feasible for early railways as a path featuring small wagons. He also did not argue that the original adoption of small wagons had been premature or that it was the result of imperfect futures markets. What he did argue was that technical interrelatedness and very high transaction costs resulting from the legacy of history had, together, locked in a path of outcomes that was highly inefficient relative to what alternative institutions in the early twentieth century might have made possible.

Videocassette Recording Systems

Another disputed case in path dependence was the competition between alternative videocassette recording systems from the mid-1970s to the mid-1980s. The VHS system, promoted by a coalition led by JVC (Japan Victor Corporation), became the standard, beating out Sony's Betamax. Arthur (1990) explained this as the result of positive feedbacks in the video film rental market: Video rental stores would stock more film titles for the system with the larger user base, while consumers would adopt the system for which they could rent more videos. Arthur suggested that, if the common perception that Betamax offered a superior picture quality is true, then "the market's choice" was not the best possible outcome.

Liebowitz and Margolis (1995) pointed out, however, that Sony had actually been first to market. If an early lead had mattered, they argued, then Sony should have won. They attributed the VHS victory to active product promotion and to the advantage of VHS in offering a longer playing time, and they offered substantial evidence against Arthur's suggestion that the winning system may have been inferior. In their view, purposeful, forward-looking behavior had driven the outcome, not path dependence.[6]

[6]Liebowitz and Margolis (1995) state that Arthur implicitly argued for their third-degree path dependence – remediable but unremedied error – in the case of VHS. Arthur, however, offered no analysis that might suggest that the outcome was remediable. It might be argued that Arthur's analysis was mistaken, but that does not amount to a claim.

In contrast, Cusumano, et al. (1992) argued that there was indeed a positive-feedback dynamic in the video rental market, but this market emerged late, after VHS had already gained a strong lead. The onset of positive feedbacks turned Betamax's small but stable market share into a fast-declining one, forcing it to exit the market.

More intriguingly, the authors attributed the earlier lead of VHS to path dependence in supplier choices, which was itself the result of suppliers' foresight regarding path dependence in consumer choices. Believing that consumer dynamics would lead to a single standard in the end, manufacturers and distributors increasingly supported VHS over Betamax as they saw others doing so, increasing their expectations that VHS, not Betamax, would later become the standard. Ultimately, the authors argued, VHS won as the result of differences in the promoters' early strategies. First, Sony initially pursued a go-it-alone strategy, while JVC built a coalition of suppliers in order to benefit from foreseeable positive feedbacks in market share. Second, JVC's partner Matsushita installed a large manufacturing capacity to solidify expectations among other suppliers. Third, Sony guessed that consumers would prefer a smaller cassette size, while JVC instead chose a larger cassette with longer playing time. In the event, a longer playing time proved more important to consumers in the early years, when only a VHS tape could record an entire American football game or a long movie. Distributors responded to this temporary advantage by joining the VHS coalition permanently.

Later, Liebowitz (2002) responded that the larger VHS cassette size offered a permanent rather than transitory advantage. A larger cassette with a longer tape, he argued, facilitated higher tape speeds and thus better picture quality for any given total playing time.

Information Technologies

Much of the attention of the literature on increasing returns and path dependence has been directed toward certain markets in information technology (Shapiro and Varian 1998). Network externalities (or network effects) in these markets make products with a larger market share more valuable to new users. That may give rise to bandwagon effects (Rohlfs 2001) that give competition a winner-take-all, or at least winner-take-most (Arthur 1996), character. Firms profit in such markets by developing and strategically promoting proprietary system "architectures" to become de facto standards (Morris and Ferguson 1993). Entering a market earlier matters, and new firms may have no opportunity to enter a market against an established incumbent (Arthur 1996).

Applying this sort of analysis, some observers have argued that either IBM or Apple Computer rather than Microsoft could have become the dominant firm in microcomputers, controlling the key architecture or system standard (Rohlfs 2001; Carlton 1997). However, only Microsoft leader Bill Gates had, and acted on, the foresight that control of an operating system standard would matter. Microsoft entered the personal computer industry as a contractor to IBM, supplying the

MS-DOS operating system to run on IBM's personal computers. Because IBM thought that its hardware expertise would matter more than software, it allowed Microsoft to retain intellectual property rights to MS-DOS and even to supply it to competing computer hardware firms. IBM thus lost control of what proved to be the essential architecture of its own product. When competing hardware suppliers proved able to match IBM's features without violating its patents, IBM was left mostly out of the market while Microsoft thrived.

Gates himself thought in the mid-1980s that Apple's Macintosh computer system, with its innovative graphical user interface, was particularly well suited to become the key system standard for the emerging personal computer market. It is perhaps particularly poignant that Gates even recommended what he considered a winning strategy to Apple executives in a famous memorandum sent in June 1985, a time when Microsoft gained most of its revenues from selling application software for Apple. The memo concluded, "Apple must open the Macintosh architecture to have the independent support required to gain momentum and establish a standard" (Carlton 1997, 40–43). In effect, Gates intuited and applied a concept of path dependence.

In the event, Apple preferred to keep tight control over the Macintosh's highly integrated hardware and software architecture, so it declined to follow Gates's advice. Gates then turned Microsoft's attention to developing its own open-architecture graphical user interface, Windows, built on the foundation of Microsoft's MS-DOS operating system. Microsoft indeed gained the independent support of partner firms to establish Windows as the key industry standard. Thus, Microsoft secured its dominant position in the market for personal computer operating systems.

Arthur's (1996) advice to information-technology firms included just such a strategy, and he later commended Microsoft for establishing a unifying platform that enabled the wider industry to develop and many firms to flourish (Tetzeli 2016). In response to industry concerns that Microsoft might use its leading position to stifle competition, Arthur (1996) said that Microsoft should not be penalized for winning its market, but it should be prevented from leveraging its dominant position in one market to gain competitive advantage in another market.

Arthur's caveat expressed a concern that motivated a growing opposition to Microsoft among some firms in Silicon Valley, as well as by the Antitrust (competition policy) Division of the US Department of Justice. Beginning in 1994 Silicon Valley-based attorney Gary Reback and colleagues wrote a series of White Papers, with consulting assistance from Brian Arthur and economist Garth Saloner, as submissions to the Antitrust Division. In one of these, the authors sought to convince the government to take action against Microsoft for bundling its Windows operating system with Microsoft's new Internet browser, Internet Explorer, arguing that this bundling prevented effective competition by the similarly new Netscape browser. They argued that Microsoft was using its dominance in computer operating systems to preemptively attain dominance in the then-emerging Internet market, stifling competition to the detriment of consumers. The White Paper received favorable attention from some antitrust scholars and attorneys as well as parts of the press, and it helped lead the Antitrust Division to pursue and obtain a consent agreement with Microsoft to restrict its competitive practices.

Liebowitz and Margolis wrote numerous op-ed articles and policy papers against the argument of the White Paper, culminating in a book (Liebowitz and Margolis 1999). Building upon their prior arguments about path dependence, they asserted that predatory bundling is an unsupportable concept. They further argued that Microsoft had gained its leading position in several software markets simply by offering superior products.

The broader basis for Liebowitz's and Margolis's (1994) argument was, in part, their perception that the network externalities that benefit users of information products are "pecuniary" externalities. Pecuniary externalities are the gains or losses that one economic agent causes for others when the agent's actions lead to changes in prices. Although other agents gain or lose, the change in price reflects the adjustment of a market to a new equilibrium outcome that is efficient. Liebowitz and Margolis argued that when externalities are mediated by market transactions with suppliers, they are pecuniary in character and thus have no implications for economic efficiency. The actions of suppliers to win their markets are, therefore, simply part of the benign process of vigorous competition in which the best products have the best opportunities to win. Restrictions on Microsoft's competitive practices would therefore harm consumers. More broadly, in the view of Liebowitz and Margolis, the literatures on path dependence and network externalities have little to contribute to the analysis of information technology markets.

A partial counterargument to Liebowitz and Margolis had essentially been offered earlier by Nobel laureate Paul Krugman (1991, 485) in a discussion of increasing returns in economic geography. In perfect competition, Krugman wrote, "pecuniary externalities have no welfare significance and could not lead to … interesting dynamics … [but] in the presence of imperfect competition and increasing returns, pecuniary externalities matter; for example, if one firm's actions affect the demand for the product of another firm whose price exceeds marginal cost, this is as much a 'real' eternality as if one firm's research and development spills over into the general knowledge pool." Expanding on Krugman's point, the familiar analysis that allows pecuniary externalities to be ignored holds only for perfectly competitive, price-taking markets with convergent dynamics and thus unique equilibria. It does not hold for the conditions that some industry analysts attribute to many segments of information technology markets, and it does not hold for the conditions that Arthur and others proposed in theories of increasing returns leading to path dependence. Under those conditions, Krugman's "interesting dynamics" might affect outcomes. Thus, in the view of Arthur and others, allowing Microsoft to bundle products and enter a market with an immediate lead in market share might prevent competitors from introducing and promoting superior products.

Economic Geography

Since 1990, there has been a flurry of applications of path dependence to explanations in economic geography. Most notably, Nobel laureate Paul Krugman (1991), Arthur (1994), and others have explained instances of industrial location as the result

of historical small events that grew in impact as a result of positive feedbacks. Agglomerations of activity such as Silicon Valley, for electronics and information technology, and Dalton, Georgia, for US carpet production, are interpreted as having begun with the settlement of a few initial firms for nonsystematic reasons rather than any inherent advantage in the location. The initial firms were joined by an increasingly thick web of specialized suppliers, industrial customers, and competitors as new firms found cost advantages in locating near other firms in the same industry. Particularly in Silicon Valley, the advantages have included sharing of technological knowledge and thick specialized labor markets that have facilitated the staffing of start-up firms while giving workers from failed start-ups quick opportunities for reemployment.

This literature has upended the traditional focus of economic geography on systematic reasons for industrial location. In modeling, agglomeration externalities (or effects) have functioned much like network externalities (or effects) in making allocation processes path dependent. Also similarly, Liebowitz and Margolis (1994, 1995) argued that the externalities are largely pecuniary in nature and/or internalized through such market mechanisms as ground rents, so that there are no substantial welfare implications for the industrial location process. Krugman (1991), by contrast, offered a model in which externalities lead endogenously to the separation of a core industrial region from a periphery, with differing productivities and incomes in the different regions. (His model employed pecuniary externalities, but he argued that it made no essential difference whether the externalities were pecuniary or not.)

In historical research, Bleakley and Lin (2012) investigated portage sites around rapids on rivers in parts of the United States. These sites became early centers of commerce and industrialization, due both to breaks in the progress of travel and to the availability of water power. More interestingly, Bleakley and Lin documented how these sites remained economically important and even increased in importance after the original advantages of the sites were made obsolete. They interpreted this finding using a model in which local increasing returns to scale gave rise to path dependence.

Nikolaus Wolf and various coauthors have written numerous papers on industrial location in Europe. N.F.R. Crafts and Wolf (2014) explained the location of Britain's cotton textiles industry in 1838 as a result of a combination of "first and second nature" features – both the original advantages of previous decades, including water power, and acquired advantages that resulted in part from sunk costs and agglomeration effects. Redding et al. (2011) addressed how the division of Germany following the Second World War caused the shift of Germany's national air transport hub from Berlin to Frankfurt. The reunification of Germany and the return of Germany's capital to Berlin have not caused any reversion of the hub to Berlin. The authors developed evidence that the relocation of Germany's air hub was not uniquely determined by economic fundamentals, but was rather a shift between multiple steady states.

Institutional Change

There is a growing literature as well on path dependence in the development of institutions. Douglass North (1990) was early in applying the concept of path dependence to institutional change, pointing to historical examples where

institutions persisted while changing conditions made them increasingly maladapted. David (1993, 1994) posited an analytical similarity of path dependence in technology, institutions, organizations, and other matters.

Eichengreen (1996) and Meissner (2005) argued that the international diffusion of monetary systems, such as the classical gold standard of the late nineteenth century, was path dependent. They interpreted the reduction in transaction costs from adopting a common monetary system in terms of positive network externalities.

Gary Libecap (2011) addressed long-term institutional path dependence in the property rights and distributional mechanisms for water supply in arid districts of the American West. He found that arrangements developed to address conditions of the late nineteenth and early twentieth centuries have reduced later opportunities to reallocate water to higher value uses or to respond to hydrological uncertainty.

Nuclear Power Reactors and Pest Control

In another early publication on path dependence, Robin Cowan (1990) argued that transitory circumstances and path-dependent learning effects led to the establishment of the prevalent "light-water" design for civilian nuclear power reactors. This design, adapted from nuclear submarines, was rushed into use during the Cold War due to the political value of demonstrating peaceful uses for nuclear technology. Thereafter, learning effects arising from engineering experience continued to make the light-water design the rational choice for new reactors. Cowan argued, however, that an equivalent degree of development would likely have made an alternative design superior.

Cowan and Gunby (1996) addressed path dependence in farmers' choices between the system of chemical pest control and the alternative system of integrated pest management (IPM). IPM uses predatory insects to devour harmful ones, and any drift of chemical pesticides from neighboring fields makes the use of IPM difficult or impossible. IPM must therefore be used on the whole set of farms that are in proximity to one another. Where this set is large, the transaction costs of persuading all farmers to forgo chemical methods often prevent adoption. In addition to these localized positive feedbacks, local learning effects also make the choice between systems path dependent. Local lock-in to each practice is sometimes ended by such developments as invasions by new pests and the emergence of resistance to pesticides.

Railway Track Gauge

Puffert's (2000, 2002, 2009) research on the emergence of regional standards for railway track gauge appears to be the most extensive case study yet undertaken of path dependence in technical choice. The large number of local realizations of the gauge selection process (in contrast to single realizations for QWERTY and other cases) helped to clarify the ways that various features of the process mattered.

The single most crucial event in the history of railway track gauge was the hiring of engineer George Stephenson to build the Liverpool and Manchester (L&M) Railway, opened in 1830. Stephenson used the 4 ft, eight and a half inch gauge of the primitive coal tramway where he had conducted early experiments with steam locomotion, while a rival team of engineers was proposing to build the line using the unprecedented 5 ft, 6 in. gauge that they thought proper for a new generation of railways. The L&M immediately became the model of best practice for other early railways in Britain, the United States, and Continental Europe, and its gauge was widely adopted. Within a decade of the L&M's opening, however, many engineers came to regard broader gauges as superior. Nonetheless, the Stephenson gauge has retained a lead in aggregate route length ever since.

Puffert identified both nonsystematic events and systematic developments that led to the introduction of particular gauges to particular regions. He addressed how positive feedbacks led to the emergence of regional standard gauges, as new railway lines generally adopted the gauges of established neighbors, while diversity in gauge emerged on a larger scale, because different regions within a country or continent initially adopted different gauges. As a result, two regional standard gauges emerged in Britain, six in North America, six in Continental Europe, three in Australia, and multiple gauges in other intercommunicating regions as well. Puffert considered how optimizing behavior subsequently led to conversions of gauge and improved network integration in some cases but not in others.

Puffert found that the cost of either coping with or resolving diversity was the main path-dependent inefficiency in track gauge, well outweighing the inefficiency of the prevalent Stephenson gauge relative to hypothetical broader standard gauges. Still, diversity was resolved most easily where it proved most costly, and the mechanism for resolving diversity was frequently the sort of coordinating behavior discussed by Liebowitz and Margolis (1995). Much of Britain's and North America's diversity, for example, was resolved by the growth of interregional rail systems under common ownership, which internalized the benefits of standardization. Still, this improvement took the form of what David (2001) called path-constrained melioration, not full adaptation to practices that engineers and railway officials regarded as optimal. Britain made the Stephenson gauge its standard at a time when the expert consensus favored gauges ranging from 5 ft (1524 mm) to 5 ft 6 in. (1676 mm). North America standardized on the Stephenson gauge when the expert consensus there favored 5 ft, the regional standard gauge that was then being turned away from in the southeastern United States. Government involvement assisted beneficial standardization in some cases while hindering it in others, sometimes for protectionist reasons.

Puffert identified roles in the gauge selection process for numerous idiosyncratic individuals making influential choices, some on the basis of acute foresight or purposeful learning, others with less thoughtfulness, but all constrained by the past while pursuing their future visions. Technology, geography, and market forces played important systematic roles, while positive feedbacks lent growing impact to nonsystematic features of the process. In contrast to what Kay (2013) wrote

of QWERTY, in a rerun of the tape of history, the Stephenson gauge would not always win.

One episode offers a particularly poignant example of how a path-dependent process might preserve and magnify the effects of an idiosyncratic error that would be easily weeded out by an ergodic allocation process. In 1850, the three main Australian colonies, New South Wales, Victoria, and South Australia, coordinated their choices of gauge in adopting 5 ft 3 in., the gauge recently chosen by the British government for Ireland. However, the engineer hired in 1852 to build the first line in New South Wales insisted on using the Stephenson gauge, and the colony's legislative council agreed. Railway companies in the other colonies protested furiously while proceeding with their plans to adopt the broader gauge, with the support of local governments. One notably perspicacious engineer in Victoria argued vigorously for adapting to the regrettable vicissitudes of New South Wales, but other railway officials and politicians stubbornly exaggerated the benefits of their preferred gauge. Both the Governor General of Australia and Britain's Colonial Office declined to intervene (despite expert advice favoring both the broader gauge and the necessity of standardization), reflecting a policy to promote autonomy in the colonies.[7]

In due course, as many indeed foresaw, Australia was saddled with a needless and costly diversity of gauge that has persisted to the present day. Efforts to standardize Australia's gauge began to bear fruit after a century of dispute, and so the national network has been rationalized to a substantial extent. However, the costs of first coping with and then partly resolving Australia's lack of network integration have been great. These costs might be regarded as a result of institutional failure – although not simply either market failure or government failure. More fundamentally, these costs are the result of an allocation process that neither systematically overrides errors nor assures an optimal technical choice or degree of network integration, even by the standards of the knowledge of the time.

Conclusion

The concept of path dependence offers a fruitful framework for investigating the impact of past choices and events on current features of the economy. It can be a tricky concept, as it requires close attention to ways that outcomes depend on both constraints arising from the past and goals oriented toward the future.

Path dependence involves both the branching of potential paths of allocation and the stability or lock-in of paths. The branching may be due in part to nonsystematic elements in the process, as well as to increasing returns giving rise to positive feedbacks. The stability may be due to coordination costs as well as the physical costs of a change in practice.

[7]Many features of this story are, of course, well known in Australia, but the recent research of John Mills (2007) clarified the interrelated choices of key actors during the early 1850s.

Some of the theorists of path dependence were arguably insufficiently attentive, early on, to ways that forward-looking behavior may override or greatly modify a path-dependent process. The skeptics have arguably been reductionistic in treating a fundamentally dynamic process as an essentially static one, while failing to see how the affirmers have their own sophisticated understandings of markets.

The test of an analytical concept, of course, is its fruitfulness in explanation. Does path dependence, along with such related concepts as increasing returns and positive feedbacks, offer insights that illuminate previously hidden facets of economic outcomes and processes of change? Are the contrasting explanatory modes of historical contingency and intentional teleology ultimately compatible and indeed synergistic? Is economic history merely the investigation of the past role of purposeful behavior? Or does economic history offer a new vision of how the economy and society work? One lesson may be that outcomes are not always the predictable result of an impersonal invisible hand – but rather the surprising result of persons whose choices take the path of history in one direction rather than another, for good or for ill or simply for different.

Cross-References

- ▶ Douglass North and Cliometrics
- ▶ Financial Systems
- ▶ Institutions
- ▶ Railroads
- ▶ Travel and Tourism

References

Arthur WB (1989) Competing technologies, increasing returns, and lock-in by historical events. Econ J 99:116–131

Arthur WB (1990) Positive feedbacks in the economy. Sci Am 262. (February:92–99

Arthur WB (1994) Increasing returns and path dependence in the economy. University of Michigan Press, Ann Arbor

Arthur WB (1996) Increasing returns and the new world of business. Harv Bus Rev 74(4):100–109

Arthur WB (2013) Comment on Neil Kay's paper: rerun the tape of history and QWERTY always wins. Res Policy 42:1186–1187

Bleakley H, Lin J (2012) Portage: path dependence and increasing returns in U.S. history. Q J Econ 127:587–644

Carlton J (1997) Apple: the inside story of intrigue, egomania, and business blunders. Times Business, New York

Cowan R (1990) Nuclear power reactors: a study in technological lock-in. J Econ Hist 50:541–567

Cowan R, Gunby P (1996) Sprayed to death: path dependence, lock-in and pest control strategies. Econ J 106:521–542

Crafts NFR, Wolf N (2014) The location of the British cotton textiles industry in 1838: a quantitative analysis. J Econ Hist 74:1103–1139

Cusumano MA, Mylonadis Y, Rosenbloom RS (1992) Strategic maneuvering and mass-market dynamics: the triumph of VHS over Beta. Bus Hist Rev 66:51–94

David PA (1975) Technical choice, innovation and economic growth: essays on American and British experience in the nineteenth century. Cambridge University Press, Cambridge, UK
David PA (1985) Clio and the economics of QWERTY. Am Econ Rev Pap Proc 75:332–337
David PA (1986) Understanding the economics of QWERTY: the necessity of history. In: Parker WN (ed) Economic history and the modern economist. Oxford University Press, Oxford
David PA (1993) Path dependence and predictability in dynamic systems with local network externalities: a paradigm for historical economics. In: Foray D, Freeman C (eds) Technology and the wealth of nations: the dynamics of constructed advantage. Pinter, London
David PA (1994) Why are institutions the 'carriers of history'? Path dependence and the evolution of conventions, organizations and institutions. Econ Dyn Struct Chang 5:205–220
David PA (1997) Path dependence and the quest for historical economics: one more chorus of the ballad of QWERTY. University of Oxford discussion papers in economic and social history no. 20
David PA (1999) At last, a remedy for chronic QWERTY-skepticism! Discussion paper for the European Summer School in Industrial Dynamics, held at l'Institute d'Etudes Scientifiques de Cargese (Corse), France, September. https://econwpa.ub.uni-muenchen.de/econ-wp/eh/papers/0502/0502004.pdf. Accessed 29 Apr 2019
David PA (2001) Path dependence, its critics and the quest for 'historical economics'. In: Garrouste P, Ioannides S (eds) Evolution and path dependence in economic ideas, past and present. Edward Elgar, Cheltenham/Northampton
David PA (2007) Path dependence: a foundational concept for historical social science. Cliometrica 1:91–114
Eichengreen B (1996) Globalizing capital: a history of the international monetary system. Princeton University Press, Princeton
Farrell J, Saloner G (1985) Standardization, compatibility, and innovation. Rand J 16:70–83
Frankel M (1955) Obsolescence and technological change in a maturing economy. Am Econ Rev 45:296–319
Gould SJ (1989) Wonderful life: the burgess shale and the nature of history. W.W. Norton, New York
Katz ML, Shapiro C (1985) Network externalities, competition, and compatibility. Am Econ Rev 75:424–440
Kay NM (2013a) Rerun the tape of history and QWERTY always wins. Res Policy 42:1175–1185
Kay NM (2013b) Lock-in, path dependence, and the internationalization of QWERTY. Scottish institute for research in economics discussion paper 2013–41
Kindleberger CP (1964) Economic growth in France and Britain, 1851–1950. Harvard University Press, Cambridge, MA
Krugman P (1991) Increasing returns and economic geography. J Polit Econ 99:483–499
Libecap GD (2011) Institutional path dependence in adaptation to climate: Coman's "some unsettled problems of irrigation". Am Econ Rev 101:1–19
Liebowitz SJ (2002) Rethinking the network economy. AMACOM, New York
Liebowitz SJ, Margolis SE (1990) The fable of the keys. J Law Econ 33:1–25
Liebowitz SJ, Margolis SE (1994) Network externality: an uncommon tragedy. J Econ Perspect 8:133–150
Liebowitz SJ, Margolis SE (1995) Path dependence, lock-in, and history. J Law Econ Org 11:204–226
Liebowitz SJ, Margolis SE (1999) Winners, losers, and Microsoft: competition and antitrust in high technology. The Independent Institute, Oakland
Margolis SE (2013) A tip of the hat to Kay and QWERTY. Res Policy 42:1188–1190
McCloskey DN (1973) Economic maturity and entrepreneurial decline: British iron and steel. Harvard University Press, Cambridge, MA, pp 1870–1913
Meissner CM (2005) A new world order: explaining the international diffusion of the gold standard, 1870-1913. J Int Econ 66:385–406
Mills JA (2007) The myth of the standard gauge: rail gauge choice in Australia 1850–1901. PhD dissertation, Griffith University

Morris CR, Ferguson CH (1993) How architecture wins technology wars. Harv Bus Rev 71:86–96. (March–April)
North DC (1990) Institutions, institutional change, and economic performance. Cambridge University Press, Cambridge, UK
Puffert DJ (1991) The economics of spatial network externalities and the dynamics of railway gauge standardization. PhD dissertation, Stanford University
Puffert DJ (2000) The standardization of track gauge on North American railways, 1830–1890. J Econ Hist 60:933–960
Puffert DJ (2002) Path dependence in spatial networks: the standardization of railway track gauge. Explor Econ Hist 39:282–314
Puffert DJ (2004) Path dependence, network form, and technological change. In: Guinnane T, Sundstrom W, Whatley W (eds) History matters: essays on economic growth, technology, and demographic change. Stanford University Press, Stanford
Puffert DJ (2009) Tracks across continents, paths through history: the economic dynamics of standardization in railway gauge. University of Chicago Press, Chicago
Redding S, Sturm D, Wolf N (2011) History and industrial location: evidence from German airports. Rev Econ Stat 93:814–831
Rohlfs JH (2001) Bandwagon effects in high-technology industries. MIT Press, Cambridge, MA
Scott P (2001) Path dependence and Britain's 'coal wagon problem'. Explor Econ Hist 38:366–385
Shapiro C, Varian HR (1998) Information rules. Harvard Business School Press, Cambridge, MA
Tetzeli R (2016) A short history of the most important economic theory in tech. Fast Company. https://www.fastcompany.com/3064681/most-important-economic-theory-in-technology-brian-arthur. Accessed 14 Apr 2019
Van Vleck VNL (1997) Delivering coal by road and rail in Britain: the efficiency of the 'silly little bobtailed' coal wagons. J Econ Hist 57:139–160
Veblen T (1915) Imperial Germany and the industrial revolution. Macmillan, London

Analytic Narratives

What They Are and How They Contribute to Historical Explanation

Philippe Mongin

Contents

Abstract

The expression "analytic narratives" is used to refer to a range of quite recent studies that lie on the boundaries between history, political science, and economics. These studies purport to explain specific historical events by combining the usual narrative approach of historians with the analytic tools that economists and political scientists draw from formal rational choice theories. Game theory, especially of the extensive form version, is currently prominent among these tools, but there is nothing

This chapter expands on work begun when the author visited Wissenschaftskolleg zu Berlin in 2015–16. For useful comments and encouragement, the author thanks Steven Brams, Bertrand Crettez, Lorraine Daston, Claude Diebolt, Françoise Forges, Luca Giuliani, Michael Gordin, Michael Haupert, Benjamin Miller, Roger Ransom, Daniel Schönpflug, Frank Zagare, as well as the participants at the workshops "The Limits and Possibilities of Narrative Explanations" (Wissenschaftskolleg, 17–18 March 2016) and "Computational Models of Narrative 2016" (Krakow, 11–12 July 2016). Many thanks also to Ben Young and Michael Haupert for assisting in the preparation of the final draft.

P. Mongin (✉)
GREGHEC, Economics and Decision Sciences, CNRS & HEC Paris, Jouy-en-Josas, France
e-mail: mongin@greg-hec.com

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_52

inevitable about such a technical choice. The chapter explains what analytic narratives are by reviewing the studies of the major book *Analytic Narratives* (Bates et al., 1998), which are concerned with the workings of political institutions broadly speaking, as well as several cases drawn from military and security studies, which form an independent source of the analytic narratives literature. At the same time as it gradually develops a definition of analytic narratives, the chapter investigates how they fulfil one of their main purposes, which is to provide explanations of a better standing than those of traditional history. An important principle that will emerge in the course of the discussion is that narration is called upon not only to provide facts and problems but also to contribute to the explanation itself. The chapter distinguishes between several expository schemes of analytic narratives according to the way they implement this principle. From all the arguments developed here, it seems clear that the current applications of analytic narratives do not exhaust their potential, and in particular that they deserve the attention of economic historians, if only because they are concerned with microeconomic interactions that are not currently their focus of attention.

Keywords

Analytic narratives · Game theory · Rational choice theory · Historical explanation · Narratives versus models · Case study method · Equilibrium analysis of institutions · Deterrence theory · Political economy · Security studies

Introduction

The expression "analytic narratives" refers to studies that are located at the academic boundaries between history, political science, and economics. These studies purport to explain specific historical states of affairs by combining the usual narrative approach of historians with the analytic approach that is familiar to economists and political scientists. Being specific, and indeed often highly specific, the historical situations, events, or actions they cover rarely overlap from one study to another. If there is any unity to analytic narratives, it does not lie in the objects but in the method of explanation, and from this angle, they have two broad principles in common. The first is that analytic narratives jointly exploit the resources of narration and analysis, the presumption being that this can result in better solutions to explanatory problems than if either technique were used in isolation. The second principle is that the analytic component is drawn from the theories of rational decision-making, prominent among which is game theory; the presumption here is that the tools they offer do fit the purpose of combining narration and analysis. More needs to be said to characterize analytic narratives, but these two principles are part and parcel of their definition.

Both principles come out most clearly in *Analytic Narratives* (Bates et al. 1998), an important collective book that popularized the expression and provided the approach with a manifesto as well as illustrative case studies. These studies belong

to the historical branch of political science, and to get the full range of the genre, one must turn to the historical parts of those other fields – to wit, military studies, security studies, and international relations (IR) studies – in which analytic narratives have also undergone autonomous development. Proximate forms of analytic narratives had circulated there before the eponymous book came out. Besides giving retrospective structure to these significant, albeit unconscious, past attempts, *Analytic Narratives* pursues a specific program on political institutions, which it proposes to reconstruct as equilibria of individual interactions, these generally being modeled by game theory. The chapter is concerned with the connections between analytic narratives and history, and although it will mention this connection with theoretical political science, this will not be developed here.

We will explain what analytic narratives (AN) are by surveying, first, the five cases in the eponymous book and then five further cases drawn from military and security studies, to which we append a case that is again borrowed from political history but uses the same techniques as those in the latter group. In general, we follow a bottom-up approach, first summarizing the cases, and then attempting to capture their methodological features. As we journey along this inductive road, we will identify a third guiding principle of AN, which is less transparent than the first two, to the effect that the narrative component does not simply provide the data against which explanatory hypotheses are to be tested, *but also contributes to the explanation as such*. The chapter takes the third principle to be definitional, just like the first two, which amounts to defining AN more precisely than is usually done.

The emphasis on the third principle, and the upgrading of the narrative component more generally, is common to this chapter and other accounts by the same author, where this component receives even more emphasis. Along the same line of analysis, we will draw an internal distinction between different forms of AN. The crucial observation here is that some AN give the final explanatory word to a narrative, while others state their explanatory conclusions in theoretical language. Thus, although we regard it as being definitional, we take the third principle to be implementable in quite different ways from one AN to another.

Besides providing a definition, the chapter assesses the extent to which AN contribute to historical explanation. For this, we use the scheme of deductive explanation, which Hempel (1965) and other philosophers of science proposed to clarify the structure of scientific explanation. This scheme is popular among some AN contributors: however, we will argue that the AN themselves conform only very roughly and imperfectly to it. Here, however, the discovery of deductive failures in the explanatory arguments functions as a positive feature, as it prepares us for the claim that the narrative component of AN complements deduction in structuring their explanations. This is how the chapter connects its two topics, i.e., the definition of AN and the account of their explanatory capacity.

The chapter develops as follows. Section "The Five Studies of *Analytic Narratives*" summarizes the work by Bates, Greif, Levi, Rosenthal, and Weingast collected in *Analytic Narratives*. Section "Some Defining Characteristics of Analytic Narratives" exploits this material to make some progress with the definition of AN. In particular, it argues that AN should involve proper formalism, but that this formalism does not have

to be limited to the game theory employed in the book. This section merely makes explicit what the authors themselves suggest. Section "Analytic Narratives from Military and Security Studies" extends the sample with two military studies by Haywood and Mongin, a group of studies on the Cuban crisis, which are included here only to facilitate comparison, several security studies by Zagare (in particular taken from his 2011 reference book *The Games of July*), and a study of post-communist political transitions by Nalepa. Section "Analytic Narratives and Deductive Explanation" discusses how AN contribute to historical explanation, by reference to the deductive scheme of scientific explanation. Section "The Role of Narration in Analytic Narratives" deals with the narrative element of AN, arguing that it can make up for some of the failures of the deductive scheme that the previous section pointed out. This concludes our assessment of the contribution AN make to historical explanation, at the same time as establishing our proposed definition for this genre. The section also sets out a taxonomy of AN based on the way narratives enter the exposition, the three categories being alternation (of the narrative with the model), local complementation (of the narrative by the model), and analyzed narratives (in which the model and the narrative are merely juxtaposed). Section "Conclusion" briefly takes stock suggesting that AN may be a tool for economic history.

The Five Studies of *Analytic Narratives*

Here we simply review the five case studies presented in *Analytic Narratives*, following the chronological order adopted in that book. The next section will use this major sample to introduce a general discussion of AN.

Case 1 Middle-Ages Genoa (Greif) In the Middle Ages, the city-state of Genoa was first governed by elected consuls (1096–1194) and then by an appointed magistrate, the *podestà*, who was chosen from outside the city (1194–1334). Under the consulate, civil peace prevailed from 1096 to 1164 (period I), and then there was civil war lasting from 1164 to 1194 (period II). Under the *podesteria*, civil peace prevailed throughout (period III). Genoa's main economic activity was long-distance trade in the Mediterranean, and this activity was prosperous concomitantly with civil peace, i.e., for periods I and III, but with a noticeable peak at the end of the former. The main actors of economic and political life were the clans, which appear to have kept their identity and relative influence fixed for much of the period under study. In view of this fact, the time sequence emerges as problematic. Why did the clans first cooperate and then fight under the politically unchanged conditions of the consulate? Why did they cooperate most efficiently at the end of the first period of civil peace? How did the institutional move to the *podesteria* contribute to reestablishing the civil peace that prevailed henceforth and why did it occur when it did? These are Greif's main explanatory questions. He notes that the historians' work fails to answer them satisfactorily, or even to raise them in full clarity.

Greif responds by constructing two classes of extensive form games of perfect information. He then investigates their subgame perfect equilibria, as is classically

done for such games.[1] The first class, which relates to the consulate regime, has two games, both of which involve the clans as players; the difference between them hinges on whether the number of maritime possessions of Genoa is taken to be exogenous or endogenous. We will report only on the simpler of the two, which is the game with an exogenous number of possessions.[2] This game explains the changes from period I to II by using the external threat posed by the German Emperor as a variable parameter. Depending on whether the threat is absent or present, Greif retains a different subgame perfect equilibrium – here relabelled as *mutual deterrence equilibrium* (MDE). The presence of this threat pushes the clans towards *mutually advantageous* MDE by the following mechanism. In general, clans compete to gain control over the consulate, which would guarantee them a higher share of trade benefits, and this competition stabilizes peacefully only because they spend on deterrence resources they could more profitably spend on joint trade; this is what MDE formally captures. Now, the controlling clan also incurs the burden of external wars when they happen, so that the external threat changes the clans' *ex ante* net benefits of conquering the consulate; this is why MDE with fewer resources spent on deterrence, and more on joint trade, arise when there is such a threat. The chapter presents his first class of games only informally; a full treatment appears in Greif (2006, Chap. 8).

The second class has a single game, which is intended for period III and has the *podestà* as a third player. Among others, it captures a clan's two strategic possibilities of accepting the *podestà*'s authority or attempting to take control of Genoa, at the risk of starting a war against the *podestà* and the other clans. For relevant parametric conditions, this game has a subgame perfect equilibrium that explains the stabilizing effect of the *podesteria*. At this equilibrium, clan 1 (which makes the first move) abstains from challenging clan 2; clan 2 (which reacts to clan 1's move) fights if challenged; and the *podestà* (who reacts to the two clans' moves) joins forces with clan 2 against clan 1 in case of a fight, but colludes with clan 1 otherwise. That the *podestà* can possibly collude with clan 1 motivates clan 2 to fight, and that the *podestà* can possibly support a fighting clan 2 motivates clan 1 not to challenge in the first place. As entailed by subgame perfect equilibrium, these two threats are credible. The parameters on which the existence of this equilibrium depends are the players' probabilities of victory and defeat and the accompanying payoffs. Most important are the parameter values for the *podestà* since they ought to match his reward scheme and military means, as described by historians. Greif's chapter presents the *podesteria* model in complete detail (see also Greif 2006, Chap. 8).[3]

[1]For the game theory discussed in this chapter, see the texts by Morrow (1994) and Harrington (2009), and at a more advanced level, by Myerson (1991) and Osborne and Rubinstein (1994).

[2]Only the second game makes it possible to investigate the clans' trade-off between fighting a civil war to gain control of Genoa and peacefully collaborating to get more maritime possessions.

[3]Critics of Greif's approach to the *podesteria* have complained that this institution originated in a decision made by the German Emperor, not by the Genoese. See, however, fn 5.

Case 2 Ancien Régime Finances (Rosenthal) A classic historical problem concerns understanding why the pace of institutional change differed between France and England in the seventeenth and eighteenth centuries, with one country keeping the absolutist monarchy until its final disruption, while the other moved gradually towards representative government. Rosenthal reconsiders the problem in the light of the two countries' difference in fiscal structure. Given that the product of taxes was mostly spent on wars, this leads him to raise another question, i.e., how a country's style of warfare relates to its political regime.

The examination is carried out in terms of an informally stated model that an appendix makes formal. There are two actors, the King and the elite (an abstraction representing the parliaments in France and England, and the provincial estates in France, where they existed), who enjoy separate fiscal resources and try to make the best of them in fighting profitable wars. By assumption, the King alone has the power to launch a war, and if he exerts it the elite decides whether or not to participate financially. Since most wars need joint funding, there is a free rider problem, which, the model shows, is more acute when the fiscal resources are shared between the King and the elite than when they are in one player's hands. This translates into the prediction that wars are more frequent, the higher the King's share of fiscal resources. For Rosenthal, France's absolutism was a case of sharing, whereas England's representative government was one of near control by the elite. Hence, he has a rough prediction to test on the two countries and can address the issue of how their warfare relates to their regimes. His model also implies the correct prediction that the overall level of taxation was higher in England than in France. However, it is unclear how it addresses the initial question of the different pace of political change in the two countries.

Case 3 19th Century Conscription (Levi) In the nineteenth century, several Western states changed their regulations of military service, moving from conscription with provisions for buying out of the duty to more or less universal conscription. Historians have usually emphasized democratization and military efficiency as being the two likely reasons for this. However, the latter is technically doubtful (a professional army would have dominated all other arrangements), and the former is objectionable in view of the timing of reforms (they indifferently took place either before or after universal suffrage prevailed). Starting from these objections, Levi compares the changes in France, the United States, and Prussia, paying attention not only to the chronological pattern but also to the variable pattern of buying out (there were three distinctive forms, i.e., substitution, replacement, commutation). She does not mean to displace the previous explanations entirely, but rather to subsume them under her own.

To do so, Levi develops an informal analysis in the spirit of formal political economy, whereby three main actors contribute to shape national decisions on the conscription regime. They are the army, which wants only military efficiency; the government, which balances it against social and economic considerations, such as

employing the population efficiently; and the legislature, which aligns itself with the coalition among three social groups (traditional elites, middle class, and workers) that make up the constituent body. With this construction at hand, the pattern of reform in each country can be explained by hypothesizing changes in the actors' motivations. Levi proposes two such changes, i.e., the increased demand from army and government for troops and the legislature's evolving preferences, both of which push in the direction of universal conscription. She relates the latter change to a reshuffle within the politically influential coalition (the pivotal middle class turning away from the traditional elites and becoming allied with the workers), as well as to an increased taste for equality among the social groups. The two main hypotheses from the historical literature appear again, though included within a systematic explanation scheme. The study draws on Levi's (1997) thorough antecedent work on the same topic and thus includes rich historical evidence.

Case 4 Antebellum Federation (Weingast) Historians of the United States have long been puzzled by the relative stability of the federation through the decades that preceded the Civil War. Classically, they argue that slavery was at first not as divisive an issue as it would become, and that the Democratic Party after Jackson successfully managed a coalition of southern and northern interests. Others have emphasized the role of local political issues and changing economic conditions. Weingast includes these factors in a narrative that stresses explicit political arrangements, especially the following *rule of balance*: slave states should remain equal in number with free states, so as to provide the South with a veto power in the Senate. The narrative records the crises that the Union underwent each time the admission of a new state threatened the balance. The first crisis led to the emergence of the compromise rule, which helped resolve the second, but did not work with the third. This ultimate failure depended on an admixture of economics and politics: to keep an effective balance despite the continuing expansion to the West, the slave economy would have had to develop beyond its feasible limits.

Weingast includes three formal models in his narrative, the first of which is of the spatial brand of voting theory. This model aims at weighing the political influences on the politics of the Union of the agrarian South, commercial Northeast, and intermediate Northwest, respectively. When these three actors differ only on the economic dimension, the Northwest acts as an electoral pivot, and the Union as a whole inclines in the direction of agrarianism, because the Northwest is closer to the South than the Northeast on this particular dimension. However, if slavery enters the political debate, the last part of conclusion does not necessarily hold because the pivotal Northwest is closer to the Northeast than the South on that dimension. The spatial model makes it possible to clarify coalitional possibilities when the political debate is so enlarged. Following a related treatment, Riker (1982) had famously claimed that it was to some Northeast politicians' advantage to introduce slavery on the electoral agenda, and it became a political issue after 1830 precisely for that reason (the so-called *Riker thesis*). The other two models used in Weingast's study are extensive form games of a straightforward sort. Comparison of their subgame

perfect equilibria shows that giving the South a veto power has the effect of blocking the compromise between the Northeast and the Northwest that would otherwise prevail. This reinforces the general point that the rule of balance was an important component of federal stability in antebellum America.[4]

Case 5 ICO (Bates) From 1962 to 1989, the International Coffee Organization (ICO) regulated the international prices of coffee by setting quotas on the exports of its members, notably Brazil and Colombia, which were the main producers. Bates accounts for the birth, regular functioning, and final collapse of this institution. This involves him using, or at least mentioning, various game-theoretic tools, but the study nonetheless follows a classic narrative structure, with a beginning, an end, and intervening steps that exactly reproduce the objective sequence. While the other studies state their explanatory problems in advance and subject their narrative parts to the solution of these problems, this one lets its explanatory puzzles and answers emerge as the story unfolds.

The birth of the ICO raises one such explanatory puzzle. As early as the 1950s, Brazil and Colombia had a cartel policy of restricting quantities and boosting prices, and tried to attract other coffee producers to this policy. However, it is only at the beginning of the 1960s that they succeeded in doing so and thus became able to establish the ICO. Bates explains why success was delayed by arguing that the proximate cause of its establishment was that the United States became favorable to the cartel policy. This is a paradoxical answer because the United States was on the consumption not the production side. Bates makes this answer plausible by relating how Brazil and Colombia, having failed in their first attempt, turned to the US State Department brandishing the communist threat and the long-term advantages of a cartel organization, and eventually met with success when some large US coffee-selling companies decided to support their lobbying. Each of these small narrative segments is followed by an allusive formal argument that clarifies the strategic situation.

Some Defining Characteristics of Analytic Narratives

The five studies of *Analytic Narratives* have a common focus on institutions, and more precisely on institutions that either belong to the internal state organization (Cases 1, 2, 3, and 4) or indirectly depend on it (Case 5, which deals with

[4]Perhaps because its topic has been heavily researched, Weingast's study seems to have aroused special attention from readers of *Analytic Narratives*. Some have complained that it is not clear whether the rule of balance in the Senate was central to the stability of the Federation, and accordingly how much its collapse contributed to the civil war. A quick answer may be that the study selects one particular sequence of actions and events for investigation, and this provides a partial but real explanatory argument (which incidentally comes out most clearly at the post-modeling narrative stage, see section "The Role of Narration in Analytic Narratives").

international relations). They also embody a common approach to the understanding of institutions. The key idea is that institutions operate not only through the formal rules that overtly define them but also through implicit rules of behavior that guarantee their function given the way participating agents respond to them. As an example, the *podesteria* can be viewed either by considering the official terms of employment of the *podestà*, or, more relevantly for an explanatory purpose, as a system of mutual threats between the clans and the *podestà* that made it possible for the latter to fulfil his role effectively. This heuristic is implemented by representing institutions in terms of equilibria of interactive processes, and it is at this juncture that game theory comes into play. Thus, Greif devises a game in which the existence of an equilibrium demonstrates that mutual threats can credibly balance each other, which secures compliance of the clans with the *podesteria* institution.

Conceptualizing institutions, and more specifically political institutions, as *equilibria of interactive processes*, whether by game-theoretic or other means, is a significant contribution to the neo-institutionalist school of thought. Elsewhere, Greif (2006, Chap. 1) clarifies the differences between this "self-enforcement" conception and those of previous neo-institutionalists. These writers had already discussed institutions from the perspective of the agents' interests, but in a somewhat more naïve fashion, often simply assuming that institutions are imposed "top-down" on the agents (what Greif calls the "institutions as rules" conception).[5] Two other major differences, even with historically oriented neo-institutionalists such as North (1981, 1990), have to do with the method of case studies adopted in *Analytic Narratives* and the specific attention this work pays to the role of narratives.

More crucially for our purposes, the five studies illustrate the two principles stated at the opening of this chapter. Each exhibits a collaboration between narrative writing and the employment of analytic tools, and each borrows these tools from the theories of rational decision-making. In the rest of this section, we expand on the second principle, using the studies as reference material. We defer an examination of the first principle to section "The Role of Narration in Analytic Narratives," where the third principle will also be considered.[6]

[5] An answer to the objection echoed in fn 3 is forthcoming along this line. The *podesteria* may well have been imposed "top-down" upon the Genoese from the outside, yet the question arises nonetheless of why they made its functioning possible, a question that the "institutions as rules" conception addresses. Clark's (2007) otherwise critical account of Greif (2006) clearly recognizes this.

[6] *Analytic Narratives* has given rise to a rather large number of discussions, which space reasons prevent us from covering here. The reader may in particular consult *American Political Science Review*, 94, 2000, no 3, and *Social Science History*, 24, 2000, no 4, which contain one or more reviews followed by a rejoinder from the five authors. Some of these discussions express strong scepticism about either the individual contributions or the methodological project itself; among the reasons for this scepticism is the routine complaint that "rational choice theory" (whether formal or not) is either flawed or inapplicable. None of these discussions – even the favorable ones – properly recognizes the special function that narratives, as against other forms of reporting of historical events, fulfil in the AN methodology.

Formal models of varying precision and sophistication occur in Cases 1, 2, 4, and 5, though not in Case 3. One may thus wonder whether or not AN in general must involve a formalism. As for the studies that do employ one, they primarily rely on extensive form games of complete information, and one may thus wonder how much this choice of models matters to AN. This section answers these two questions by arguing that (i) *AN do require formal models* and (ii) *AN can borrow these models from any formal branch of the theories of rational decision-making*. These two answers spell out what the five authors have only suggested (in *Analytic Narratives*) or briefly stated (in their informative rejoinders to critics; see Bates et al. 2000a, b). Thus, they acknowledge that their "restriction of models to extensive form games limits the range of issues [they] address" (2000b, p. 691), and they endorse "the requirement of a formal model" (p. 693). This amounts to making claims (i) and (ii).

One argument for restriction (i) is that ignoring it would take the edge off the AN methodology. Historians already borrow from common-sense ideas on individual rationality to confer explanatory value on their narratives. But they rarely make these ideas explicit, and this may be for two reasons: they may regard them as being too banal to be stated, or they may consider that a fuller statement would break the narrative flow. As they are not subject to the same discursive constraints as ordinary narratives, AN can unfold the terse suggestions contained in the latter and thus attempt to enhance their explanatory value. To do that, AN bring in specialized concepts of individual rationality: but if they eschew formalization of these concepts, they may be hardly different from the scholarly expansions that historians append to ordinary narratives in the introductions, conclusions, and appendices of their works. Interestingly, when revisiting *Analytic Narratives*, Levi (2002, p. 109) claims that the essays in this book "do not represent a breakthrough." It seems more fruitful, however, to allow that they do introduce something new – for better or worse. With (i) included in their definition, AN create an unusual tension between the narrative and the formal modeling. How this tension can be managed is the most exciting problem AN raise for the methodologically minded social scientist.[7]

There is actually a more direct reason for supporting restriction (i). The AN methodology crucially relies on the concept of equilibrium, understood one way or another, and the full development of this concept plainly requires a formalism. Levi rightly observes that *Analytic Narratives* makes extensive use of comparisons of equilibria: "the emphasis is on identifying the reasons for the shift from an institutional equilibrium at one point in time to a different institutional equilibrium at a different point in time" (2002, p. 111). This is the method of *comparative statics*, which economists implemented and made famous before passing it on to other social scientists. However, the method will remain at a heuristic level as long as one does not select a formal theory – e.g., that of extensive form games of complete information – and specify a model within that theory – e.g., by fixing a set of players, a set of strategies and preference orderings that fully define the game to be studied. The comparative statics

[7]One of the first works to explore this tension is the collection by Grenier et al. (2001); it does not yet refer to AN.

exercise, which is quantitative by essence, is possible only if some of the data of the modeling stage – e.g., the preference data of the game – are stated parametrically, with a range of numerical values set for each parameter. The exercise then consists in deducing how the equilibria change as a consequence of the parameters varying within their ranges. Even more clearly than in *Analytic Narratives*, this will be illustrated in Zagare's (2011) *The Games of July* (see Cases 9 and 10 in next section).[8]

Having defended restriction (i), we move to the generalization proposed in (ii). A common thread between Cases 1, 2, 4, and 5 is their reliance on non-cooperative games of extensive form. In their introductory manifesto, the authors defend this particular form on the ground that they focus on "sequences of actions, decisions, and responses that generate events and outcomes" (Bates et al. 1998, p. 9; see also Levi 2002, p. 111). The implicit claim is that the sequence of moves in the game is capable of paralleling a concrete sequence of actions and reactions by historical actors. One may, however, doubt that such nice parallelism really takes place. Consider the *podesteria* game; it involves moves such as "to challenge," "to fight," "to prevent," which can hardly represent genuine actions by the clans or the *podestà*. If the moves are idealizations, their sequential ordering cannot represent the passing of historical time; and indeed, if the *podestà* plays last rather than first in the game, this is for theoretical convenience, not descriptive accuracy. Of course, this semantic observation does not entail that extensive form games are unimportant to AN but only that this genre has no privileged association with them. Cases 6 and 7 in the next section will show that games in normal form can be equally relevant. On a different note, Cases 1, 2, 3, and 5 share a limitation by only considering extensive form games of *perfect information* (or possibly extensions of these to exogenous uncertainty). This conveniently guarantees that solutions to the games can be found by backward induction, as the subgame perfect equilibrium requires in this case. As next section will also show (Cases 9, 10, and 11), AN can support the more sophisticated formalism of extensive form games of *incomplete, hence also imperfect information*, along with the perfect Bayesian equilibrium concept, which is commonly used to solve these games.[9]

AN can be developed in other technical directions than noncooperative game theory. Cooperative game theory may provide appropriate models when it comes to analyzing the formation of coalitions, as in the ICO case.[10] Furthermore, not every historical state of affairs that involves multiple individuals calls for a game-theoretic analysis: individual decision theory, whether of the expected utility form or others,

[8]In this paragraph, we imply the familiar conception of a *model* as a construct mediating between theories and real objects. The alternative conceptions canvassed in recent philosophy of science could also be brought into relation to AN.

[9]On the logical relation between the concepts of perfect information and complete information, see the texts by Fudenberg and Tirole (1991), Harrington (2009), Myerson (1991) and Osborne and Rubinstein (1994). The first text in this list is the *locus classicus* for the perfect Bayesian equilibrium concept.

[10]Bates suggests considering the Shapley value for this purpose. For this and other concepts of cooperative game theory, see Myerson (1991) or Osborne and Rubinstein (1994).

may be sufficient for the modeling purpose. This is the case by definition when the multiple individuals face natural uncertainty, but also and more subtly when social uncertainty can acceptably be represented *as if it were natural*. When commenting on Clausewitz's military narratives, Mongin (2009) argues that some of his judgments can plausibly be reconstructed as expected utility comparisons; this is so despite the fact that the situations are strategic in an intuitive sense. There seems to be no rule to determine when game theory is indispensable to the analysis of interaction and when it is not. It is useful to register this indeterminacy, and then avoid restricting the technical apparatus of AN beforehand.[11]

Analytic Narratives from Military and Security Studies

While the previous AN belong to the historical part of political science, those covered in this section mostly belong to the historical parts of military studies (Cases 6 and 7) and security studies (Cases 8, 9, and 10). We distinguish between the last two fields as follows: military studies is concerned with actual war actions – battles, campaigns, guerillas, information wars, and the like – and security studies with actions taken under the shadow of war, i.e., facing the possibility of wars that may or may not break out. As two game-theoretic contributors to security studies write, "the games we analyze are not war games as such, but the choice that players make may precipitate conflict that leads to war" (Brams and Kilgour 1988, p. 3). This basic distinction is sometimes overlooked, which is unfortunate because military and security studies have different conceptual orientations.[12]

Case 6 World War II Battles (Haywood 1954) Shortly after World War II, Haywood, then a colonel in the US Air Force, discovered von Neumann and Morgenstern's (1944) work and made the first ever application of game theory to war events, publishing a sketch in Haywood (1950) and a detailed version in Haywood (1954). For this application, he selected a 1943 naval battle in the Pacific War and a strategic turning point in the 1944 Normandy campaign. His main concern was to connect the US military doctrine of decision, which was prescriptive, with von Neumann and Morgenstern's Min-Max solution for 2-person zero-sum games, which he also regarded as being prescriptive. He argued that the military doctrine of decision was right in prescribing officers to act on an estimate of the enemies' capabilities, not on a guess of their intentions – the argument being that if the game has no Min-Max solution in pure strategies, guessing intentions leads to an infinite regress of strategic calculations. Despite this prescriptive orientation, Haywood's inquiry has some bearing on historical explanation.[13]

[11]Schiemann (2007) promotes a further extension of AN to behavioral economics and illustrates this by a study of an event from the Yugoslav civil wars in the 1990s.

[12]See Betts (1997) for a more thorough discussion that includes a history of security studies.

[13]Haywood has rather mysteriously disappeared from the academic scene, despite Brams's (1975) and Harrington's (2009) supportive reviews of his contribution.

In the Bismarck Sea battle of February 1943, the US Air Forces destroyed a naval Japanese convoy that was sailing from Rabaul on New Britain Island to Lae on the New Guinea coast. Of the two possible routes, north of New Britain and south of it, the Japanese commander had chosen the former. Unaware of this, the US general in charge had to choose between concentrating his reconnaissance flights on one route or the other, and he actually took the northern option, whence his crushing victory. Haywood argues for the rationality of both moves, including the Japanese one. Having pointed out that the northern route was mistier than the southern one, he computes the number of bombing days associated with each of the four possible outcomes and solves the resulting 2×2 zero sum game using von Neumann and Morgenstern's solution. Here Min-Max reasoning on the two sides leads to an outcome that fit the facts, which confers some explanatory value on this reasoning. The other application is the Avranches battle fought between US General Bradley and the German general von Kluge in August 1944. Haywood analyzes it in terms of a 3×2 matrix, which this time has no pure strategy solution. As he does not quantify the payoffs, he cannot exhibit mixed strategies solutions, and this leaves one in doubt about what he achieves in terms of explanation.[14]

Surprising though it seems, game theory rarely enters military studies as properly defined. Applications of a prescriptive or instrumental nature certainly exist, the best-known being those pursued in the 1960s and 1970s for the RAND Corporation and some US military agencies (see, e.g., Erickson 2015). But it seems as if the historical part of military studies has no genuine game-theoretic application to offer besides Haywood's and the next case to be reviewed. This is not to say that scholars in this area have no interest in history. To the contrary, there is a long tradition among military strategists, which dates back to Jomini and Clausewitz, of basing their thinking on a careful examination of past battles and campaigns. However, this tradition is almost entirely narrative in the ordinary sense; so much so that it acted as a foil to the anti-narrative position epitomized by the *Annales* school historians in the middle of twentieth century, notably Braudel (1969).[15] By revisiting the Waterloo campaign, Mongin (2008, 2018) attempts to show that it is possible to turn even a worn-out example of military narrative into an AN.

Case 7 The Waterloo Campaign (Mongin 2008) As is well known, Napoleon's return to power in 1815 ended with his resounding defeat by Wellington and Blücher on the battlefield of Waterloo in Belgium. On June 16, the campaign began favorably for him, with the French beating the Prussians at Ligny, near Charleroi. On June 17, Napoleon decided to send a large detachment under Marshal Grouchy against the defeated Prussians, and he took the rest of his army to Waterloo, near Brussels, where the English and Dutch were ready for a defensive battle. On June 18, the French failed to break through the enemy lines and were eventually crushed when

[14]At any rate, a later historical discovery showed that Bradley had in fact been cognizant of the orders received by von Kluge, as the Allies had broken the German Enigma code (see Ravid 1990).

[15]Clark (1985) conveniently summarizes the position taken by post-war *Annales* historians; see also Stone's (1979) critical discussion.

the Prussians came as additional help. Despite innumerable histories, an explanatory gap remains in this sequence: why did Napoleon decide to send out Grouchy's detachment? By doing so, he ran the risk of not having it on his side when he faced Wellington, or, much worse, Wellington and Blücher together if they managed to join forces, a possibility that effectively materialized.

To make progress with this explanatory question, Mongin proposes a zero-sum game in normal form with two players, Napoleon and Blücher, allowing for uncertainty in several ways. First, both Napoleon and Blücher are uncertain of which battles will result from their independent decisions. Specifically, Napoleon can keep his army united, dispatch Grouchy for a pursuit of Blücher, or dispatch it for interposition between him and Wellington; and on his part, Blücher may either retreat to Germany or try to join Wellington at Waterloo. This is nothing but standard strategic uncertainty. However, a second form of uncertainty enters, since Napoleon does not know Blücher's type – here whether or not the latter was badly weakened after Ligny – and this means that the game is one of incomplete information. Third, external circumstances matter besides the players' decisions, and both are *ex ante* uncertain of the issue of each given battle. This is nonstrategic information, which is treated here as if it were objective and amenable to common expected utility calculations by Napoleon and Blücher.

Given suitable parameter restrictions, von Neumann and Morgenstern's Min-Max solution delivers a unique equilibrium, which involves pure strategies. As in Haywood's example, this arguably delivers not only an equilibrium but also rational choice recommendations. Napoleon should choose to dispatch Grouchy for interposition, and Blücher should try to join Wellington. That Napoleon effectively chose interposition, rather than mere pursuit, can only be conjectured from the historical record, but the game reinforces this hypothesis. The *ex post* failure is not an objection since it could result from an unfavorable resolution of objective uncertainty and from Grouchy misapprehending the plan – some historical evidence points in these two directions. Overall, the study exemplifies how an analytic narrative can be both formal and interpretive, since assumptions and conclusions are assessed in terms of evidential reports that are always incomplete, equivocal, and, given the high stakes, unavoidably biased. The conclusions adjudicate among existing positions, indeed by reinforcing classic pro-Napoleonic arguments against equally classic anti-Napoleonic ones.

Case 8 The Cuban Missile Crisis (Various Authors) Few diplomatic events have raised more scholarly interest than the crisis that took place from October 16 to 28, 1962, between the United States and the USSR. On 16 October 1962, President Kennedy was shown U2 photographs demonstrating that the Soviet Union was building missile bases in Cuba. Kennedy and his advisers pondered over several options, which included doing nothing, making a diplomatic move, bombing the missile sites, and blockading Cuba with the US Navy. Deliberation and further investigation led to the blockade decision of 22 October. Khrushchev was notified and the blockade was then publicly announced to the nation. In the ensuing days, the

crisis deepened, with some secret diplomacy nonetheless taking place. It was eventually resolved on 28 October, when Kennedy and Khrushchev managed to coordinate on a compromise solution. Essentially, in return for the USSR removing its missile systems from Cuba, the United States would lift the blockade, pledge not to invade Cuba, and – this was a later and secret part of the deal – remove missiles from Turkey.

Innumerable accounts of this famous sequence have circulated, with the flow being sustained by the appearance of declassified secret material (see, e.g., Allison's 1999 revision of his classic 1971 study). Among the accounts based on, or inspired by, game theory, none seems to us sufficiently rich in narrative content to qualify as an AN. Rather, they treat the Cuban Missile Crisis as a mere application of theoretical ideas, and, if we include this topic here, this is because it offers a touchstone of *deterrence models*, which recur in the AN literature. As Zagare (2014) has explained, the game-theoretic literature on the Crisis has gone through three essentially different stages. While the first authors, like Schelling (1960), gestured towards game theory rather than actually used it, a second wave from the mid-1970s onwards exploited 2×2 normal form games such as the Prisoner's Dilemma and Chicken, sometimes adding ingenious variations to them. A third wave, which began in the mid-1980s, adopted extensive form games, whether of complete or incomplete information, to analyze deterrence. Examples of the second wave appear in Brams's (1975) *Game Theory and Politics*, and at a more advanced level, in his *Superpower Games* (1985).[16] For the third wave, which he associates with a "sea change," Zagare (2014) mentions an early model by Wagner (1989) and his own "perfect deterrence theory" (as developed in Zagare and Kilgour 2000), both of which involve extensive form games of incomplete information. This leads us now to a specific examination of Zagare's contributions to AN.

Case 9 The Moroccan Crisis of 1905–1906 (Zagare 2015) After they had overcome their conflict over Sudan in 1898, France and Britain moved towards an alliance that materialized in formal agreements in April 1904 (the so-called *Entente Cordiale* agreements). The most important of them involved a trade of influences, with France supporting Britain's leading role in Egypt, and Britain supporting France's freedom of action in Morocco (with some role conferred on Spain in the northern part). Having not been consulted at all over Morocco, and also responding to the Sultan's wish to counter France's threatening influence on his country, Germany championed its sovereignty and an "open door" policy for foreign trade and investments. The German diplomatic pressures on the French

[16]In the latter work, Brams introduces a 2×2 game that schematizes the American and Soviet choices (Blockade and Air Strike, Withdraw and Maintenance, respectively) and applies his "theory of moves" to find that the Compromise issue (Blockade, Withdraw) emerges as a "non-myopic" equilibrium.

government led in 1905 to the resignation of foreign minister Delcassé and the reluctant acceptance by Président du Conseil Rouvier of an international conference. Officially devoted to the economic and administrative reforms that Morocco needed, the conference took place from January to April 1906 in Algeciras. While Germany hoped to drive a wedge into the Entente Cordiale and score a diplomatic victory over France, Britain supported its ally and Germany ended up almost entirely isolated, getting only limited concessions that would not suffice to curb France's colonial activism.

Zagare (2015) revisits the 1905–1906 events by appealing to the *Tripartite Crisis Game*, which belongs to his more general "perfect deterrence theory." This is an extensive form game of incomplete information with three players acting sequentially as follows. Challenger can either keep to the *status quo* or make a demand on Protégé, who can either concede or hold firm, in which case Defender enters the stage by either supporting or not supporting Protégé. If Defender has supported Protégé, Challenger plays again by either backing down or accepting a conflict, and if Defender has not supported Protégé, the latter plays again by either backing down or realigning on Challenger's side. Information is incomplete in that each player has two possible types: Challenger may be "determined" or "hesitant," Protégé "loyal" or "disloyal," and Defender "staunch" or "perfidious." This figurative terminology captures the fact that, for each nation player, some of its preferences over the terminal nodes are unknown to the other two players. To deal with the 1905–1906 crisis, Zagare specializes the Triple Crisis Game by assuming that Challenger – here Germany – is "determined"; this technically means that, at the last stage, Challenger prefers to accept a conflict rather than to back down. Thus, incomplete information is limited to Protégé and Defender – here France and Great Britain, respectively. Technically, Protégé is "loyal" if, at the last stage, it prefers backing down to realigning, and Defender is "staunch" if it prefers reaching the node where Challenger accepts the conflict to reaching the node where Protégé realigns. Fixing initial probability values for Protégé being "loyal" and Defender being "staunch," Zagare shows how they get revised at the perfect Bayesian equilibria he computes. In the Moroccan study as well as in other recent articles and in his 2011 book, Zagare explicitly claims to be using the AN methodology.

Case 10 The July 1914 Crisis (Zagare 2011) *The Games of July* (2011) investigates the diplomatic events that decisively contributed to the outbreak of World War I, particularly emphasizing four historical turning points. The first deals with a remote but influential decision made by Bismarck in 1879 to offer a military alliance to Austria, despite the tension this created with Russia, which was the main target of this arrangement. The second relates to the unqualified support – or "blank check" – Austria obtained from Germany in early July 1914 to crush Serbia, and the third to the escalation of conflict with the other powers once Austria began taking action. The fourth is devoted to the British decision to maintain an ambiguous policy during the July crisis, a decision that may have

misled Germany in believing in Britain's neutrality and thus may have contributed to the outbreak of the war. Each turning point raises specific explanatory problems that a brief narrative and review of historical literature helps locate. The book answers them through the instrumentality of game-theoretic modeling, in accordance with a methodology that the author distils in preliminary chapters and identifies with that of AN (see also Zagare 2009).

Each case relies on a game of its own, although all are taken from the common shelf of "perfect deterrence theory." The first, second, and fourth sequences are handled by means of relevant variations of the Tripartite Crisis Game, and the third by means of the *Asymmetric Escalation Game*, which also belongs to "perfect deterrence theory." We focus on the fourth case, which is concerned with British policy, because this permits comparisons with the Moroccan case, in which this policy had already played a crucial role. The Liberal Grey, who had succeeded the Conservative Lansdowne at the Foreign Office in the midst of the Moroccan crisis, essentially pursued his predecessor's policy of supporting France without making any military commitment to it. The persisting problem for the British was to secure peace on the continent by combining deterrence (of the Germans) and restraint (of the French, especially in their support to the Russians), and this led them to foster ambiguity on their final intentions. Whether it was to the point to maintain this carefully balanced policy in July 1914, as Grey did, is a major historical question. As is well known, it eventually needed Germany's invasion of Belgium on August 4 for Great Britain to engage militarily on the side of France and Russia.

As in Case 9, the game-theoretic treatment proceeds from the Tripartite Crisis Game. Germany, France, and Great Britain still occupy the roles of Challenger, Protégé, and Defender, but this time Defender is "staunch" and Challenger may be either "determined" or "hesitant" (the latter means that, at the last stage, Challenger prefers to back down rather than to accept a conflict). As the "blank check" issued to Austria was not known to the other players, endowing Germany with two types appropriately represents this uncertainty, but it is not obvious that the staunch type describes how Britain was perceived in July 1914.[17] However, interesting mixed equilibria occur even under this limiting assumption, and they are consistent with Grey's "straddle strategy," as Zagare (2011, p. 160) aptly designates it, thus providing the British diplomacy with a rationale (see also Zagare and Kilgour 2006). As these equilibria seem compatible with the strategic situation more broadly, they might serve to capture the protagonists' effective interaction. Supposing they are indeed the historically relevant ones, the war would have broken out not because of Britain's ambiguity, which had a serious

[17]The assumption made regarding Britain's type appears to be connected with a mathematical difficulty. The Triple Crisis Game can currently be solved only in limiting cases. Concerning the Moroccan crisis, the restriction was that Challenger was "determined," and here it is that Britain is "staunch" (Zagare 2015, p. 335, fn 7).

intent even though it was a gamble, but because in 1914, unlike in previous earlier crises such as the Moroccan one, the gamble turned out badly.

We close this section with another case that does not belong to either military or security studies but rather to historical political science. We include it here because it involves an extensive form game of incomplete information and the use of perfect Bayesian equilibrium as in Cases 9 and 10. The game belongs to the class of deterrence models, which bridges the work in security studies with some of the work in political science.

Case 11 (Nalepa 2010) During the transition from communism to democracy experienced by eastern European countries, communist officials had to choose, roughly, between opposing the trend as much as they could and seeking a deal with the democratic opposition. In this deal, they would retreat from power in return for a promise that they would not later be banned from public functions. Nalepa (2010) starts from the observation that, in some countries, the communists reached such a compromise with the democrats and the democrats kept their promise (at least by and large and for some period of time). This is puzzling because the democrats had every reason to renege on the compromise. However, as the author argues, the communists did have a way of avoiding this. The communist secret police had once infiltrated opposition movements, so the democratic parties, which were heir to these movements, were themselves in some danger of falling prey to a ban or similar "transitional justice" measures. Only the communists knew the extent of the infiltration, and this gave them an informational advantage over their opponents. An explanation of the historical puzzle is forthcoming along these lines, and Nalepa, who claims to be using the AN methodology, substantiates it by game-theoretic modeling.

The first model, which she attributes to Przeworski, is a perfect information extensive form game in which the communists anticipate the democrats' disavowal and choose opposition rather than compromise. This model is of course a strawman, since it never allows for the possibility of compromise. The second model, which is Nalepa's, introduces asymmetric uncertainty and comes close to a signaling game. By assumption, the communists know exactly the percentage of infiltration among democrats, who, being entirely ignorant, form a uniform probability on this parameter. If the communists choose to compromise rather than oppose, democrats read this move as a signal that they are infiltrated to a significant extent and revise their probability accordingly. This informational exchange is captured in terms of the perfect Bayesian equilibrium concept, which we have already encountered in Cases 9 and 10. The study closes by comparing some of the available equilibria obtained from this concept with historical situations. Unlike in Czechoslovakia, where the communists opposed the democrats until they collapsed, compromises prevailed in Poland and Hungary, and for relevant parameter values, there exist equilibria related to these situations.

Analytic Narratives and Deductive Explanation

The studies covered in sections "The Five Studies of *Analytic Narratives*" and "Analytic Narratives from Military and Security Studies" suggest some generalizations regarding the explanatory potential of AN. First, with the exception of Case 5, which belongs to recent history, and Case 6, which similarly belonged to recent history when it was written, they rely on an extensive scholarly record they use not simply to determine the factual data but also to suggest problems to be solved. The record is usually of the traditional narrative brand, and they identify the problems by noticing explanatory gaps within it. For example, Case 1 revisits the alternation of civil war and peace in Genoa with a view to explaining it, which had not really been done before. Case 3 revisits the establishment of universal conscription with a view of synthesizing explanations that hitherto had only been partial; and Case 7 revisits the Waterloo campaign with a view to arbitrating a classic disagreement among historians. Moreover, the problems appear to have been selected by carefully considering what the importation of analytic tools could add to the more traditional treatment. As Bates et al. (1998, p. 13) write, "our cases selected us, rather than the other way around." The only exception here appears to be Case 2, the topic of which – the Ancien Régime finances in comparative perspective – is arguably too wide for the analysis to get much grip on it. Mongin (2008, 2018) goes as far as to claim that starting from the extant historical literature, defining problems based on the lacunas therein, and restricting the models to limited fragments of it are preconditions for one's adopting the AN methodology.

One may observe, however, that the problems are not exclusively of an explanatory nature.[18] The Waterloo study aims not only at ranking competing explanations but also at substituting some missing data – what instructions Napoleon gave to Grouchy – with a deduction from the model. Here the gaps in the earlier narratives concern *the facts of the matter*, and not the explanation by itself. Less ambitiously than this substitutive role, though, AN can orient factual research in novel directions, as do other forms of problem-inspired history, like that promoted by the *Annales* school. However, original scholarship has thus far been exceptional among AN contributors, and they do not seem yet to have moved existing scholars towards new agendas.

From the angle of the philosophy of explanation, AN seem naturally to connect with the *deductive scheme* proposed by Hempel (1965), Nagel (1961), and many others.[19] Both the *Analytic Narrative*s team and Zagare mention this well-known scheme (though without detail).[20] Broadly speaking, it postulates that to explain a particular fact is to deduce the statement that the *explanandum* fact occurs from statements that other facts occurred, along with statements of generalities so as to effect a connection – these facts and generalities being offered as the *explanans*.

[18]Discussions of AN rarely make this point; see, however, Downing (2000, p. 91).

[19]Useful critical summaries appear in Salmon (1992) and Bird (1998).

[20]See Bates et al. (1998, p. 12; 2000a, p. 697) and Zagare (2011, pp. 5–7).

Besides requiring the logical correctness of the deduction, the scheme places epistemic requirements on the constitutive statements, and philosophers of science have dissenting formulations here. However, they all agree on the two basic points that the *explanandum* statement must be known to be true and the *explanans* statements must be, if not necessarily known to be true, at least empirically well supported. The rest of this section discusses the extent to which AN explanations fit with the deductive scheme; we will exhibit significant discrepancies, and thus prepare our ultimate claim that narratives are an essential component of these explanations. This discussion first singles out the *deductive requirement* of the scheme (not to be confused with the scheme itself) and then proceeds to the *epistemic requirements* that the scheme also involves. Since we mean to follow the existing literature, we focus on the use of game theory; but our conclusions can to a degree be generalized to the other formal theories discussed in section "Some Defining Characteristics of Analytic Narratives."

If the deductive requirement is to apply to AN effectively, it needs to be adapted to the distinction between *statics* and *comparative statics* that runs across them. One may consider a game either *specifically*, i.e., for fixed values of its parameters, or *generically*, i.e., by not restricting the parameters or (more commonly) restricting them minimally.[21] Although not entirely sharp, this distinction points towards two different possibilities for deduction. What can be deduced in the case of a specific game is that given outcomes occur as equilibria of that game, and in the case of a generic game, that different outcomes occur as equilibria when the instantiation of that game changes with the parameter values. This is the statics versus comparative statics distinction, as it emerges from the use of game theory.[22] Correspondingly, one may either explain a given historical fact by associating it with an equilibrium of a specific game or explain a change in historical facts by associating it with a change in the equilibria of a generic game. Comparative static explanations are logically more powerful than static explanations and should be preferred in principle. However, the AN literature makes it plain that comparative static explanations are not easy to come by. An exceptionally clear example appears with Zagare's multiple versions of the Triple Crisis Game, each of which is investigated in terms of comparative statics.[23] Haywood's simple analysis of two specific games represents the polar opposite case of a merely static explanation. Although more sophisticated, Greif's, Rosenthal's, and Mongin's analyses fall more on the static side, since they let parameters vary only in relation to given equilibria, so as to secure sufficient conditions for the existence of these equilibria, and they do not study the dependence of equilibria on parameter values across the full range of possible values.

[21]Game theory has another technical sense for the word "generic"; this will not be considered here.

[22]Other formal theories would specify the distinction somewhat differently.

[23]On the definitions given above, the Triple Crisis Game is not a generic game, but rather *a set* of such games. For instance, the specialized version for the Moroccan crisis is one such generic game, and the specialized version for Grey politics is another.

Having clarified this preliminary distinction, we can explicate two difficulties AN must face in seeking to satisfy the *deductive requirement*. The first has to do with the *multiple equilibria* occurring under most equilibrium concepts used by AN. When the multiplicity occurs in a specific game or – for some fixed values in the parameter range – in a generic game, the game-theoretic assumptions do not suffice for a definite conclusion, and the deductive machine needs supplementing by some external selection procedure. When the multiplicity occurs in a generic game simply because the parameters change, the situation turns out for the better; now the deductive machine works autonomously. Then the next step will be to compare the equilibria and their underlying parameter values with the available historical evidence. Contributors who claim that the deductive scheme is relevant to AN seem to have this favorable case in mind.[24]

Second, there is the troubling problem of deciding what in the games plays the role of the *generalities* that the *explanans* must contain if the deductive requirement is to come into effect. In a static exercise, the natural candidate for this role is the chosen equilibrium concept, e.g., subgame perfect equilibrium, the Min-Max solution, and perfect Bayesian equilibrium. In a comparative static exercise, it makes sense to consider as generalities not only the equilibrium concept but also the generic game (or, at a higher level, the class of generic games, the Triple Crisis Game being such a class). So, we do find general statements in AN explanations. However, they are general only in the sense of being expressible as logically universal statements, not necessarily in the deeper sense of being *nomological*, that is of counting as putative laws of nature. Hempel's (1965) paradigmatic version of the deductive scheme proposes various conditions, besides the logical form, for a generality to be nomological.[25] The question arises whether the game-theoretic statements singled out above meet these conditions, but we will not try to answer it here, being content with stressing its relevance. AN contributors who allude to Hempel (Zagare 2011, approvingly; Bates et al. 1998, disapprovingly) do not seem to have gone far into it. The common view among them is, more loosely, that the game-theoretic pattern uncovered in one study can be transferred with some success to other studies. Some go farther and claim to have at their disposal theories that apply across a significant range of historical states of affairs. This is the case with Greif (whose "theory of endogenous institutional change" includes the Genoa study as a particular application) and Zagare (whose "perfect deterrence theory" encompasses most of his studies relative to World War I and, as he suggests, can also be applied to some contemporary events).

We now consider how AN satisfy the *epistemic requirements* of the deductive scheme of explanation. As noted above, AN typically draw their problems from the extant historical literature and use little more than this corpus for checking their solutions empirically. To pass this empirical test, they need to answer three questions

[24]See Bates et al. (1998, p. 15) and Zagare (2011, p. 16).

[25]These and other conditions have been thoroughly discussed in the philosophy of science; see, e.g., Bird (1998, Chap. 1).

in the affirmative. (i) Do the equilibria of the games approximate what historians have observed concerning the *explanandum*? (ii) Do the game-theoretic assumptions that constitute the *explanans* draw support from what historians have observed concerning the circumstances of the *explanandum*? (iii) Is the *explanans* independently supported, i.e., does it also draw support from what historians have observed concerning other states of affairs than those under current investigation? We will review these questions in turn.

Regarding (i), there appears to be a gap between Cases 1, 2, 3, and 4 of section "The Five Studies of *Analytic Narratives*," and Cases 6, 7, 8, 9, and 10 of section "Analytic Narratives from Military and Security Studies." The *explananda* of the second group are narrowly circumscribed in time and space, directly bear on interactive decisions, and often if not always involve designated individuals, such as Bradley, Napoleon, or Grey. By contrast, the *explananda* of the first group extend rather widely across time, space, or both, bear on institutional or organizational facts rather than interactive decisions as such, and without exception involve collective actors, such as clans, political elites, or regions. To be linked to game-theoretic equilibria, the observable *explananda* of the first group need to undergo a more thorough abstraction process than those of the second group. This makes their explanations prima facie more debatable than the others are. Pushing this line, Mongin (2018) recommends applying the AN methodology preferentially to *explananda* that share with military and security *explananda* the convenient properties of being spatiotemporally well defined, and involving recognizable decisions made by recognizable historical actors. However – as also pointed out by Mongin – such recommendations threaten to trivialize AN. The explanations of the first group are more challenging than those of the second, which may remain too close to the historians' accounts to bring much illumination. It seems as if a balance needs to be struck between the two dangers of arbitrariness and pedestrianism.

When it comes to (ii), the question of the identity of players arises again, and there are now the further questions of endowing them with relevant *strategy sets* and *preference orders*. AN keep the number of players to a bare minimum. This may be easier to accept when players are hypothetical constructs, as in the first group of studies, than when they are identifiable historical figures, as in some studies of the second group. Indeed, somewhat shockingly, Grouchy does not enter the Waterloo game, and the games for the July crisis never include all major powers together.[26] Technical convenience explains these lacunas: thus reduced, the Waterloo game can accommodate some informational complexity, and the July 1914 games can be resolved despite their rich informational structure. For similar reasons, AN tend to rely on rather small sets of pure strategies. To allow for mixed strategies enlarges the players' possibilities, but like much of game-theoretic economics, the AN literature is reluctant to take this option; only Cases 9 and 10 make a significant exception.

[26]There were five at the time: Britain, France, Germany, Austria, and Russia. Zagare and Kilgour (2006, p. 635) address this objection.

Historians will no doubt complain that the definition of both players and strategies in AN impoverishes or distorts the historical evidence.

The definition of the players' preferences is even more problematic. A modestly sized set of strategies, and hence of outcomes, is already enough to turn preferences into complex objects. Thus, *podesteria* with seven outcomes, and Triple Crisis with six, induce 7! and 6! possible orderings, not small numbers. Moreover this computation assumes there are no indifferences. Zagare (2015, p. 332) contrasts two inferential methods to define preferences sensibly: one can try to infer them either from the historical actors' observable choices or from plausible general assumptions (such as the standard monotonicity and dominance assumptions of decision theory). Preferences are said to be "revealed" in the former case (which is loosely reminiscent of the revealed preference method in economics) and "posited" in the latter. We understand these two ways as being complementary rather than exclusive. Observed choices, even repeated under different historical circumstances, can hardly provide enough data, since preferences typically make counterfactual comparisons, and general assumptions are unlikely to be sufficient either. Whatever the chosen balance between "revealed" and "posited" preferences, historians will still no doubt complain – this time, though, not by arguing that AN *subtract* too much from the available evidence, but rather that they *add* too much to it.

Question (iii) plays an essential role in all formulations of the deductive scheme of explanation. As Nagel (1961, pp. 43–43) writes, for instance, the point here is "to eliminate explanations that are in a sense circular and therefore trivial because one or more of the premises is established (and perhaps can be established) only by way of the evidence used to establish the [*explanandum* statement]." To require that *all* statements in an *explanans* be tested independently would be exacting, but even the mild form of the requirement with "some" instead of "all" turns out to be challenging. Statements referring to historical particulars are the most recalcitrant, because of the paucity of historical data. Thus, one of the games in Case 1 postulates that clans strike a trade-off between the benefits of gaining control of Genoa and the costs of becoming responsible for its external security, but the sparse historical record does not contain any independent evidence for this assumption.

Explanans statements that are akin to generalities have a better chance of being tested independently. The Triple Crisis Game of Cases 9 and 10 illustrates this possibility. It underlies the *explanantia* proposed for no less than four different historical *explananda* (Germany's choice of an alliance with Austria in 1878, its diplomatic failure at Algeciras in 1906, its "blank check" to Austria in 1914, and finally, Britain's ambiguous policy in 1914). As Zagare and Kilgour (2000, 2003) argue, what is central to the Triple Crisis Game is *the assumption that Protégé can realign with Challenger at the final stage*. Strategically, this gives Protégé leverage over Defender while enlarging the room for maneuver of Challenger, and the Triple Crisis Game thus acquires a flexibility that makes it applicable across various historical situations. This assumption constrains the four *explanantia* above and thus provides a way of testing any of these individual *explanantia* by the empirical success or failure of its neighbor. This establishes that Zagare's explanations meet the independent testability condition at least in part. However, one should of course not confuse

independent testability with successful independent testing. The central assumption of the Triple Crisis Game runs into the historical problem that it applies more convincingly to the 1878 and 1906 contexts, in which Protégé's threat to realign was plausible, than to the 1914 contexts, where this threat made limited sense.[27]

To sum up this section, we have borrowed the classic deductive scheme from the philosophy of explanation and used it as a thread to investigate how AN contribute to historical explanation. This scheme recommended itself because game theory has a deductive machinery, and also because contributors to the field of AN often lay claim to it. We found that AN do not always involve proper deductions, and that they meet the epistemic conditions of the deductive scheme only imperfectly. The next section shows that the narrative component of AN can alleviate these failures. More generally, it considers the role of this component in fuller detail.

The Role of Narration in Analytic Narratives

Consider first the deductive failure connected with the *multiplicity of equilibria*. Authors of AN are aware of this difficulty, and typically resolve it by appealing to their narratives to decide among the possible equilibria.[28] This sketch of an answer needs to be refined by distinguishing between *different kinds of multiplicity*, as we did in section "Analytic Narratives and Deductive Explanation." Suppose the author of an AN wishes to devise an explanation in terms of some generic game. Narrative information has already established what the *explanandum* consists of and is now expected to say what parameter values of the generic game actually prevailed in the circumstances of the *explanandum*. If the generic game associates a unique equilibrium with these values, a dichotomy straightforwardly follows: either the equilibrium agrees with the *explanandum*, and the explanation can proceed further, or there is no agreement, and the explanation has failed. Now consider the case in which the generic game associates several equilibria with the historically relevant parameter values, and exactly one of these equilibria agrees with the *explanandum*. It is not clear whether one may still hope for an explanation. A standard move in applied game-theoretic work, for instance in industrial organization, is to check whether the unsuitable equilibria can be discarded on intuitive grounds. This informal procedure has sometimes led to new developments in game theory. However, AN contributors can regiment it less technically *by letting the narrative speak*. The pieces of narrative information to use for the selection task may overlap, but should not be identical with, those which have already served to determine the *explanandum*; otherwise, a gross circularity would result.

[27]Zagare and Kilgour (2006) and Zagare (2011, pp. 161–162) show awareness of this problem. Indeed, it would have been extraordinary if in July 1914 France had threatened Britain that it might align with Germany.

[28]See Bates et al. (1998, p. 15): "Repeated games, for example, can yield a multiplicity of equilibria. To explain why an outcome occurred rather than another, the theorist must ground his or her explanation in empirical materials." It is for "the narrative" to provide these "materials."

The last move illustrates how the recourse to the narrative may complement an imperfect deductive explanation. The selection it operates is conceptually different from that which consists in fixing parameter values. However, writers of AN are not always clear about which kind of multiplicity, and hence which kind of selection, they are concerned with. The reason for this seems to be that extensive form games of incomplete information entered the field only belatedly. Under complete and otherwise perfect information, backward induction provides the extensive form game with an essentially unique equilibrium once the parameters are fixed. Under incomplete, hence also imperfect information, backward induction is no longer available, and subjective beliefs are part of the definition of equilibria, which tends to make them non-unique even for fixed parameter values. The *Analytic Narratives* contributors were not yet in a position to clarify this necessary distinction, which, by contrast, comes out well in the introductory comments to *The Games of July*.

Let us now return to the problem of preference assumptions. Although it seems a good strategy to combine "revelation" (from choices) and "position" (of commonsensical comparisons), this will not always be sufficient to determine the players' preferences, and here again the narrative can help. For one thing, by granting that the historical actors have some internal stability, it enlarges the set of choice data on which "revelation" depends; for another, again granting stability, it offers a means of cross-checking what "position" suggests. To illustrate, Napoleon's preferences in June 1815 cannot be guessed only from his choices at the time plus the trivial notion that he preferred victory to defeat. His preferences included his risk attitudes, and to assess the latter, it is best to adopt some temporal distance and remember that he had been a bold and generally lucky gambler throughout his career. Thus, enlarging the narrative beyond the initial scope limits the arbitrariness of the preference assumptions in AN. This illustrates how the narrative can facilitate compliance with the deductive scheme – this time, when an epistemic, not a logical requirement, is concerned.

Still in the epistemic sphere, consider the problem of independently testing a generic game. One way AN contributors address this problem is by applying the game to spatiotemporally disconnected historical states of affaires, so as to explain them jointly. (This actually exceeds the usual demand for an independent test, which demands that one resort to control cases, but not to the point of devising a full-fledged explanation for these cases.) Besides Zagare's repeated employment of the Triple Crisis Game, we can exemplify the procedure with Nalepa's generic game, which she uses to account for three different democratic transitions (Case 11). Interestingly, Nalepa (2010) reinforces her joint explanation by selecting new facts from the narrative. She mentions that the communists began negotiating with the democrats at an early stage in Poland and Hungary, and belatedly in Czechoslovakia. This suggests that the communists believed in their bargaining power more strongly in Poland and Hungary, a suggestion that connects with another fact of the matter: they had infiltrated democrats more deeply in these two countries. Thus, the temporal pattern of negotiations indirectly supports the main explanatory point, which is that the degree of infiltration was crucial to the communists' success and failure in defending their position. Revisiting the narrative with an eye on independent testing has turned out to be productive.

Although it points in a clear direction, the previous analysis is not sufficiently specific, because it does not make clear why it is *narration*, rather than any other way of presenting historical evidence, that helps fill explanatory gaps. This suggests a more general question: why do AN contributors so strongly value this particular mode of exposition? The primary reason seems to be that they are concerned with *interactions*, whether directly or indirectly, and historical reports of actions typically come under narrative guise.[29] Now, it is still another question whether AN should *themselves* preserve the form of their existing sources. Arguably, by doing so, they are more objective than they would be if they reshaped these sources in non-narrative form. Presumably, reshaping would sometimes add and sometimes suppress too much information. Mongin (2008, 2018) illustrates this point with the sources on the Waterloo campaign. They all consist of narratives, from the witnesses' unelaborated testimonies to the military strategists' highbrow accounts, with a number of contextual variations in between, such as those of popular military history. To summarize this evidence in any other way than narration would distort it. Moreover, the problem addressed in the Waterloo study precisely consists of a gap found in earlier narratives, and a natural way to make this clear is to devise a summary narrative in which the gap becomes shockingly visible.

Related to these points is the fact that the formal theories of rational choice have privileged connections with the narrative mode of exposition. Models formulated with these theories typically come with *stories* to make their technical points salient or even simply intelligible. In the case of AN, these stories may or may not be realistic to the point of mapping onto parts of a historical narrative. In section "Some Defining Characteristics of Analytic Narratives," we argued that the actions in extensive form games were idealizations of concrete actions. Although complete similarity is beyond reach, the stories behind rational choice models, and in particular games, have sufficient common ground with genuine histories for relevant interchanges to take place between the two.[30]

There are many other reasons why it may be justified to value the narrative mode of expression, but they are not specific to the methodology of AN and belong rather to that of history in general, so we will not review them here.[31] However, one of these reasons deserves to be singled out. Philosophers of history often argue that narratives, properly understood, encapsulate causal claims in their reports of temporal succession. Here is a famous didactic example: "The king died, and then the queen died of grief." Whether causal claims such as that made by this sentence are satisfying from an explanatory viewpoint is a matter of dispute. Some philosophers, like Danto (1985), think that narratives are explanatory by themselves, while others, like Dray (1971), think that narratives are only occasionally so. An intermediary

[29]The novelist Philip Roth is said to have made this pronouncement: "Everything that matters comes to us in the form of a narrative." At least, every *action* that matters comes to us in this form.

[30]More on this in Grenier et al. (2001) and Mongin (2008).

[31]See in particular Roberts's (2001) collection, with classic pieces by Dray, Mink, White, and others and the 1985 collection of Danto's works in the philosophy of history, *Narration and Knowledge*.

position, which is probably White's (1984, 1987), is that the causal content of a narrative can always be extracted and subjected to separate scrutiny, so that the narrative will or will not be explanatory, depending on how the examination of the content turns out. This intermediary position seems promising for the methodology of AN. By stressing the possibility of extraction, it opens the door to the modeling stage of AN, and by making this extraction relative to causality, it reorients their assessment from their deductive towards their causal performance, an enrichment of the current discussion of AN.

We now complete our examination of the narrative component of AN by paying special attention to their *expository feature*s; for the main, this is borrowed from Mongin (2016). We will recognize three distinctive expository patterns for AN.

Case 1, concerning Genoa, follows a chronological order extending from the consulate period, with its succession of civil peace and war under the consulate, to prolonged civil peace under the *podesteria*.[32] The exposition of the consulate period follows a remarkable pattern. First, a standard narrative records the main facts and introduces the *explananda*; then comes a game-theoretic model with relevant variations, which suggests the *explanans* hypotheses; and finally a narrative consolidates the explanation. Unlike the first, this narrative borrows theoretical terms from the modeling part, e.g., "mutual deterrence equilibrium," and serves to clarify and empirically support the *explanans* hypotheses, thus assuming the function of problem-solver. Despite its special features, this is a narrative all right, so we do have an *alternation* pattern. This pattern also appears, though a little less transparently, in the rest of Case 1, as well as (albeit with some differences) in Cases 4 and 5.

In Case 6, on Waterloo, the exposition begins with a campaign narrative in the style of military history, which introduces the main facts and the (here unique) *explanandum*. Then, a game-theoretic model delivers the explanatory hypothesis, and a discussion follows that introduces more historical evidence. As a distinctive feature, this study considers the initial narrative as being essentially satisfactory, except for the explanatory gap it draws attention to. Thus, the model and its discussion are parenthetical, and the initial narrative can be resumed once the gap is filled. This pattern of *local supplementation* differs from alternation in being less ambitious, since it does not involve creating a new narrative. However, the two patterns locate the final explanation in a narrative, and this feature is more important than the difference between them.

Case 11, on "transitional justice," goes through the following expository steps. It introduces the historical problem of "transitional justice" in the early post-communist years, puts forward a theoretical hypothesis both informally and formally, proceeds to a narrative history of transitions, and finally compares facts from

[32]One may note the dramatic quality of this sequence, which reminds one of the triadic plot structure in many dramas or fictional stories: an initially stable situation, a conflict between the characters, and a positive or negative resolution of this conflict (see Freytag 1863, elaborating on Aristotle's *Poetics*).

this narrative with the theoretical hypothesis. Although well developed, the narrative here is only a provider of data, and the whole study obeys a standard hypothesis-testing scheme. What customizes it is that it gives a narrative form to its empirical evidence. We will call this expository pattern *analyzed narrative*. Unlike in the first two patterns, it does not entrust the final explanation to a narrative, whether ordinary or revised, but rather states it abstractly and theoretically.[33]

Starting from this contrast, one may conceive of AN in two different ways. In a *restrictive* view, they count as such only if they follow the alternation or local supplementation patterns; in a *liberal* view, they may also follow the analyzed narrative pattern. A reason for preferring the former is that it seems best to emphasize what is most specific about AN; we have already used such an argument in section "Some Defining Characteristics of Analytic Narratives." The thought-provoking move is to *make* narratives analytic, and this necessitates the return of the narrative at the end of the study. As they simply *juxtapose* the analytic and narrative components, analyzed narratives are less novel. For two reasons, however, this argument may be too stringent.

First, as we have mentioned, the three patterns share the feature of bringing in historical evidence narratively, and this is by itself an important specification, since not every work in economic or political history does that. There have even been voices in these fields, as well as in history more generally, calling for narration to be downgraded, an attitude that conflicts with the way it enters analyzed narratives. Well-known representatives of this anti-narrative stance are the members of the *Annales* school, who championed "problem-oriented" against "narrative-oriented" history, and the intransigent "new economic historians" whose flags were economic modeling and econometric techniques.[34] One reason for preferring the liberal view of AN is that they clearly illustrate the opposite stance of the "revival of narrative," to borrow a famous phrase by Stone (1979). Second, the beginning of this section has pointed out several means by which narration can rescue flimsy explanations, and these means are also available in the third pattern. In particular, Case 11 selects equilibria from narrative information in a manner no different from Cases 9 and 10. In terms of the principles stated in the introduction, the third one, whereby the narrative actively contributes to historical explanation, appears to be common to AN

[33]Crettez and Deloche's (2018) treatment of Cesar's death further illustrates the subgenre of analyzed narratives. Following the general AN methodology, they carefully review the historical evidence and extract from it a problem they solve with the aid of a formal model. How plausible is the suggestion made by Suetonius and others that Cesar was aware of the plot to murder him when he went to the Ides of March meeting of the Senate? The authors' two-person game of normal form has a single Nash equilibrium that is mixed, which in their view suggests a negative answer to this question. Here the narrative provides both the evidence and the problem, but the solution is stated in theoretical, non-narrative terms.

[34]The firm contrast that *Annales* postulates between narrative- and problem-oriented history appears among others in Furet (1981). The anti-narrative stand is also present among new economic historians, e.g., Kousser (1984), who defends "quantitative social scientific history" against a "revivalism" of narrative. Not every cliometrician has adopted this stance; witness the open attitude of the editors of this *Handbook*.

broadly understood, although the alternation and local supplementation patterns apply it more systematically and, as it were, more interestingly than the analyzed narrative pattern does.

These reasons can tilt the balance in favor of the liberal view of AN, and we will adopt this view here, thus completing our attempt at defining the AN genre. To make this definition more transparent, we may cite two groups of studies it does *not* cover. (i) Some studies are concerned with specific historical events, involve a significant amount of narrative information, and base their explanations on the outcome of complex interactions, but refrain from adopting a formalism and thus provide only promising sketches of explanation. Besides Case 3, Myerson's (2004) discussion of the Weimar disaster is a good example – and all the more so given that his informal comments are evidently made with a possible modeling in view.[35] Works like these are *proto-analytic narratives*. (ii) Other studies are also concerned with specific historical events, base their explanations on the outcome of complex interactions, and do develop these explanations by means of properly formalized models but do not confer an explanatory function on the narrative, nor even prioritize it among the sources of historical information. Two studies by Greif that antedate his adherence to AN methodology can serve as examples. Greif (1993) investigates the community of Maghribi Jewish traders who operated in maritime commerce in the eleventh and twelfth centuries, and Greif et al. (1994) investigate the connection between the merchant guilds of medieval Europe and long-distance trade. These studies focus on the commitment and coordination problems that traders faced in their dealings with official rulers or other traders, and they use game-theoretic models to show that well-designed informal (in the Maghribi example) or formal (in the guild example) institutions could overcome these problems. Their exposition mixes theoretical elements with historical evidence, which is only occasionally narrative, in a dialogue that clearly differs from the alternation pattern implemented in the Genoa case. Despite their emphasis on interactions and game theory, which likens them to AN, they are closer to other formalized works in historical political science or economics. Let us designate them *analytic non-narrative histories*.[36]

Conclusion

This chapter has defined AN in terms of three principles, the most intriguing of which is that AN call upon a narrative also at the explanatory stage. We have pursued our definitional investigation at the same time as making progress with the other topic of the chapter, i.e., how AN contribute to historical explanation. In this latter

[35]For instance, Myerson (2004) suggests treating the events of 1930–1933 in terms of a signaling game between the Allies and the German conservative leaders. To get rid of the reparations burden, the latter would try to impress the former by pushing forward Nazism as a political force (a dangerous game if ever there was one).

[36]More examples could be found in Greif's (2002) survey of game-theoretic economic history.

discussion, we have selected the deductive scheme of scientific explanation as a benchmark: overall, analytic narratives exhibit more deviation from, than conformity to, the deductive scheme, and this is precisely why they call upon the narrative for help. A more complete account of their explanatory performance would have clarified the kind of causal connections they can hope to establish, and this would have led us also to investigate the kind of counterfactual history they develop. The necessary brevity and thematic unity of this chapter made it impractical to go in these directions. Similarly, we refrained from explicitly defining what a narrative consists of. This would have required us to compare the narrative mode of discourse with the other modes, such as exposition, argumentation, and description, which historians also use, and thus to delve in the recent work of narratologists as well as the more traditional concerns of rhetoricians and literature teachers.

Thus far, political scientists have paid more attention to AN than other social scientists. This is easily explained by the fact that the two main currents that have shaped the development of AN, i.e., the equilibrium approach to institutions and the deterrence approach to national security, are primarily of concern in political science. But these disciplinary associations are in part a matter of contingency, and it is anyhow the case that "analytic narratives should have no boundaries with respect to subject or evidence" (Bates et al. 2000b, p. 690). In particular, there is no reason why AN could not also have a significant place in economic history. What might restrict their use therein is that they are concerned with fine patterns of actions and events, like the formal theories of rational choice they borrow from, and are thus unable to handle long-term historical processes, such as Britain's Industrial Revolution, or large-scale sets of social and economic relations, such as slavery in nineteenth century USA. But these wide topics are of course the bread and butter of today's economic historians; and if AN can teach them anything, it would be precisely by directing their attention towards the fact that it is possible to approach some microscopic structures no less rigorously than these topics, albeit by different formal means.

Cross-References

- ▶ Cliometric Approaches to War
- ▶ History of Cliometrics
- ▶ Institutions
- ▶ Political Economy
- ▶ The Antebellum US Economy

References

Allison GT (1971) Essence of decision, explaining the Cuban Missile Crisis. Little, Brown and Co, Boston (2nd revised ed. co-authored with P. Zelikow, Addison Wesley Longman, New York, 1999)

Bates RH, Greif A, Levi M, Rosenthal JL, Weingast B (1998) Analytic narratives. Princeton University Press, Princeton
Bates RH, Greif A, Levi M, Rosenthal JL, Weingast B (2000a) The analytic narrative project. Am Polit Sci Rev 94:696–702
Bates RH, Greif A, Levi M, Rosenthal JL, Weingast B (2000b) Analytic narratives revisited. Soc Sci Hist 24:685–696
Betts R (1997) Should strategic studies survive? World Polit 50:7–33
Bird A (1998) Philosophy of science. UCL Press (reprinted by Routledge, London, 2000)
Brams SJ (1975) Game theory and politics. Free Press, New York
Brams SJ (1985) Superpower games: applying game theory to superpower conflict. Yale University Press, New Haven
Brams SJ, Kilgour M (1988) Game theory and national security. Oxford University Press, New York
Braudel F (1969) Ecrits sur l'histoire. Flammarion, Paris (Engl. trans. On history. University of Chicago Press, Chicago, 1980)
Clark S (1985) The Annales historians, ch. 10. In: Skinner Q (ed) The return of grand theory in the human sciences. Cambridge University Press, Cambridge, pp 178–198
Clark G (2007) A review of Avner Greif's institutions and the path to the modern economy: lessons from medieval trade. J Econ Lit 14:727–743
Crettez B, Deloche R (2018). An analytic narrative of Caesar's death: suicide or not? That Is the Question. Ration Soc 30:332–349
Danto A (1985) Narration and knowledge. Columbia University Press, New York
Downing BM (2000) Economic analysis in historical perspective. Hist Theory 39:88–97
Dray W (1971) On the nature and role of narrative in historiography. Hist Theory 10:153–171
Erickson P (2015) The world the game theorists made. University of Chicago Press, Chicago
Freytag G (1863) Die Technik des Dramas. S. Hirzel, Leipzig
Furet F (1981) De l'histoire-récit à l'histoire-problème. In: L'atelier de 1'histoire. Flammarion, Paris, pp 73–90. Eng. tr. in G. Roberts (2001) (ed), ch. 17, pp 269–280
Fudenberg D, Tirole J (1991) Game theory. The MIT Press, Cambridge, Mass
Greif A (1993) Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. Am Econ Rev 83:525–548
Greif A (2002) Economic history and game theory, ch. 52. In: Aumann RJ, Hart S (eds) Handbook of game theory with economic applications, vol 3. Elsevier, Amsterdam, pp 1989–2024
Greif A (2006) Institutions and the path to modern economy. Cambridge University Press, New York
Greif A, Milgrom P, Weingast BR (1994) Coordination, commitment, and enforcement: the case of the merchant guild. J Polit Econ 102:745–776
Grenier JY, Grignon C, Menger PM (eds) (2001) Le modèle et le récit. Editions de la Maison des sciences de l'homme, Paris
Harrington J (2009) Game, strategies, and decision making. Worth Publishers, New York (2nd ed., 2014)
Haywood OG Jr (1950) Military decision and the mathematical theory of games. Air Univ Q Rev 4:17–30
Haywood OG Jr (1954) Military decision and game theory. J Oper Res Soc Am 2:365–385
Hempel C (1965) Aspects of scientific explanation. Academic, New York
Kousser JM (1984) The revivalism of narrative: a response to recent criticisms of quantitative history. Soc Sci Hist 8:133–149
Levi M (1997) Consent, dissent, and patriotism. Cambridge University Press, Cambridge
Levi M (2002) Modeling complex historical processes with analytic narratives. In: Mayntz R (ed) Akteure-Mechanismen-Modelle. Schriften aus dem Max-Planck-Institute für Gesellschatsforschung Köln, vol 42. Campus, Frankfurt am Main, pp 108–127
Mongin P (2008) Retour à Waterloo. Histoire militaire et théorie des jeux. Ann Hist Sci Soc 63:39–69
Mongin P (2009) Waterloo et les regards croisés de l'interprétation. In: Berthoz A (ed) La pluralité interprétative. Odile Jacob, Paris

Mongin P (2016) What are analytic narratives? In: Miller B, Lieto A, Ronfard R, Ware SG, Finlayson MA (eds) Proceedings of the 7th workshop on computational models of narrative (CMN 2016), open access series in informatics. http://drops.dagstuhl.de/opus/volltexte/2016/6714/pdf/OASIcs-CMN-2016-13.pdf
Mongin P (2018) A game-theoretic analysis of the Waterloo campaign and some comments on the analytic narrative project. Cliometrica 12:451–480
Morrow JD (1994) Game theory for political scientists. Princeton University Press, Princeton
Myerson RB (1991) Game theory: analysis of conflict. Harvard University Press, Cambridge, MA
Myerson RB (2004) Political economics and the Weimar disaster. J Inst Theor Econ 160:187–209
Nagel E (1961) The structure of science: problems in the logic of scientific explanation. Routledge, London
Nalepa M (2010) Captured commitments. An analytic narrative of transitions with transitional justice. World Polit 62:341–380
North DC (1981) Structure and change in economic history. Norton, New York
North DC (1990) Institutions, institutional change, and economic performance. Cambridge University Press, Cambridge
Osborne MJ, Rubinstein A (1994) A course in game theory. MIT Press, Cambridge, MA
Ravid I (1990) Military decision, game theory and intelligence: an anecdote. Oper Res 38:260–264
Riker WH (1982) Liberalism against populism: a confrontation between the theory of democracy and the theory of social choice. W.H. Freeman, San Francisco
Roberts G (ed) (2001) The history and narrative reader. Routledge, London
Salmon W (1992) Scientific explanation, ch. 1. In: Salmon MH, Earman J, Glymour C, Lennox JG, Machamer P, McGuire JE, Norton JD, Salmon WC, Schaffner KF (eds) Introduction to philosophy of science. Prentice Hall, Englewood Cliffs, pp 7–41
Schelling TC (1960) The strategy of conflict. Harvard University Press, Cambridge, MA
Schiemann JW (2007) Bizarre beliefs and rational choices: a behavioral approach to analytic narratives. J Polit 69:511–524
Stone L (1979) The revival of narrative: reflections on a new old history. Past Present 85:3–24
von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton (2nd ed., 1947)
Wagner H (1989) Uncertainty, rational learning, and bargaining in the Cuban missile crisis. In: Ordeshook P (ed) Models of strategic choice in politics. University of Michigan Press, Ann Arbor, pp 177–205
White H (1984) The question of the narrative in contemporary historical theory. Hist Theory 23:1–33
White H (1987) The content of the form: narrative discourse and historical representation. Johns Hopkins University Press, Baltimore
Zagare FC (2009) Explaining the 1914 War in Europe. An analytic narrative. J Theor Polit 21:63–95
Zagare FC (2011) The games of July: explaining the great war. University of Michigan Press, Ann Arbor
Zagare FC (2014) A game-theoretic history of the Cuban Missile Crisis. Economies 2:20–44
Zagare FC (2015) The Moroccan crisis of 1905–1906: an analytic narrative. Peace Econ Peace Sci Public Policy 21:327–350
Zagare FC, Kilgour M (2000) Perfect deterrence. Cambridge University Press, Cambridge
Zagare FC, Kilgour M (2003) Alignment patterns, crisis bargaining, and extended deterrence: a game-theoretic analysis. Int Stud Q 47:587–615
Zagare FC, Kilgour M (2006) The deterrence-versus-restraint dilemma in extended deterrence: explaining British policy in 1914. Int Stud Rev 8:623–641

Spatial Modeling

Florian Ploeckl

Contents

F. Ploeckl (✉)
School of Economics, The University of Adelaide, Adelaide, SA, Australia
e-mail: florian.ploeckl@adelaide.edu.au

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_56

Abstract

Spatial modeling is a systematic approach to understand the spatial configuration of economic activity from a local to a global scale. This chapter begins with an overview of empirical tests of spatial correlation, including testing of spatial randomness, join-count statistics, and Moran's I. This leads to a discussion of spatial concentration measures and the usage of such tests and measures in the context of economic history. An overview of the development of spatial modeling, from von Thünen to Hotelling to the new economic geography demonstrates the abstract nature of modeling, its application in a wide range of settings, and the usefulness of economic history to validate theoretical approaches. The final section goes over a range of modeling components used in quantitative spatial economics and their relevance in economic history. The chapter concludes with an outlook to alternative modeling approaches, including spatial point processes and networks.

Keywords

Economic Geography · Agglomeration · Spatial Concentration · Spatial Correlation

Introduction

"Everything is related to everything else, but near things are more related than distant things." This quote from Tobler (1970) is commonly referred to as the first law of geography. Even if most empirical analyses in economics take this reciprocal influence to be completely negligible for all practical purposes, the question of spatial relationships has received much more attention again in economics and the economic history literature over the last three decades, which has triggered a renaissance in explicit spatial modeling.

While economists had not paid much attention to geographical aspects, the field of geography was alive and well. Nevertheless, when economists started to pay more attention, the approaches, questions, methods, and tools used in both fields continued to diverge. Economists focused on highly technical, theoretical, and quantitative tools to draw general conclusions. Geographers, on the other hand, moved to qualitative studies focusing on specific cases, which were strongly shaped by the field's "cultural turn"(Krugman 2010). While there is a literature that brings the approach of geographers to historical analysis, this chapter focuses on spatial modeling as understood by economists and quantitative economic historians. "The Development of Spatial Modeling" presents a history of formal spatial modeling from this viewpoint over the last two centuries and discusses its relationship and interaction with economic history.

This overview of the historical development of spatial modeling demonstrates that it has an extreme breadth without a clear theoretical center. The core theories of international trade, Ricardo, Heckscher-Ohlin, and the new trade theory with

increasing returns and monopolistic competition, are important components, yet they are only pieces of the puzzle. Preferences and amenities, production functions and local productivity, and the flows of people and ideas are just a few of the additional pieces. Consequently, this chapter does not present formal theories but utilizes a recent overview of the various components (puzzle pieces) used in location theoretic modeling in economics to place their relevance within the application of spatial modeling in economic history.

To start off, however, a more basic question is discussed, namely, how to detect spatial correlation and therefore a potential need for spatial modeling? After a short discussion and clarification of the relevant structures, in particular actors and space, the main empirical approaches to detect spatial correlation are introduced. This covers the distribution of events over space, the spatial distribution of binary characteristics of a set of locations, and similarly, the distribution of continuous characters. It concludes with a discussion of other statistical measures of spatial concentration and specialization and their use within economic analysis in economic history.

Basic Structures

Monads and Dyads

Probably the most famous model explaining an explicitly spatial economic phenomenon is the *gravity equation*. In practice, it resembles the physics theory modeling the force of attraction between spatially distinct objects. Applied to trade relationships, the equation explains the trade flows between two countries by modeling it based on the characteristics of origin and destination and the distance between them. The focus on flows between two countries implies that the modeling uses dyads, the combination of two actors, as the unit of observation. There are a number of other spatial relationships, such as trade agreements to military conflicts, where dyads are an appropriate modeling approach. For quantitative economic history, another important application is the analysis of *market integration*, focusing on the correlation of prices between two locations.

Lampe and Sharp (2016) thoroughly discuss foundations and the use of the gravity equation in economic history. Market integration has been discussed by a number of authors, including Jacks (2005). Consequently, this chapter focuses on individual actors, modeling the origin and impact of spatial relationships with monads, i.e., individual actors, rather than pairs of them.

Actors

The advent of modern data processing and computing also led to the development of *geographic information system* (GIS) software. Its purpose is to collect, combine, manage, and transform spatially referenced data. GIS data are categorized into two

different types, namely, *vector data* and *raster data*. Although each individual data series has to be in one of the two formats, it is possible to combine data from both categories as well as convert series from one to the other. These data formats reflect the nature of the underlying data and as such correspond to the units of observations representing actors within modeling approaches.

Vector data are geographic shapes mirroring empirical units of observations and come in three different types, namely, points, lines, and polygons. *Points* are usually used to locate individuals or individual units like firms or cities in space. *Lines* model links between points, for example, infrastructure, borders, or communication ties. *Polygons* represent areas, most prominently countries and regions.

Raster data use a different approach by parting the complete surface into a rectangular grid where each cell contains a single data point. Although each cell is associated with a single value, that value can actually imply a combination of others. One example is that each cell has a color value; however each color is a unique combination of the three base colors. The format resembles digital pictures, and important applications include mapping *nighttime light intensity* or geographic characteristics like elevation.

In terms of spatial modeling, it is straightforward to see vector data as representing units of observation and therefore actors in most modeling approaches. Raster data, in contrast, does not fit particularly well as representations of actors, whose actions are modeled. In this context it serves mainly as a data source that allows the determination and calculation of characteristics of vector-based units observations. As such it can be used as outcome data, for example, satellite data about nighttime light intensity is used to demonstrate modern consequences of historical events or long-term developments involving a number of spatial phenomena like borders (Michalopoulos and Papaioannou 2018).

Distance

Spatial modeling and empirical studies using spatial approaches, however, not only use *geographic distance* based on a two-dimensional representation of physical space but can accommodate other approaches to space and distance as well.

As shown in the discussion of Hotelling's model in "The Development of Spatial Modeling," one possible difference is the number of dimensions. Simple models might use one-dimensional distances, which can usually be represented by a single line and the distance between actors located on that line, while more complicated ones use multidimensional setups, where the relative locations of actors differ along multiple axes. Modeling policy differences between countries is an example, where the dimensionality of space depends on the number of policy areas under comparison, ranging from a single area to multiple policy fields.

Another possibility is to move away from physical distance and approach the relative spatial position of actors in different ways. The main approach is the idea of binary distance, where two actors are either directly linked to each other or not at all. In practical empirical terms, that type of relative positioning is for actors

representing areas, for example, countries or counties, which often exhibit some form of contiguity, for example, two countries sharing a common border. A related approach is thinking in terms of network structures. *Networks* are modeled as nodes and edges between these nodes, so if the node represents an actor, then the existence of an edge with another node implies that the two actors are connected, while the absence of an edge means that they are not.

Modeling Spatial Correlation

Although Tobler's law stated above is an assumption that "everything is related to everything else," it's not clear how strong those connections are and whether the correlation really matters or can even be empirically detected.

Any empirical exploration of *spatial autocorrelation* obviously requires a method to determine a summary statistic characterizing the presence, direction, or strength of any such correlation. The different types and characteristics of actors, different approaches to distance, and different statistical assumptions imply that there is no single way to characterize, determine, and test for spatial correlation. There is, however, a substantial literature in regional science and other geography disciplines that provides empirical approaches to explore spatial correlations that are adopted and used in quantitative economic history.

The following goes through commonly used measures in three different settings, namely, spatial correlation between the locations of actors, spatial correlation between binary characteristics of actors, and spatial correlation between continuous characteristics of actors. The application is practically demonstrated with a data set of villages and towns in Saxony and their population in the early nineteenth century. The data are based on Ploeckl (2012, 2017) and contains 3515 settlements, including 140 legal towns, with their respective population in 1834.

Spatial Randomness

The first considers cases where the interest is whether observations are located randomly on a two-dimensional surface. As Baddeley et al. (2015) describe, if the spatial distribution is indeed random, the underlying *spatial point process* is characterized by two key properties:

- *Homogeneity*: the points have no preference for any spatial location.
- *Independence*: information about the outcome in one region of space has no influence on the outcome in other regions.

Mathematically this can be formulated for a point process X and an intensity value λ as:

- The number $n(X \cap B)$ of random points falling in a test region B has mean value $En(X \cap B) = \lambda|B|$.

- For test regions $B1$, $B2$, .., Bn which do not overlap, the counts $n(X \cap B1)$, .., $n(X \cap Bn)$ are independent random variables.

In combination with the assumption that the counts within the test regions follow a Poisson process, these characteristics are referred to as *complete spatial randomness* (CSR). CSR implies that the expected number of observations within any area is linear in the size of the area. The constant coefficient, λ, is referred to as *intensity* (Baddeley et al. 2015).

The context of a point process and its intensity is relevant for a number of situations, probably more prominent in fields like biology (e.g., the abundance of virus particles on a cell surface), zoology (the location of ticks in a forest), forestry (spatial distribution of particular tree species), and climatology (lightning strikes). But it is also useful for economic history, from the productivity of agricultural crop production to the risk and occurrence of particular behavior, such as crime, to the presence of natural resources and other relevant endowments.

If the spatial distribution is not random and CSR does not hold, then there are two alternatives. First, the pattern can be more dispersed, which makes it appear more regularly spaced, and second, it can be grouped, which makes it appear more clustered.

As Baddeley et al. (2015) mention, there are a number of statistical approaches testing spatial randomness. A basic one is using *quadrat counts*, where the whole area is split into equal-sized quadrats. The number of points in each is then used in the test statistic. This also works with other shapes as long as they are of equal size and nonoverlapping. With m regions and $n = \sum_i n_i$ the total number of points, the formal test statistic is: $X^2 = \sum_i \frac{\left(n_j - n/m\right)^2}{n/m}$, which has under the null hypothesis a distribution that is approximately $\chi 2$ with $m - 1$ degrees of freedom.

The location of settlements within a region is an applicable situation where a test of spatial randomness is informative. Applying the quadrat counting test to the historical distribution of settlements in Saxony shows for 20 tiles a $\chi 2$ value of 407 and a p-value below 0.01. Clearly settlements are not randomly distributed. Specifying the alternative as one-sided and either regular or clustered shows that these spatial locations look as if they are clustered.

The exploratory tests about spatial randomness can only diagnose in which direction any violation runs but don't specify which of the two main assumptions, the featureless surface and the independence between observations, is the source of the deviations. The final section revisits approaches analyzing these types of statistical processes and structural factors shaping the spatial distribution of observations.

Join-Count Statistics

Most economic history analyses concern situations where the spatial location of actors, from individuals to firms, towns, counties, or countries, is taken as given. The interest is then in the spatial correlation of a particular actor characteristic, so in its distribution over the existing locations of the observations.

If the outcome of interest is binary, one approach is the *join-count statistic*. Examples for such outcomes are which firms went bankrupt and which survived, which cities adopted a particular policy or technology, or which countries were on the gold standard. The statistic assumes a static situation, so this approach is appropriate for empirical situations where the relevant binary data are given by a cross section.

Following the theoretical exposition by Cliff and Ord (1981), this can be illustrated for the case of countries on and off the gold standard. The central idea is to look at contiguous country pairs and count the connections (hence the "join") categorized by the status of the two states. Practically, the basis are separate counts of the number of neighboring country pairs where both are on gold (GG), where one is on gold and the other one is off (GO), and where both are off gold (OO). Based on the probability that a country is on or off gold, an expected number of joins for each of the counts can be derived. If the observed GG count and the corresponding expectation are close, the gold standard is randomly distributed; if the count is significantly higher, gold standard countries are clustered; and if it is negative, they are dispersed. The results for the three counts are not necessarily symmetric, so a spatial correlation of being on the gold standard does not imply a spatial correlation of being off gold. The difference between the expectation and the actual counts also indicates how strong such a spatial correlation effect is.

Cliff and Ord (1981) provide a generalized theoretical formulation as well as the necessary moments for significance testing. The latter requires two assumptions, first about the sampling process underlying the observations and second about the nature of the joins. The first has two possible specifications that mirror sampling with and without replacement: free and non-free sampling. *Free sampling* means each observation has the same probability of being on the gold standard, which is usually based on the sample average of countries on the gold standard. *Non-free sampling* is when each observation has again the same probability, but the total number of countries on the gold standard is restricted to that observed in the sample.

The other assumption is about *spatial weights*. Cliff and Ord (1981) show that the formulation holds for general weights, not only for the special case of binary zero – one joins as represented by contiguity. The decision which spatial weights to apply, however, is left to the researcher. The discussion about distance measures above indicates potential choices, while the nature of the data provides a certain guidance. Network models have an obvious binary link interpretation, area data like countries or counties point toward contiguity measures, and observations within regular space, such as firms, fit well with distance-based weights. The main point, however, is that for the same outcome data, different weights can result in different conclusions, since any spatial correlation is dependent on the weighting scheme.

The choice of a weighting scheme depends on the purpose of the calculation. If the intention is exploratory to determine the nature of any potential spatial correlation, then testing different weights to see whether any result exhibits spatial correlation at all is one approach. If the intention is to investigate a particular transmission mechanism or channel that should cause spatial correlation, then the weighting scheme should be specified accordingly.

The empirical example demonstrates the impact of the weighting scheme. The set of settlements in Saxony is split into towns and villages depending on whether or not a location had official, legal town rights in 1834. This results in 140 settlements designated as towns. The join-count test is conducted twice, once with binary weights based on the distance between the two settlements as less than 10 km and once based on the inverse distance as weight. The vicinity-based weights result in a statistic of 3.55 versus an expectation of 2.77 with a p-value below 0.01, indicating that there was positive, spatial clustering between towns. The distance-based weights result in a statistic of 225.96 with an expectation value of 235.94 and a p-value of 0.27. This indicates no significant evidence for that spatial correlation in this form.

There is also an extension of the join-count approach to discrete data with more than two outcomes referred to as J_{tot} (Cliff and Ord 1981).

Moran I

The counterpart to the join-count statistic for continuous outcome data is the so-called *Moran's I* (Cliff and Ord 1981). This test statistic is defined as follows where n is the number of observations and w_{ij} is the weight of the link between i and j:

$$I = \frac{n}{\sum_{ij} w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

As with the join-count statistic, the included weights w_{ij} have to be specified exogenously. The related discussion above about the specification and impact of the weight matrix applies here as well, so its precise form depends on the purpose of the calculation and should be informed by the research question and any implicit or explicit theoretical framework about the phenomenon in question.

The value of Moran's I ranges between -1 and 1. If there is no spatial autocorrelation present, it takes on the value of zero, while positive values indicate positive spatial autocorrelation (*clustering*), and negative values indicate negative spatial autocorrelation (*dispersion*).

As for the join-count statistic, it is possible to derive expectation and variance for Moran's I, which allows standard approaches to test the statistical significance of any empirical results.

Usage

The main usage of these indicators is to explore the possible presence and direction of *spatial autocorrelation* within a data set. Such an exploration can also be used to determine whether the spatial implications of a particular transmission mechanism are reflected in the observed data.

A further area of exploration is to understand changes in spatial correlation over time, between different outcomes or between different configurations. In the cases of join-count statistics and Moran's I, it is possible to trace and compare developments over time as long as the utilized distance weights remain constant. Because the development over time is only a special case of a comparison between different data sets, the same logic applies to such a situation; as long as the spatial weights are identical, it is possible to draw comparative conclusions about the relative strength and direction of spatial autocorrelation of the different data sets. The situation changes when spatial weights differ. The comparisons of the strength are only useful for identical data to determine whether a particular mechanism results in spatial correlation.

The three indicators only test whether spatial autocorrelation is present in the full sample. They do not identify characteristics of particular subsets, so, for example, they do not identify whether the same correlation holds for subregions or where particular clusters are located. There are approaches to explore local variations over space, in particular *local indicators for spatial autocorrelation* (LISA) (Anselin 1995).

Measures of Specialization

The described measures focus on the spatial aspects of the geographic distribution of some economic outcome. There are, however, also approaches that focus primarily on the strength of *spatial concentration* and have similarities with those looking at *market concentration*, like the Herfindahl index.

Krugman (1991) formulates a measure of *regional specialization*:

$$SI_{jk} = \sum_i \left| \frac{E_{ij}}{E_j} - \frac{E_{ik}}{E_k} \right|$$

where E_{ij} is the level of employment in industry i for region j, E_j is the total industrial employment for region j, and E_{ik} and Ej are the same for region k. The index compares regions j and k, which are completely despecialized if the index is zero and fully specialized if it is 2. Kim (1995) uses this to document the regional specialization of manufacturing in the USA between 1860 and 1987 and finds the average specialization increases until the interwar period, before it consistently decreases throughout the postwar period.

Kim also analyzes *localization*, the question of whether individual industries are more or less concentrated than manufacturing in general. He uses Hoover's *coefficient of localization*, which is based on a region's employment share relative to total employment in individual industries as well as total manufacturing employment shares, to show that the average localization level in the USA followed a similar trend over time as regional specialization.

Ellison and Glaeser (1997) take up this question of localization and incorporate firm and plant-level concentration into the calculation. Similarly, Duranton and

Overman (2005) propose an approach that shifts the analysis from regional units to continuous distance underlying the definition of agglomeration patterns. Another step is to modify the approach to determine *co-agglomeration* patterns between different industries (Ellison et al. 2010).

The calculation of these indices has a predominantly descriptive approach. It demonstrates the nature and level of spatial *agglomeration* in specific industries or economic activity in general. It is not only applicable to modern data but also allows a consistent description of historical and long-term developments.

The indices can also be used to infer either evidence about spatial models or particular economic factors driving location and specialization patterns. Kim (1995) shows that his results provide empirical support for models of *regional specialization* based on scale economies and resource use, Ellison and Glaeser (1997) show support for spillovers and natural advantage leading to agglomeration, and Ellison et al. (2010) provide evidence for the impact of three types of *Marshallian externalities* on industry agglomeration pattern, namely, input-output linkages, labor market pooling, and knowledge spillovers. In comparison to these conceptually oriented analyses, studies can also use such measures to understand historical developments. Gutberlet (2014), for example, provides a historical study demonstrating how the transition from water to steam power influenced and contributed to the geographic concentration of manufacturing in the German empire during the late nineteenth century.

The Development of Spatial Modeling

> Imagine a very large city in the midst of a fertile plain not traversed by any navigable river. The plain's soil is of uniform quality and capable of cultivation everywhere. At a great distance from the city the plain turns into an uncultivated wilderness separating this state from the rest of the world.
> The plain does not contain any other cities besides that large city. Consequently, it therefore produces all crafts and manufacturing products for the whole country while the city is supplied with food from the surrounding plain only.
> Mines and salines, which satisfy the demand for metals and salt, are assumed to be in the vicinity of this central city, which will be referred to as'the city' as it is the only one.
> The question is this: under these conditions what kind of agriculture will develop and how will the distance to the city affect the use of land if this use is chosen with the utmost rationality?
> von Thünen (1875) as translated by Beckmann (1972) and the author

Von Thünen's opening paragraph of his "Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie" succinctly lays out the question and a set of assumptions for an economic model focusing on the spatial distribution of economic activity. Written in the early nineteenth century, his work is usually seen as the origin of formal spatial modeling, laying the groundwork for understanding the nature of spatial economic patterns in an abstract and conceptual way, yet keeping the analysis fully grounded in his agricultural and practical knowledge (Fujita 2010). His model assumes that:

- There is a single city representing the whole of the economy.
- The economy is closed and in autarky, no trade with other economies can take place.
- The plain surrounding the city is featureless, so no productivity differences for agriculture or differences in transaction and transportation costs.
- Only agricultural production is considered; all other sectors are located in the city and excluded from the analysis.
- Transport costs only depend on the distance from the city.
- Agents are rational and only maximize monetary profit.

Von Thünen first discusses the answer based on his knowledge about different agricultural knowledge but then formalizes his model using mathematical relationships.

He defines the relevant variables, measured either in weight as *Scheffel* or monetary terms as *Thaler*:

The return in grain is x, the net return is ax Thaler, the sowing costs are b Thaler, the cultivation costs are c Thaler, q is the share of costs in gross returns, p is the share of costs in money terms, and h is the local grain price.

He calculates the *Landrente*, the rent from producing a particular agricultural product depending on the associated costs and benefits, as:

$$\left(\frac{ax}{h} - \frac{b + (1-p)c + c(1-p)aqx}{h}\right) \text{ Scheffel and } p(aqx + c) \text{ Thaler.}$$

Switching from conceptual into actual calculations, von Thünen combines this rent expression with the *transport costs* to the central market. He formulates these transport costs in terms of grain consumed by the horses, which foreshadows the iceberg trade costs assumption used in the modern trade literature, where a certain amount of goods "melts away." Costs are assumed to be linear in distance, with a doubling of the distance leading to a doubling of transport costs in absolute terms. Given that locations are assumed to be identical in terms of production characteristics, the location rent falls with distance from the city.

Von Thünen consequently derives the distance at which the land rent, including transport costs, equals zero, which delineates the areas where the crop in question can be profitably grown and where such a land use would be unprofitable. Repeating the calculations for other crops and agricultural products results in a *land rent curve* for each product, which shows the profit for each crop depending on the distance from the city. At any distance the curve of one product has the highest value, so that product is the most profitable to be produced at that distance. The resulting pattern is the well-known *system of rings* around the city where the production in each ring is dedicated to a particular crop or product. The assumptions of the model are strongly simplifying, yet the model has a clear result explaining the spatial patterns of economic activity observed by von Thünen almost two centuries ago and still observable today.

Von Thünen's modeling approach with rent curves resulting in a set of rings around a central location has been used in a wide range of applications and fields, most prominently in agricultural economics, urban economics, and other fields in economics, as well as regional science. The model is further applied in a wide range of settings in other social sciences, from sociology to marketing to business.

Economic history is also influenced by von Thünen's ideas and modeling approach. Fernand Braudel (1981), for example, takes up the idea of rings in his Civilization and Capitalism, using them to analyze European development before the Industrial Revolution. The model, however, is not just taken up in qualitative analysis but is obviously used in cliometrics and quantitative economic history. Von Thünen himself does so with a quantitative, comparative analysis of agriculture in Mecklenburg and Belgium. Not only does he analyze the returns for agriculture in the two regions, but using a similar approach as Wrigley (1985), he also uses relative productivities and Mecklenburg's population density to derive how many inhabitants Belgium should be able to sustain with its agriculture. His prediction is not far off of Belgium's actual population at the time, demonstrating the viability and suitability of his approach for quantitative economic analysis.

One recent example of taking up his ideas and model in quantitative economic history is Kopsidis and Wolf (2012). They analyze the spatial pattern of agricultural productivity in Prussia to determine the role of urban demand in shaping agricultural development. As an aside, their analysis uses data from 1861 to 1865, just before Mecklenburg, and the area containing von Thünen's original estate became part of the Norddeutsche Bund and Prussia's control in 1866. They build a model based on von Thünen's logic to explain agricultural productivity as reflected by rent data from Prussian agricultural statistics. Using county-level data, they show that *market potential*, measured as distance-weighted urban population, has a substantially stronger explanatory effect than alternative hypotheses, such as institutional structures. Their model has further predictions about the spatial patterns of rents, labor intensity, farm gate prices, and crop mix with certain details clearly resembling von Thünen's ideas. Testing these hypotheses empirically confirms that Prussian agriculture was strongly shaped by urban demand, explaining thereby not only the spatial pattern around individual cities but also the spatial pattern of agricultural development on a national scale.

Although von Thünen's work was written in the first half of the nineteenth century, spatial modeling developed further only at the turn of the twentieth century and the following decades according to Fujita (2010). The next wave of models started to deal with *market area analysis*, delineating the boundaries of the market areas served by different firms. This was eventually extended to *spatial price policy*, which started to jointly determine spatial locations and prices of firms. In von Thünen's model, the farmers simply receive a certain price for their products independently of their own decisions. The new models incorporated this joint determination in a number of ways, with the model introduced by Hotelling (1929) being one of the most influential or at least well-known examples.

Hotelling's (1929) "Stability in Competition," in contrast to von Thünen, was not primarily set out as a spatial analysis but to investigate and demonstrate certain

aspects and characteristics of competition within a duopoly setting. He focused on the speed of adjustments in the market and reaching equilibrium. This was based on the discrepancy between the characteristics of influential models of competition that featured firms immediately capturing large parts of a market with small price changes and the real-world observations of substantial persistence and inertia in consumer behavior in such a situation.

His analysis starts with consumers uniformly distributed over a fixed length line and two firms randomly located at two points on the line. The firms sell identical goods to the consumers, who do not distinguish between either firms' goods except for transport costs, represented by the distance between the consumer and the respective firm. Hotelling then derives profit functions for both firms in terms of both their prices. With these functions he demonstrates that if each firm optimizes its profit by setting its price in response to the price of the other, both prices converge toward an equilibrium and with each step only a small subset of consumers shifts between the two firms.

The spatial aspect becomes more central in the second part of his analysis, where he considers not only the spatial extent of the market of each firm but also their location decisions. The analysis starts with one firm randomly located on the line and then considers where the second firm will locate in relation to the first. Hotelling argues that the second firm will choose to be as close as possible to the first while remaining on the side of firm one that constitutes the larger line segment. If firm one is in the middle of the line, so in the center of the complete market, firm two is indifferent between locating on either side. The reason for that location choice is that in this way firm two can capture as much as possible of the whole market by offering lower transportation costs for that segment than firm one.

Hotelling further addresses the location question in the contexts of *central planner* and *market outcomes*, or as he labels it: "capitalism v. socialism." He argues that to minimize the social costs of transportation, the two firms need to be located symmetrically at the quartiles of the market. This outcome, with the market split evenly across space between the two firms, implies that no consumer has to travel further than 1/4 of the length of the whole market to reach a firm. This is contrasted with the market outcome described above, in which some consumers have to travel at least 1/2 of the length of the market to reach a firm.

This difference between the two firms locating close to each other under market competition to maximize their profits and the regularly spaced distribution to minimize social transportation costs is magnified through additional firms entering the market. In a free-market situation, the entering firm will follow the same logic as the location decision of the second firm, consequently locating very close to the existing firms and thereby creating a cluster of firms. The social planner, however, would redistribute the firms along the market in a regular pattern. This demonstrates one force leading to agglomerations within free market economies.

Hotelling then describes different situations where the logic of the model is applicable. He starts with cider merchants who have to decide on the sourness of their product, so locating their firms within a particular product space. The measure of sweetness replaces distance, and transport costs are the mismatch between product

sweetness and consumer preferences. More generally, he argues that although it is not the main force leading to the standardization of many products, ranging from furniture to houses, clothing, automobiles, and even education, it is driving the effect that changes to goods are mainly small deviations designed to capture the market in order to get between one's competitors and a mass of customers.

Generalizing the approach further, he argues that real-world settings can be seen in terms of mathematical spaces, where firms locate in multidimensional spaces to maximize the mass of customers closest to their position. He acknowledges a number of conditions that could change the shape of the outcome and push firms apart, including an uneven distribution of consumers, particular shapes of transport costs, pricing structures, and some elasticity of demand considerations.

His generalization also demonstrates that his approach is not just suitable for "economic life" but for most diverse situations with competitive activity. Describing the application to one such area he writes:

> In politics it is strikingly exemplified. The competition for votes between the Republican and Democratic parties does not lead to a clear drawing of issues, an adoption of two strongly contrasted positions between which the voter may choose. Instead, each party strives to make its platform as much like the other's as possible.

Hotelling's model has been adopted in a range of fields and applications; the use in the political space he foreshadows is probably the most influential and well-known case. Downs (1957) takes up Hotelling's idea and formalizes its application in a political setting. The following list shows his main assumptions about the setup and explains how Hotelling's spatial modeling aspects are retained, yet framed in political terms:

- *The political parties in any society can be ordered from left to right in a manner agreed upon by all voters*.
  This assumption is equivalent to using a single, fixed line as space.
- *Each voter's preferences are single-peaked at some point on the scale and slope monotonically downward on either side of the peak.*
  This assumption implies that the only factor influencing consumer/voter utility is the distance to the firm/party location, modeled by Hotelling as transport costs and by Downs as closeness of the preferences.
- *The frequency distribution of voters along the scale is variable from society to society but fixed in any one society.*
  This assumption relaxes Hotelling's assumption of a uniform distribution of consumers /voters over the line but retains that the distribution remains fixed while firms/parties make their location decisions.
- *Once placed on the political scale, a party can move ideologically either to the left or to the right up to but not beyond the nearest party toward which it is moving.*
  This assumption states that firms/parties can move freely along the space. The "no jumping over" restriction is because of consistency and gradualness of change, as

it prohibits small position changes from leading to large changes in voting behavior.

- *In a two-party system, if either party moves away from the extreme nearest it toward the other party, extremist voters at its end of the scale may abstain because they see no significant difference between the choices offered to them.* This assumption models that there is a distance beyond which transport costs/differences in preferences are too expensive or prohibitive and consumers/voters prefer no action over purchasing/voting. It counteracts the above "no jumping" rule and moves parties from clustering at extremes to the center.

The consequences of these assumptions depend on the political system, in particular the number of parties, again mirroring Hotelling's discussion of the implications of the entry of more firms beyond the initial two. The best known case is majority voting, two sides vying for a victory by positioning themselves along the political spectrum. This results in the *median voter theorem*, which states that in the case of majority voting and single-peaked preferences, the outcome will be the one preferred by the median voter, so the point of the policy space that splits the population in half.

Around the same time as Hoteling and others developed models with non-competitive firms, a number of other relevant spatial approaches arose as described by Fujita (2010). Marshall presented his work on *industrial agglomerations*, stressing the role of *externalities* in the process. Lösch (1940) sets out to build a systematic model of *location theory*, including the development of market areas and economic regions. The *central place theory* of Christaller (1933) provides a description of a conceptual theory of central places with a spatial hierarchy of locations and markets.

Fujita (2010) also describes the further development of spatial economics, especially the attention paid to general equilibrium and competitive markets.

This did lead to the result of a *spatial impossibility* theorem, which, according to Fujita's description, implies that one of the following three assumptions needs to be made to describe and understand actual phenomena about agglomerations, regional specialization, and the general spatial distribution of economic activity:

1. Space is heterogeneous.
2. Externalities in production and consumption exist.
3. Markets are imperfectly competitive.

Spatial models that were developed in the wake of this often combine more than one of these, but each has certain implications.

- *Comparative Advantage Models*: These assume that space is heterogeneous and differs in some form, for example, endowments (immobile resources like natural resources), amenities (e.g., climate), or uneven transportation costs. This differentiation allows for comparative advantage, and consequently trade, between locations even under constant returns and perfect competition.

- *Externality Models*: Nonmarket interactions between firms and/or households (knowledge spillovers, communication, and interactions) generate endogenously spatial agglomeration and thereby trade, again allowing constant returns and perfect competition.
- *Imperfect Competition Models*: Firms no longer are price-takers, and spatial distribution patterns influence price policies, potentially resulting in agglomerations. There are two approaches:
  - *Oligopolistic competition*: These models assume a finite number of large agents (firms, governments, developers), who interact strategically including location decisions.
  - *Monopolistic competition*: Firms are price-setters and usually assumed to produce differentiated goods under increasing returns. Models tend to assume a continuum of firms, thus minimizing strategic interactions.

Although models in all of these directions have been developed in the last three decades, the most prominent is the last one, which started out by using a modeling framework based on the Dixit-Stiglitz model of monopolistic competition introduced by Dixit and Stiglitz (1977). Labeled the "new economic geography" (NEG), the emphasis was on developing a unified approach of general equilibrium modeling focusing on the interactions of increasing returns, transport costs, and the movement of production factors. The spatial scope of this approach was very wide, from the organization of cities, systems of cities, and regional structures to international and global models.

Paul Krugman (1991), one of NEG's pioneers and later Nobel Prize recipient, explains his interest in developing formal modeling approaches to geographic questions with the almost complete absence of spatial considerations in the contemporary international trade literature and economics more generally. Setting out to remedy that, he points toward methodological advances on market structures, in particular monopolistic competition models and *increasing returns* approaches, that had substantially reshaped international economics in the 1980s and which seemed well suited to overcome the constraints that had pushed economic geography to the back of the minds of economists.

Krugman cites a number of empirical phenomena as motivation: the large amount of domestic trade, the decrease in border relevance in a unifying Europe, and especially the existence of concentration and agglomerations, for example, the "manufacturing belt" in the USA. Nevertheless the resulting NEG and geographically oriented economics literature developed with a strong technical and methodological focus (Krugman 1991, 2010). This development was in sharp contrast to the work of economic geographers in geography and related disciplines, who moved away from mathematical modeling and quantitative methods. For example, Martin, as cited by Krugman (2010), claims that "economic geography proper" involves a rejection of abstract models in favor of "discursive persuasion" involving "a firm commitment to studying real places (the recognition that local specificity matters) and the role of historico-institutional factors in the development of those places."

While acknowledging that some criticism of the project is justified, Krugman does argue that this strong rejection is misguided by making two points. First, a central, practical intention was to reintroduce spatial thinking into the economics mainstream, and the formal modeling approach was the most successful way to do that. Second, a more conceptual argument for the approach was that it puts a stronger emphasis on "what-if" questions. Formal modeling allows the development of counterfactuals and scenarios, something that ultimately is substantially more suitable to analyze and evaluate policies, interventions, and other changes than an approach emphasizing the uniqueness of each case (Krugman 2010).

History and First and Second Nature Geography

The development of new economic geography modeling was strongly driven by the desire to understand agglomeration patterns of economic activity and population. What are the reasons that people cluster at certain locations rather than spread out evenly over the landscape?

The literature began to categorize the modeling components that were used to generate these patterns into two categories, namely, *physical geography* and *agglomeration forces*. Krugman (1993) popularizes these as "first nature geography" and "second nature geography" based on ideas by historian William Cronon in his economic and social history of Chicago, *Nature's Metropolis: Chicago and the Great West* (Cronon 1991). He outlines Cronon's argument that Chicago's rise to prominence is not primarily based on a distinctive natural advantage. The city is on a flat plain, its river barely navigable, and its harbor inadequate, but due to a self-reinforcing strength, Chicago established itself as a central market and focal point for transportation and commerce. Krugman summarizes:

> As Cronon puts it, the advantages that "first nature" failed to provide the city were more than made up for by the self-reinforcing advantages of "second nature": the concentration of population and production in Chicago, and the city's role as a transportation hub, provided the incentive for still more concentration of production there, and caused all roads to lead to Chicago. (Krugman 1993)

The new economic geography literature, and economic history research utilizing this modeling approach, usually associates "first nature" with physical geography, describing location characteristics and environments that are also referred to as endowments. These physical attributes, such as natural resource deposits, rivers, mountains, or rich agricultural soils, are usually location specific and immovable, thereby clearly distinguishing different locations in terms of their presence or extent.

"Second nature" refers to mechanisms underlying the spatial distribution of population and economic activity that are only dependent on the existing shape of the distribution. The main aspect is agglomeration and the emergence of spatial clusters. Krugman (1991) and Kim (1995) show that industries in the USA show strong agglomeration structures, rising from the middle of the nineteenth century

until a peak in the early twentieth century, and a consistent flattening out over time after that. Spillovers and externalities are potential sources, though the NEG approach focuses predominantly on the effects of *increasing returns* to cause agglomerations, in particular in combination with the impact of transportation costs and *market access*.

The question of first and second nature and the use of new economic geography models has two main contact points with economic history. The first is the use of historical developments and events to test the validity of model specifications, in particular the relevance and predictions of particular first and second nature phenomena. The second is the application of such models to understand the underlying reasons and consequences of historical developments. These uses of modeling are similar to the usage of spatial concentration indices mentioned above.

The modeling of the new economic geography is usually focused on cross sections and eventual comparative statistics; however this restricts the possibilities to empirically validate a number of model predictions and mechanisms. Consequently, a number of authors have turned to economic history, long-term developments, and historical events to test a number of models and mechanisms in quantitative economic geography.

One important example is Davis and Weinstein (2002). They use thousands of years of Japanese regional population density as well as population growth throughout the twentieth century, especially up to and in the wake of the Second World War and its devastating impact on the country. They don't build an explicit, formal spatial model but derive a number of stylized facts about spatial and other aspects of the characteristics of the regional population distribution and its change over time. They compare their empirical observations with the predictions coming from theoretical approaches in the literature grouped in three explanatory theories, namely, *increasing returns*, *random growth*, and *location fundamentals*. Their label "increasing returns" references a set of results predominantly from the new economic geography that stresses "second nature" mechanisms and the role of agglomeration forces. "Random growth" refers to an approach that sees the population growth as a random process rather than a structurally determined one. The theory is focused on explaining Zipf's law, a mathematical regularity of the city size distribution, by arguing that that distribution can be the outcome if city growth is a random walk and Gibrat's law, the independence of city size and growth, holds. "Location fundamentals" focuses on the effect of the "first nature" geography, so explains the spatial distribution with usually permanent geographic characteristics of these locations.

Davis and Weinstein summarize the fit of the stylized facts of Japan's spatial population history with these three theories as follows (Table 1).

The authors argue that no single theory clearly dominates. The existence of substantial spatial variation of Japan's population density for a long time, especially also before a tight market integration between regions and the rise of the industrial production, points to a substantial role for location fundamentals, while the increase in variation with the Industrial Revolution points to a role for increasing returns in shaping the modern distribution. For the random growth approach, however, they argue that the observed pattern after the shock of the Second World War, which

Table 1 From Davis and Weinstein (2002)

Stylized fact	Increasing returns	Random growth	Location fundamentals
Large variation in regional densities at all times	–	+	+
Zipf's law	–	+	+
Rise in variation with the Industrial Revolution	+	–	–
Persistence in regional densities	?	?	+
Mean reversion after temporary shocks	?	–	+

showed a mean reversion and a reversal of the shock, strongly contradicts the core assumption of the theory. If the growth process really were a random walk, then such a substantial shock should have left a permanent impact rather than a temporary one.

Redding and Sturm (2008) take a different approach. They focus primarily on the relevance of *market access*, a central aspect of increasing returns and the main agglomeration mechanism used in the new economic geography, construct a formal theoretical model, and then test its predictions empirically with a historical episode. Their formal approach is a multilocation version of a model initially developed by Helpman (1998) to explain the size of regions. They combine two agglomeration mechanisms: increasing returns in production and lower living costs due to a stronger competition in larger markets, with two dispersion forces – stiffer competition in larger markets and a fixed amount of a local, immovable amenity. They identify the effect of change in market access through a change in transportation costs while assuming that the level of amenity in each location is fixed. Empirically they test the model prediction, a stronger reduction in population in locations that see a larger decrease in market access, with the German separation after the Second World War. They calibrate their model with German data before the war, then introduce the shock to market access caused by the separation of the two German states, and demonstrate that the observed change in city populations is explained by that change in market access.

Redding and Sturm focus their model and analysis on the effect of agglomeration forces and the importance of market access without paying much attention to the role of endowments and location characteristics. Ploeckl (2012) takes this up and utilizes the model applied to the case of Saxony from the sixteenth to the nineteenth century to show empirically that the implied amenity values of the model can be explained by particular location characteristics. Conducting the analysis for a number of distributions over time shows that the relative importance of particular location characteristics changes. Similarly, Ploeckl (2015) uses the model and the German Empire of the late nineteenth century to empirically show that geographical characteristics like rivers, which are usually used to proxy for the "first nature" effect, can also influence the impact of market access and therefore the presence and strength of the "second nature" mechanisms. Both studies provide evidence for Davis and Weinstein's conclusion that first and second nature are not exclusive and that both together shape the spatial distribution of population and economic activity.

Explicit spatial modeling can also be used to revisit historical analyses that include spatial aspects without them being the main question of interest. Fogel's (1964) analysis of the impact of the railroad on American GDP uses a spatial approach in some parts. His idea to build a *counterfactual* based on changes in land value and use through the proximity of railroads echoes von Thünen's approach to land rents and the relative proximity to demand from the city. Donaldson and Hornbeck (2016) take up this spatial aspect and construct a spatial model that links the value of agricultural land in each US county with its market access in the rest of the country. A counterfactual analysis with market access based on a scenario of transport links without railroads results in a substantial decrease in agricultural land values and an economic loss in GDP terms only modestly larger than Fogel's comparable estimate.

Formal Modeling

As the median voter theorem and the growth impact of American railroads demonstrate, explicit modeling of space is done in many fields and situations. This wide range of application implies that the range of modeling approaches and included mechanisms is extremely wide. Even if the focus is only on location decisions, the discussion above about first and second nature geography implies that the relevant mechanisms are manifold and formal modeling needs to incorporate quite a number of them.

Redding and Rossi-Hansberg (2017) provide a survey about model components and assumptions, the building blocks used in models in quantitative spatial economics. They group their overview into a number of main aspects, namely, *preferences*, *endowments*, *production technology*, *trade*, *technology and idea flows*, *labor movements*, and *equilibrium* characteristics. The following discussion goes through each point listed and explained by Reading and Rossi, combining a short summary and discussion of the relevance and applicability of the these aspects for economic history.

Preferences

Homogeneous Versus Differentiated Goods

A first important implication of the structure of consumer preferences is the question of the homogeneity of goods. As indicated above, new economic geography models emphasize *product differentiation* and love of variety by consumers.

The international trade literature has shown the quantitative welfare gains from new varieties (Broda and Weinstein 2006), while economic history also looks at the gain from new goods, products, or technology from a social savings perspective (Leunig 2011). The relevance of modeling differentiated goods, however, has a dual purpose besides capturing these gains, namely, to include a mechanism that generates trade between locations without being purely based on endowment or technology differences.

Single Versus Multiple Sectors

Another dimension of the structure of consumer preferences is the range of goods produced in the economy. Technical constraints, in particular tractability of the model, mean that many models only use a single production sector or maybe large aggregate sectors like agriculture and manufacturing. With advances in modeling and computation, it is now feasible to expand on this and incorporate a larger number of sectors in the models.

The decision about the number of sectors involves trade-offs between the precision of the model and its complexity and tractability. Certain important economic developments relevant for economic history, for example, the structural transformation of the economy from agricultural to industrial as demonstrated for the USA by Michaels et al. (2012), fit well with just aggregate sectors. Consequently, an extension to a larger number is only preferable when the interaction between sectors is the central research question and there is enough differentiation in data and sector characteristics to exploit these.

Exogenous and Endogenous Amenities

Early new economic geography models assumed no real differences between the characteristics of locations, similar to the featureless plain of von Thünen. This implies that there are no ex ante differences for consumers in the locations themselves that could explain location decisions and the spatial distribution of economic activity. Certain geographic characteristics, for example, access to water or scenic views, however, are potentially relevant for productivity differences as well as affecting the utility of consumers directly. These influential factors, *amenities*, can therefore either be exogenous or arise endogenously, and recent modeling advances allow for the presence of both.

Fixed Local Factors in the Utility

A particular case of an exogenous "amenity" is an exogenous factor with a fixed supply. The central example for such a factor is residential land. The main contrast to regular amenities is the difference in access. This factor is rivalrous while regular ones are not. If such a factor is included in the utility of actors, then its fixed nature in conjunction with location size will act as an attraction or as a congestion or dispersion force, depending on the relevant size of the population to the amenity.

This particular case of an exogenous amenity serves primarily as a simple representation of amenities. This allows their inclusion in the modeling process and empirical estimation without requiring an explicit specification of which amenities to include (Redding and Sturm 2008; Ploeckl 2010). This is especially relevant for economic history as it allows for the inclusion of amenities without requiring a full set of relevant amenities or complete data for all observations.

Common Versus Idiosyncratic Preferences

This assumption models whether agents have identical *preferences* or whether there are idiosyncratic differences between them. These differences are usually directly

over locations, though differences in preferences over location amenities should have similar consequences. In combination with perfect mobility and no arbitrage, this assumption results in real wage equalization across locations in cases of common preferences and real wage differentials, if incomes paid to attract agents to a location differ with the heterogeneity of their preferences.

Differences in preferences and tastes have not yet received a lot of attention in economics and economic history. A major reason is that in contrast to differences in skills and productivity, the origins and measurement of these are rather problematic. One recent approach to explain regional differences in food preferences is presented by Atkin (2013), who models habit formation of consuming locally prevalent foods leading to a "home" effect similar to that in trade.

Production Technology

Constant Versus Increasing Returns

As indicated above, a central mechanism of new economic geography models is the presence of increasing returns to scale as a way to generate self-reinforcing agglomeration forces. This mechanism is usually linked with love of variety to result in specialization by and exchange over spatial locations. Recent modeling advances however have demonstrated that constant *returns to scale* production technology in combination with other mechanisms, in particular an Armington differentiation by location of origin and Ricardian technology differences, can result in a similar specialization.

The nature and shape of the production function, including the implied returns to scale, are the subject of interest in a number of fields, including economic history. The main motivation for the choice depends on the focus of the question, knowledge about empirical evidence, estimates of returns to scale in the sector, or the desired mechanism generating trade in the produced goods. One possibility is to model increasing returns as the combination of fixed and constant marginal costs (Redding and Sturm 2008), which allow the relative importance of fixed costs, and consequently the relative strength of increasing returns, to vary.

Exogenous and Endogenous Productivity Differences

Similar to the possible presence of exogenous and endogenous amenities, locations can differ in their *productivity* in both ways as well. The starting point of a featureless plain obviously is linked to a stronger focus on endogenous differences, for example, knowledge spillovers. International trade models and empirical observations, however, do emphasize the case for a role of exogenous productivity differences, such as access to local mineral resources. This is an important underlying mechanism for the differentiation of first and second nature effects, mirroring the modeling choice between exogenous and endogenous amenities.

This choice also affects a substantial number of other mechanism components. If the model includes endogenous productivity differences, the process underlying these differences needs to be included, for example, the transfer of ideas between

locations. Economic history has established productivity differences. However, the question about their endogeneity or exogeneity depends on the particular context, production process, and focus of the research question.

Input-Output Linkages

Assumptions about *input-output linkages*, especially over locations, are an important aspect explaining how locations are linked. They strongly shape how productivity shocks in particular regions and sectors spread through the wider economy. Such linkages also provide an additional mechanism for agglomeration.

Input-output linkages have been an important topic in economic history, for example, in Fogel's analyses of the railroad and its interaction with the iron industry in the USA (Fogel 1964). Another aspect is the importance of natural resources such as the availability of water power and access to coal for industrial development (Crafts and Wolf 2014). Baldwin (2016), however, contrasts the emergence of long value chains with intermediate goods as a result of modern communication technology with the first globalization with its comparatively short chains, clustered production, and transportation of final, tradeable goods. If the focus is on manufacturing as a sector rather than differences between individual industries, certain input relationships, in particular the relevance of local natural resources, can be modeled as exogenous regional productivity differences in the aggregate sector.

Fixed Local Factors in Production

Again, similar to utility, there can also be a fixed local factor in production. A fixed supply of land can apply not only to residential but also to commercial land used in production. This mechanism has similar consequences and can act as a congestion and dispersion force.

This mechanism is the production side equivalent to the fixed amenity factor. The main application in an economic history setting appears to be agricultural land, where the assumption of a fixed amount available as input for production sounds reasonable. The modeling purpose is then to include agricultural land availability as either an attraction or dispersion force.

Trade

The inspiration for the renaissance of spatial modeling in economics came from the field of international trade, and central trade theories remain at the core of the literature (Krugman 2009). The rise of NEG has focused attention to distinguish the relevant importance of endowment-based models linked to *Heckscher-Ohlin trade* structures and the increasing returns and *monopolistic competition models* underlying NEG. The afore described results by Kim (1995) and Davis and Weinstein (2002) indicate that a combination of both is applicable. Similar results are found by Wolf (2007) for Poland in an analysis following Midelfart-Knvarvik et al. (2000) who develop a model designed to test for H-O and NEG mechanisms. Recently, new models based on *Ricardian trade mechanisms* have been developed

that have seen successful adoption to economic history (Donaldson 2018). As trade itself is the result of the structure of the production process, endowments, and preferences, this section is focusing on trade costs and frictions affecting those trade patterns.

Variable Versus Fixed Trade Costs

Modeling the exchange of goods between locations requires some assumption about the cost structure of that trade. Commonly this is done by including so-called iceberg variable *transport costs*, which require that $d_{ij} > 1$ units are shipped from location i such that one unit arrives in location j (so that the difference "melts" away en route).

Von Thünen already included variable trade costs into his calculations by specifying them as the weight of the grains displaced by the fodder necessary for the horses pulling the grain carts to market (von Thünen 1875). Although there are a number of industries and sectors where fixed costs may seem more appropriate, most economic and economic history studies are utilizing variable trade costs. One reason is that the gravity equation, with its prediction of lower flows over longer distances, seems to hold even for industries that have distance-independent trade costs. One example is the communication sector. Individual phone calls and letters do not really cause substantially different costs depending on distance. This has led to the introduction of distance-independent transaction costs for consumers, most notably the penny post, and yet the volume of letters and calls decays with distance (Lampe and Ploeckl 2014).

Geographic Versus Economic Frictions

Differences and variations in trade costs arise in general due to some type of friction. The sources for these can be geographic (e.g., mountains with their elevations) or economic (e.g., institutional rules like borders or infrastructure like road and rail networks). The presence of frictions implies that a full set of transport costs between any two locations is usually not directly available. Empirical studies usually rely on geographic information system approaches to estimate these, either based on finding the cheapest path through a network (e.g., a railroad network) or the least-cost path over a cost surface (e.g., an elevation map).

One historical example for the estimation of transportation costs with GIS methods incorporating geographic frictions is Ploeckl (2010), who applies a least-cost path algorithm over a cost surface that includes geographic cost factors, such as elevation and rivers, and economic ones, like the road network. Jaremski (2012) and Donaldson and Hornbeck (2016) take a network approach by determining transport costs between two points by finding the shortest path through a network between linked locations. This requires a specification of costs for traversing each network link, and both papers assign those values based on transport infrastructure. This means that the geographical length of each connection is multiplied with a per mile cost based on available transport technology for the link, for example, wagons, canals, or railroads. Although the basic choice of different costs based on infrastructure type is an economic friction, geography has a potential influence through which links are available at all. A simple extension to include them more explicitly is to modify cost values for a link based on the geographic conditions of its location.

Economic frictions can have a wide range and depend on the scale of the analysis. One less obvious example is underlying social differences, like the ethnic composition of location populations, as Schulze and Wolf (2009) demonstrate for the Habsburg Empire. Language is another potential friction, even in the case of differing dialects (Lameli et al. 2015).

Asymmetric Versus Symmetric Transport Costs

One aspect of *transport costs* is the assumption that they are symmetric (so costs of transport from i to j are the same as those from j to i), which has consequences for patterns of trade and income as well as equilibrium outcomes. As indicated above, the standard formulation of iceberg trade costs is symmetric as d_{ij} only depends on the geographic distance between i and j, which is obviously identical to that between j and i. This is similar to the symmetry of the gravity equation, where the size of trade flows is also explained by simple distance between the two trade partners.

Geographic as well as economic frictions, however, may not have symmetric effects. Going up a mountain is different than going down; one country's tariff rates usually do not equal those of trade partners. There are instances where this is ignored, for example, many market integration studies use the correlation between prices even if the directions of underlying trade flows, and therefore applicable transport costs, are not identified. Depending on the setting, differences might be negligible. For example, Ploeckl (2010) demonstrates that transport costs in both directions converge to the same level, well approximated by plain distance, as certain frictions become less relevant the further is the total distance covered.

Role of Non-traded Goods

Another special case is the presence of non-traded goods, which represents a situation where the trade costs of these goods between locations is (or approaches) infinity. The presence of these goods nevertheless affects input-output linkages and the effects of productivity differences and shocks on expenditures.

Non-traded goods have a larger importance in international trade. On a more regional scale, they can be used as a model mechanism to retain production and population in certain locations.

Technology and Idea Flows

The incorporation of *technology and idea flows* into spatial modeling works predominantly through affecting local productivity. This implies that there must be productivity differences between locations that are at least partially endogenous.

Knowledge Externalities and Diffusion

One aspect is whether there are *externalities* with regard to knowledge and ideas and if so how and how far they diffuse. Going back to Marshall, such externalities are commonly seen as knowledge spillovers, externalities due to thick labor markets, and backward and forward linkages (Fujita 2010).

The economic history literature has established Marshallian externalities in a number of settings, especially in urban history and the development of particular industries and regions. An important modeling decision, in combination with the scale of the analysis, is the spatial extent of externalities.

Innovation

A related aspect is the question of *innovation*, which might affect local productivity as an endogenous, intentional investment. The main criteria for such investment in innovation is whether the returns can be successfully appropriated. This does depend on the speed with which ideas and innovations diffuse to other agents and locations. Modeling advances link the profitability of innovation investment to the size of the markets accessible to firms and agents, capturing the extent to which they can distribute the costs across consumers. Practical modeling is utilizing a competitive market and a potential capitalization of innovation returns in land rents as a mechanism.

Including innovation into the spatial model is predominantly relevant for situations where intentional research is assumed to play an important role. This points toward situations with larger geographical scopes, in particular countries. Historically, individual inventors have played substantial roles for technological progress, but spatial modeling is suited better for aggregating their inventive activity in locations or counties. For the return to innovation aspect, market access provides an empirical estimation fitting the historical evidence (Sokoloff 1988).

Transferability of Ideas

Another aspect is whether the *diffusion* of ideas between agents and locations is costless or not due to the presence of frictions. This is linked to international trade, where ideas transferred through foreign direct investment might not result in the same productivity improvement in other countries, or to mechanisms that require firms to face adjustment costs for transferring production innovations developed for a particular location.

The main aspect here in contrast to externalities is the intentional transfer of technology, a special version of innovation and the endogenous increase of productivity. Although technical knowledge, such as patents, was traded historically (Burhop and Wolf 2013), it is primarily relevant for questions dealing with multisite firms, including modeling the decisions between exporting or foreign direct investment. In other contexts it can be subsumed into the characteristics of the innovation mechanisms underlying endogenous productivity changes.

Labor Movements

Migration Costs

The ability of agents to move and change locations is an important condition for an integrated spatial model. The cost of such *migration*, however, can have implications. The presence of such frictions and moving costs provides a mechanism

explaining real wage differentials between locations, which requires some of the above described mechanisms such as idiosyncratic preferences to exist in the absence of migration costs. Empirically, moving costs and frictions are more strongly linked to settings with an international scale rather than local and regional analyses.

One particular setting where explicit moving costs become relevant is the modeling of international migration, especially the aspect of destination choice. Transatlantic mass migration in the twentieth century is one setting where real wage differentials act as an incentive to move, and particular moving cost components, for example, chain migration, affect location choice over different countries as well as locations within countries (Hatton and Williamson 1998).

Commuting

A more local, urban aspect of labor movements is the potential presence of *commuting*, so a separation of workplace and residence or production and consumption. Models explaining the shape of cities traditionally started with locating production activity at the center, while residential location and linked land prices are mainly determined by commuting costs to the center. More recent advances expand the analysis to non-monocentric patterns but retain a strong link between agglomeration forces and commuting costs. Commuting costs are consequently also a potential mechanism to explain differences in city sizes within systems of cities.

Commuting is the local equivalent of migration costs for regional and international settings. It is especially relevant as a separate mechanism if the focus is on urban settings and the explicit separation of residential and production location (Ahlfeldt et al. 2015). Historically, commuting is less of a concern due to constraints on individual mobility that made significant commuting very costly and fairly small. If the focus is not on commuting directly, it can be subsumed into local productivity differences, for example, von Thünen discusses additional production costs due to distance between estate and fields, or as a local amenity where commuting time influences the utility residents receive from particular locations.

Skills and Heterogeneity

Migration and commuting decisions are influenced by modeling decisions about common or idiosyncratic preferences and productivities of agents over locations. Idiosyncratic differences can be used to explain the empirically observed characterization of migration and commuting as following a gravity equation pattern. A closely related mechanism is whether there are systematically different types of agents who value location characteristics differently. Such a mechanism leads to equilibrium distributions of economic activity that are characterized by *spatial sorting*, in which agents of a given type self-select into particular locations based on the characteristics of the locations.

Spatial sorting according to particular migrant and location characteristics can be observed in transatlantic mass migration, where selection based on the presence of other immigrants from the same ethnic or cultural background occurred. Another important example is the spatial sorting based on preferences, as described by

Schelling's model of segregation (Schelling 1971). Racial segregation is one important application for this (Boustan 2010, 2012).

Congestion in Transportation

Another choice for models with explicit flows of people is congestion. Are travel costs linked to the volume of flows? Can infrastructure provision reduce congestion? In urban models this relates to issues concerning whether an increased provision of highways leads to an increase in vehicle travel without an impact on congestion and whether public transport significantly affects travel delays and congestion.

This is the labor flow equivalent of trade costs previously discussed and an endogenous version of commuting costs. It has a substantially more limited application in economic history as its main relevance is for scenarios about urban infrastructure, in particular in the context of residential choice.

Endowments

A number of previously described mechanisms build upon specific characteristics of locations. Their starting values need to be exogenously specified. *Endowments* in this context refer therefore to the initial distribution of these characteristics underlying the incorporated mechanisms.

Spatial Scope and Units

The model requires decisions about its scope and spatial units. Is it concerned with a local setting in a single city, a systems of cities, regions within countries, or international, even global structures? Is space a set of locations or continuous? And is it one- or two-dimensional? Spatial units are often discrete, based on available data as indicated above, but questions of aggregation need to be addressed. Clearly spatial scope and units are linked, and modeling decisions and data availability for each constrain available choices for the other. A related decision is the question of mobility of factors, which are mobile over spatial units and which are fixed, and if so at what level of aggregate units.

Population and Skills

The above described mechanisms covering idiosyncratic preferences, differential labor productivities, or multiple types of agents need to be reflected in the endowments in terms of population and the distribution of types and characteristics over locations.

While the differences in characteristics like skill level and productivity are implied by choices about other mechanisms above, the endowment structure and starting distribution of the characteristics can have a substantial impact on empirical economic history analyses. Particular aspects are spatial sorting mechanisms, for

example, the urban-rural distinction, racial segregation, and migration destination decisions that depend on the initial distribution for a sound understanding.

Capital and Infrastructure

The presence of *capital* as a production factor allows for these to be mobile across locations. For tractability reasons such physical capital is, however, usually assumed to fully depreciate in each time period. Besides capital as a production factor within locations, the structure, and therefore endowment, of an economy might contain *infrastructure* between locations. Depending on assumptions and modeling choices about movements of goods, ideas, people, and capital, the nature of costs and frictions about these movements can be linked to the infrastructure and transport networks in the economy, which are usually treated as exogenous in current models.

The impact of this infrastructure between locations, however, can be subsumed into the productivity or amenity differences of affected locations or into transportation and congestion cost frictions. This might be especially relevant for economic history since data requirements increase substantially with the number of locations, and necessary historical information may be incomplete or even unavailable.

Equilibrium

The *equilibrium* of spatial models may require a number of technical assumptions. These are contingent on the selection of particular mechanisms and components, but the most relevant are market structure, general versus partial equilibrium, rent distribution, and trade balance. Market structure is linked to, and needs to be consistent with, the returns to scale assumption, so increasing returns are usually connected with monopolistic competition. The second concerns the completeness of the imposed equilibrium conditions, where partial equilibrium allows for an exogenous outside of the model. The question of rent distribution concerns whether the distribution of land rents is fully incorporated in the model or takes a partial equilibrium approach using an "absentee landlord" outside the model. Finally, because most models include completely balanced exchange, yet empirical data show the existence of persistent imbalances, the balance of trade between locations must be considered. There are a number of fixes, though they are exogenously imposed rather than a full endogenous modeling of local consumption and savings.

Another aspect is the uniqueness of equilibrium. Spatial models may have multiple equilibria without a clear selection (Fujita et al. 1999). While in more theoretical settings the interest is on a particular statistic or other equilibrium characteristics that may not depend on selecting a particular equilibrium, the application of spatial models to economic history is usually focused on an empirical outcome and therefore requires the selection of an equilibrium. This might be done through specifying an initial distribution to determine the outcome, which does require a quantitative specification of values for the mechanism described in the Endowment section above.

Implementation

This long list of different potential model components obviously requires some approach or criteria in order to select among them. There are some obvious combinations, where the choice in one aspect determines the choice in another aspect.

A starting point is the relevance of the model for the available data. Unless the model is designed for primarily theoretical reasons, the empirical setting, the nature of the overarching research question, and the structure and extent of relevant data should motivate the model component selection.

Even if sufficient data are available, the formal model should remain analytically tractable. The inclusion of more components increases the difficulty in obtaining an analytical solution, ascertaining the uniqueness of an equilibrium, and deriving comparative statics. This also applies to computational aspects, where technological advances in computing power and speed have made it possible to solve large-scale problems. Nevertheless, it is not always clear a priori what the required computing resources are to achieve a solution.

The question of data and model complexity also includes assumptions about structural parameters required by certain modeling choices. A number of modeling choices require either the specification of constraints or even direct specification of parameter values. For example, the selection of standard iceberg trade costs assumes that the parameter specifying the "melting" costs is negative in the theoretical model. An empirical application of the model then further requires a specific value for the parameter. For some parameters, for example, those concerning trade costs, there might exist estimations of suitable values for historical contexts, while for others economic historians have to rely on modern estimates or even on informed guessing in their parameter choice. There is also the possibility that the model is designed to calibrate certain parameters as a first step. However, this does increase the complexity and complicates tractability.

A final decision is the delineation about where outside interventions, in particular policy interventions, affect components and which parts remain invariant. For example, a partial equilibrium setup might include an exogenous world price, which consequently requires an assumption about whether price changes in the domestic market as a result of some intervention have any implications for the world price. And if productivity is taken as invariant to interventions, is it completely exogenous, or do other modeling choices imply externalities and spillovers that could affect productivity as a result of an intervention?

Alternative Approaches

The discussion has focused on spatial modeling centered around location theory, that is, explaining the cross-sectional spatial distribution of people, firms, or economic activity more generally. However, as alluded to earlier, spatial modeling has a very broad application, and consequently there are other approaches that essentially model spatial relationships.

The most basic approach to spatial modeling is that spatial correlation is noise and nuisance, something to be ignored in formal analysis and simply corrected for in empirical estimation. This can be done through correction for spatial correlation in econometric analysis or through the inclusion of spatial variables, such as particular regional dummies or geographic coordinates.

As indicated there are tools and methods to analyze the randomness of events in space more extensively and systematically. The so-called *spatial point process* approach differs from the location theory approach in the main assumption about the nature of the outcome (Baddeley et al. 2015). Location theory sees the outcome as the choice of actors, while spatial point processes conceptualize it as the realization of an underlying random process. The methodology has substantial applications in fields like biology and epidemiology where the observed outcomes are phenomena and events, from disease cases to tree locations, that are not economic actors locating in space. Mirroring the first and second nature aspects, the methodology incorporates underlying covariates, usually in the form of surfaces like elevation or temperature, to represent the influence of endowments. As a result, the likelihood of an event at a particular location depends on the particular characteristics of that location, and the attraction process between events, where the likelihood that an event occurs at a particular location depends on the locations of the other events.

Similarly, the discussion about distance measures above mentions *networks*, which have begun to receive substantial attention within economics and economic history (Jackson 2008). Networks differ in their approach to space by focusing only on the relative position between the nodes rather than on the absolute position on a line, plane, or multidimensional space. While networks can reflect actual geographic links, for example, railroad tracks, the approach is usually much more suitable for spatial modeling within non-geographic spaces. These can be varied, from networks of family ties in medieval towns to interlocking directorates between banks to trade agreements between countries. Edges and nodes of such networks allow us to determine a wide range of spatial aspects, most prominently the importance of actors and how central they are within the larger network.

To conclude, one particular aspect that has not been explicitly addressed in this chapter is the question of dynamics. Spatial modeling is not only useful to understand cross-sectional structures but also explicitly dynamic developments. It has the potential to explain the dynamic decision processes underlying a large range of economic phenomena and developments from technology diffusion to infrastructure development to international agreement formation.

Cross-References

► Cliometric Approaches to International Trade

References

Ahlfeldt GM, Redding SJ, Sturm DM, Wolf N (2015) The economics of density: evidence from the Berlin Wall. Econometrica 83(6):2127–2189
Anselin L (1995) Local indicators of spatial association—LISA. Geogr Anal 27(2):93–115
Atkin D (2013) Trade, tastes, and nutrition in India. Am Econ Rev 103(5):1629–1663
Baddeley A, Rubak E, Turner R (2015) Spatial point patterns: methodology and applications with R. CRC Press, Boca Raton
Baldwin R (2016) The great convergence information technology and the new globalization. Harvard University Press, Cambridge
Beckmann MJ (1972) Von Thünen revisited: a neoclassical land use model. Swed J Econ 74(1):1–7
Boustan LP (2010) Was postwar suburbanization "white flight"? Evidence from the black migration. Q J Econ 125(1):417–443
Boustan LP (2012) Racial residential segregation in American cities. In: Brooks N, Donaghy K, Knaap G (eds) The Oxford handbook of urban economics and planning. Oxford University Press, Oxford, pp 1–25
Braudel F (1981) Civilization and capitalism, 15th–18th century. Volume 1. The structure of everyday life. Harper and Row, New York
Broda C, Weinstein DE (2006) Globalization and the gains from variety. Q J Econ 121(2):541–585
Burhop C, Wolf N (2013) The German market for patents during the second industrialization, 1884–1913: a gravity approach. Bus Hist Rev 87(1):69–93
Christaller W (1933) Die Zentralen Orte in Su¨ddeutschland. Gustav Fischer Verlag, Jena
Cliff A, Ord J (1981) Spatial processes: models & applications. Pion, London
Crafts N, Wolf N (2014) The location of the UK cotton textiles industry in 1838: a quantitative analysis. J Econ Hist 74(4):1103–1139
Cronon W (1991) Nature's Metropolis: Chicago and the great west. W W Norton & Co Inc, New York
Davis DR, Weinstein DE (2002) Bones, bombs, and break points: the geography of economic activity. Am Econ Rev 92(5):1269–1289
Dixit AK, Stiglitz JE (1977) Monopolistic competition and optimum product diversity. Am Econ Rev 67(3):297–308
Donaldson D (2018) Railroads of the raj: estimating the impact of transportation infrastructure. Am Econ Rev 108:899–934
Donaldson D, Hornbeck R (2016) Railroads and American economic growth: a "market access". Approach Q J Econ 131(2):799–858
Downs A (1957) An economic theory of political action in a democracy. J Polit Econ 65(2):135–150
Duranton G, Overman HG (2005) Testing for localization using micro-geographic data. Rev Econ Stud 72(4):1077–1106
Ellison G, Glaeser EL (1997) Geographic concentration in U.S. manufacturing industries: a dartboard approach. J Polit Econ 105(5):889–927
Ellison G, Glaeser EL, Kerr WR (2010) What causes industry agglomeration? Evidence from Coagglomeration patterns. Am Econ Rev 100(3):1195–1213
Fogel R (1964) Railroads and American economic growth: essays in econometric history. The Johns Hopkins Press, Baltimore
Fujita M (2010) The evolution of spatial economics: from thünen to the new economic geography. Jpn Econ Rev 61(1):1–32
Fujita M, Krugman P, Venables AJ (1999) The spatial economy: cities, regions, and international trade. MIT Press, Cambridge
Gutberlet T (2014) Mechanization and the spatial distribution of industries in the German empire, 1875 to 1907. Econ Hist Rev 67(2):463–491. https://doi.org/10.1111/1468-0289.12028
Hatton TJ, Williamson JG (1998) The age of mass migration: causes and economic impact. Oxford University Press, New York

Helpman E (1998) The size of regions. In: Pines D, Sadka E, Zilcha I (eds) Topics in public economics: theoretical and applied analysis. Cambridge University Press, Cambridge, pp 33–54
Hotelling H (1929) Stability in competition. Econ J 39(153):41–57
Jacks DS (2005) Intra- and international commodity market integration in the Atlantic economy, 1800–1913. Explor Econ Hist 42(3):381–413
Jackson MO (2008) Social and economic networks. Princeton University Press, Princeton
Jaremski M (2012) Estimating antebellum passenger costs: a hub-and-spoke approach. Hist Methods 45(2):93–101
Kim S (1995) Expansion of markets and the geographic distribution of economic activities: the trends in U.S. Q J Econ 110(4):881–908
Kopsidis M, Wolf N (2012) Agricultural productivity across Prussia during the industrial revolution: a Thuenen perspective. J Econ Hist 72(3):634–670
Krugman P (1991) Geography and trade. Leuven University Press, Leuven
Krugman PR (1993) On the relationship between trade theory and location theory. Rev Int Econ 1 (2):110–122
Krugman P (2009) The increasing returns revolution in trade and geography. Am Econ Rev 99(3):561–571
Krugman P (2010) The new economic geography, now middle-aged. Presentation to the Association of American Geographers, 16 Apr 2010
Lameli A, Nitsch V, Südekum J, Wolf N (2015) Same same but different: dialects and trade. Ger Econ Rev 16(3):290–306
Lampe M, Ploeckl F (2014) Spanning the globe: the rise of global communications systems and the first globalisation. Aust Econ Hist Rev 54(3):242–261
Lampe M, Sharp P (2016) Cliometric approaches to international trade. In: Diebolt C, Haupert M (eds) Handbook of cliometrics. Springer, Heidelberg, pp 295–330
Leunig T (2011) Social savings. Economics and history: surveys in cliometrics. Wiley, Somerset, pp 21–46
Lösch A (1940) Die Räumliche Ordnung Der Wirtschaft. Gustav Fischer Verlag, Jena
Michaels G, Rauch F, Redding SJ (2012) Urbanization and structural transformation. Q J Econ 127(2):535–586
Michalopoulos S, Papaioannou E (2018) Spatial patterns of development: a Meso approach. Ann Rev Econ 10:383–410
Midelfart-Knvarvik K, Overman H, Redding S, Venables A (2000) The Location of European Industry. Technical report, Directorate General for Economic and Financial Affairs, European Commission
Ploeckl F (2010) Borders, market access and urban growth; the case of Saxon towns and the Zollverein. Documents de Treball de l'IEB 2010(42)
Ploeckl F (2012) Endowments and market access; the size of towns in historical perspective: Saxony, 1550–1834. Reg Sci Urban Econ 42(4):607–618
Ploeckl F (2015) It's all in the mail: the economic geography of the German empire. University of Adelaide School of Economics Working Papers 2015(12)
Ploeckl F (2017) Towns (and villages): definitions and implications in a historical setting. Cliometrica 11(2):269–287
Redding SJ, Rossi-Hansberg E (2017) Quantitative spatial economics. Ann Rev Econ 9(1):21–58
Redding SJ, Sturm DM (2008) The costs of remoteness: evidence from German division and reunification. Am Econ Rev 98(5):1766–1797
Schelling TC (1971) Dynamic models of segregation. J Math Sociol 1:143–186
Schulze MS, Wolf N (2009) On the origins of border effects: insights from the Habsburg empire. J Econ Geogr 9(1):117–136
Sokoloff K (1988) Inventive activity in early industrial America: evidence from patent records, 1790–1846. J Econ Hist 48:813–850
Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46:234–240

von Thünen JH (1875) Der isolirte Staat in Beziehung auf Landwirthschaft and National¨okonomie, 3rd edn. Wiegandt, Hempel & Paren, Berlin
Wolf N (2007) Endowments vs. market potential: what explains the relocation of industry after the polish reunification in 1918? Explor Econ Hist 44(1):22–42
Wrigley EA (1985) Urban growth and agricultural change: England and the continent in the early modern period. J Interdiscip Hist 15(4):683–728

Historical Measures of Economic Output

Alexander J. Field

Contents

Abstract

Following an introduction this chapter provides an overview of the logic and conceptual underpinnings of national income and product measures (Part I). Part II describes developments beginning in the 1930s that led to the modern approaches and conventions regarding how we should measure these aggregates. Part III reviews contributions made by quantitative economists, new economic historians, and cliometricians to our understanding of economic epochs prior to the second half of the twentieth century. The principal focus is on the United States, although there is some reference to developments in other countries.

Keywords

National income · National product · Economic growth · National income and product accounting · The United States

A. J. Field (✉)
Department of Economics,
Santa Clara University, Santa Clara, CA, USA
e-mail: afield@scu.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_38

Introduction

This chapter assesses the contributions of cliometrics and cliometricians to historical measures of economic output. The emphasis is on the United States. There has been considerable recent and exciting work on Europe, Latin America, Africa, and Asia. But we have available an excellent survey of these contributions (Bolt and van Zanden 2014), and it is unlikely that can be improved upon it at this point. Bolt and van Zanden are central participants in the Maddison Project, an attempt to keep alive and extend decades of work undertaken by Angus Maddison to develop estimates of both output and population back in time and to new areas of the world. After Maddison died in 2010, a number of his colleagues and other scholars banded together to make his estimates widely available and to provide updates as the result of new work. These data are available at Maddison Historical Statistics (2018), and the Bolt and van Zanden paper provides an accessible summary of the latest emendations, which involve the use of new methods as well as new data.

Maddison's basic approach was to start with a single modern cross-national comparison of income/output levels (his final benchmark year was 1990) and extrapolate backwards using country level estimates of rates of growth. Bolt et al. (2018) describe limitations of this method. As one gets further and further away from the benchmark, comparative levels in the past can become increasingly unreliable. These authors have thus developed two separate databases, one optimized to provide the best cross-national comparisons at different moments of time, the other optimized to provide the best growth rate measures for individual countries or regions *over* time.

Comparing output levels between countries at a moment of time (cross-nationally) can be challenging. Simply using exchange rates to convert output in the two countries may mislead. Valuing the outputs by a common set of prices can be better. But which prices? Comparing the product of a less developed country with one that is more developed using the less developed country's prices will generally exaggerate the income differences between them, whereas using the developed country's prices may make the gap seem unrealistically small. The trade-offs reflected in developing these new approaches to Maddison's legacy are a reminder of the challenges faced in generating historical measures of economic output, both cross-nationally and over time.

The focus of this chapter will be on developments within the United States. It was in the United States that the modern framework for national income and product accounting originated, although its progenitors can be found in a number of countries. The architecture of this system was to a considerable degree the work of Simon Kuznets, an immigrant from the Soviet Union. Kuznets won the Nobel Prize in Economics in 1971 for his efforts. He died in 1985.

In defining this chapter's scope, it matters how broad a net we cast in terms of who are to be considered cliometricians. Must the individual, for example, be or have been a card-carrying member of the Economic History Association or the Cliometrics Society? Cliometrics can be understood most generally as the application of statistical data and methods to historical studies of growth and

development. The term is also sometimes used synonymously with the new economic history, an intellectual movement that began in the late 1950s and flourished during the 1960s. New economic historians were certainly cliometricians, but they went beyond simply the examination of quantitative as well as qualitative data, also typically applying sophisticated econometric methods to such data, allowing hypotheses motivated by economic theory to be tested and model parameters estimated. Notable questions explored in the US context included the profitability and efficiency of southern slave agriculture and the contribution to US economic growth of the steam railroad (this involved estimating the railroad's "social saving").

This chapter casts a relatively broad net in terms of who is considered a cliometrician. Thus, Simon Kuznets and his students John Kendrick and Robert Gallman are considered cliometricians. So is Angus Maddison. Each applied statistical data and methods to the study of the historical process of growth and development. So too are Paul David, Peter Lindert, Christina Romer, Jeff Williamson, and Tom Weiss, among many others.

Cliometricians played central roles in developing national income and product accounting systems, in the process judging what should and should not be included and how certain production and payment flows should be handled. From the late 1930s onward, as the logic, methods, and objectives of these systems became more widely understood and accepted, the task of measuring economic output, previously undertaken by individual scholars or, in the early twentieth century, private research organizations, was assumed by government statistical agencies. Should government economists and statisticians be considered cliometricians? I think the answer, in general, is no, because their primary concern is to illuminate current, not historical conditions. But this distinction is not without ambiguity, because byproducts of government statisticians' efforts have been data that eventually illuminate economic history.

The entire twentieth century, in particular, is now history. The databases maintained by the U.S. Department of Commerce's Bureau of Economic Analysis (BEA), which contain quarterly macro data going back to 1947, annual data series extending back to 1929, fixed capital stock data back to 1925, and investment flow data back to 1901, greatly facilitate exploration of this period. A primary concern of policy makers is indeed to produce accurate and timely data, but they also have a subsidiary interest in history. Historical macro data is used to construct econometric models that may be used for forecasting or policy simulations. In some cases, interest in history goes beyond that, as policy makers try to reason by analogy in comparing current challenges to episodes in the past.

Nevertheless, the key driver in the transition to government responsibility was the desire of policy makers for higher frequency data available with a much shorter time lag. Government economists' and statisticians' first responsibility is to produce *current* estimates, because without such estimates, with preliminary numbers available soon after the end of a quarter, national income and product estimates are unlikely to have great value for those determining fiscal or monetary policy.

The priorities of government statisticians are thus not exactly the same as those of cliometricians. But even though refining the historical record is not a central concern of government statisticians, as a consequence of the work they do, our understanding of economic history has been and is greatly enriched, at a minimum, because contemporaneous estimates of product and income come eventually to define the (recent) historical record for us. The numbers may not be produced with the objective of helping us write economic history, but they ultimately serve this purpose. The same can be said of the decennial censuses of population in the United States, undertaken to determine representation in the lower house of Congress, as well as government censuses of agriculture, manufacturing, and other sectors. They did not have as their primary objective assisting future economic historians. But they have nevertheless done so, providing most of the raw materials from which estimates of product and income in the nineteenth and early twentieth century have been constructed.

This chapter begins with an overview of the logic and historical development of national income and product measures (part I). Part II describes the developments beginning in the 1930s that led to the modern approaches and conventions regarding how we should measure these aggregates. Part III reviews contributions made by quantitative economists, new economic historians, or cliometricians to our understanding of economic epochs in the United States prior to the second half of the twentieth century. Concluding section follows.

Part I: The Logic and Early History of National Income and Output Estimation

The best overview of the early history of national income and product accounting is still to be found in Kendrick (1970). Kendrick, drawing heavily on Studenski's (1958) book-length compendium of efforts in different countries, divided the intellectual history into two periods, with World War I the demarcation line. In the earlier (and much longer) period, estimates were constructed almost entirely by individuals and were limited to a few relatively advanced economies. With developing agreement on concepts and methods, and improved statistical data, responsibility was eventually shifted to teams of government statisticians and expanded to many other countries. After the Second World War, the United Nations, building on League of Nations efforts, played an important role in standardizing and diffusing these systems across the globe.

In surveying the history of national accounting systems, Kendrick, following Studenski, distinguished two main flavors, material product and comprehensive. Ultimately, it is the latter that has become the internationally accepted standard, but the former, which focused (as the category name would suggest) on tangible output, informed the statistical systems of the USSR and other COMECON countries throughout much of the twentieth century. Comprehensive frameworks aimed (and aim) to include not just physical goods, such as business equipment and structures and consumer durables and nondurables, but also services delivered to

final consumers. Examples of the latter include shelter (housing) services, personal care, and legal, educational, and medical services. Technically speaking, both transportation and the electricity producing sectors, for example, should also be considered part of services production, as are wholesale and retail distribution, finance, insurance and real estate, and communication. None of these sectors produces tangible goods. The COMECON countries included services that contributed to final goods production, but not those consumed directly by households: thus freight but not passenger transport, communication services purchased by firms but not individuals, energy consumed in the production of physical goods but not by households, etc.

Modern comprehensive national income and product accounting systems approach the task of measuring output flow using three distinct approaches. The first involves aggregating value added by individual economic units. Value added is defined as gross sales less purchased materials and services (what is subtracted is referred to by the Bureau of Economic Analysis (2017) as "intermediate purchases"). Intermediate purchases include raw or semi-processed inventories as well as fuel or energy inputs. They also include labor services provided by outside contractors or businesses. In considering the deduction for services bought, it is important to distinguish between those provided by employees of the organization itself and those purchased or rented from other individuals or organizations. Only the latter are to be deducted from gross receipts. Thus legal services purchased from an outside law firm would be deducted from gross sales in calculating value added whereas those provided by in-house counsel would not.

If one calculates value added in this fashion for every unit engaged in producing goods and services for the market, and then aggregates, one will have an approximate measure of gross domestic output. It is approximate because a few imputations for nonmarketed services, such as the housing services produced in owner-occupied residences, are, by convention, added to this. It is gross because it includes that portion of private investment necessary to maintain the physical capital stock against the ravages of wear and tear and other forms of depreciation. It is domestic because it calculates value added irrespective of the nationality of the owners of the factors of production. National, or citizenship measures, such as gross national product, exclude value added attributable to foreign-owned factors of production but include value added by factors of production, both labor and capital, owned by nationals but situated outside of the country.

This value added approach is sometimes called the production method of measuring output. It measures value added as it is generated by corporations, partnerships, and individual proprietorships. Economic organizations add value to purchased materials and services by combining these with the services provided by the organization's employees and owned physical capital. The goods or services produced are sold forward, ultimately (perhaps after further transformation or change in physical location) reaching a final consumer.

With proper aggregation, the approach yields data on the respective shares of different sectors (e.g., manufacturing, transportation, agriculture) in gross value added or output.

Kendrick and others call this, perhaps confusingly, the "income originating method," because it is out of the flow of value added at each stage of production and distribution that organizations generate flows of income to the labor they employ as well as the households that own the economic entity, and thus its capital assets. The identity of gross income originating with value added for each economic unit guarantees the equality between aggregate measures of gross domestic product (GDP) and gross domestic income (GDI). An economic unit makes wage and salary payments to the labor it employs. If a corporation, the entity will, on behalf of its owners, hold title to buildings, equipment, stocks of inventories, and other nonhuman assets such as patents or trademarks. In addition to the wages and salaries it pays out, it will ultimately generate income flows to the owners of these assets.

Subtracting wage and salary payments and indirect business taxes (taxes on production and imports less subsidies) from value added yields what the BEA calls gross operating surplus. Deducting corporate income tax and capital consumption (depreciation allowances) takes one to net after tax income. Deducting net interest payments arising from debt finance yields net after tax corporate profits, which will be distributed to owners of the corporation as dividends or retained as net saving done on behalf of the households that own the company. Unincorporated businesses/individual proprietors make periodic withdrawals from business accounts to owners' accounts, distributions which reflect compensation both for the labor services provided to the business and a return to the capital invested.

Thus, as economic units produce goods and services reflected in and measured as value added, they simultaneously generate income flows to the owners of factors of production. Measuring value added as it is paid out to owners of factors of production is sometimes called the income method, or as Kendrick and others put it, the "factor income" method. This second method, which underlies the construction of "social tables" described below, requires measuring and then aggregating the income flows generated by each organization as they are actually received by households. This approach to tracking value added produces aggregates for wage and salary income, including pension and health insurance contributions, as well as the employer portion of social insurance payments. It also includes payments to nonlabor factors or production, income to capital (and a small amount to land). Value added less employee compensation yields the gross flow of income to capital. With aggregation and deduction for capital consumption, this second approach allows the calculation of the share of labor or capital in national income.

The BEA defines gross operating surplus as value added less employee compensation less taxes on production and imports (see also Sutch 2006). These indirect business taxes are taxes not levied directly on corporate or personal income. They include excise taxes, import duties, state and local sales taxes, and local property (real estate) taxes. Subsidies, such as those paid to farmers or to some public housing authorities, are subtracted from indirect business taxes: in a sense, they are the opposite of indirect business taxes. Why? Taxes are payments made to government for which no directly identifiable excludable good or service in the current period is provided in return. Subsidies or other transfers are the opposite, for they are payments from the government to households for which no corresponding good or service is tendered, at least during the current period.

Subtracting capital consumption allowances from gross operating surplus leaves net operating surplus. Subtracting the costs of debt finance (net interest payments) yields corporate profits, which ultimately flow into one of three bins: corporate income tax payments, dividends to equity holders, and retained net business earnings, which represent corporate saving on behalf of the households that own the corporation. Rent paid to real estate corporations shows up as part of the gross operating surplus of such corporations. The remainder flows directly to households as rental income of persons. Income to unincorporated businesses, which represents a return to both labor and invested capital, is listed separately in the accounts as proprietors' income. At least for the modern period, product measured from the output or expenditure side is generally considered more accurate than that from the income side. The totals developed using the income method tend to come in slightly below those reached using the product or expenditure methods (see below); the difference is treated as a statistical discrepancy.

National income is simply gross national income less consumption of fixed capital, or gross domestic income plus net factor income from abroad less consumption of fixed capital. (Note: Contrary to what is stated in many textbooks, taxes on production and imports are now included in national income, rather than part of the wedge separating national income from gross national income. See BEA NIPA Table 1.7.5, for example.) It should be noted from the previous discussion that not all gross income flows immediately and directly to households, or is available to them for consumption, or is actually consumed. Taxes leak out, and these include indirect business taxes less subsidies, corporate income tax payments, and personal income tax or payroll tax flows. Saving is another leakage and includes firm depreciation allowances, a component of national (gross) saving. Nor will all after-tax corporate profits necessarily be distributed to households as dividends. Some may be retained as net business saving, a component, along with personal saving, of private saving. Not all of a business unit's gross interest payments will necessarily find their way to households, since businesses borrow and lend amongst each other.

Personal income measures what actually flows to households. Like national income, it includes all labor compensation, all proprietors' income, and all rental income of persons. It also includes personal income receipts on assets, which is the form 1099 dividend and interest income actually received by households, and personal current transfer receipts, which consist mostly of government transfer payments (Social Security and Medicare benefits, interest on the national debt, etc.). Thus, a substantial portion of the payroll taxes from the household sector to the government are remitted to the (consolidated) household sector as social insurance transfer payments. Similarly, a portion of the remaining gross government tax receipts (corporate and personal income taxes, etc.) is remitted to the consolidated household sector as interest on the national debt. In both cases, the government effects interhousehold transfers: from those working to those aged or retired, and (largely) from those working to bondholders. Unlike national income, personal income does not include corporate profits, indirect business taxes, payroll taxes for government social insurance programs, except those distributed as dividends, and a couple of other minor items.

Personal income ultimately resolves itself into one of three bins: personal income taxes, consumption, or personal saving.

The third method of reckoning output is the expenditure approach. It is not obvious why gross expenditure should necessarily be equal to gross domestic income, given the leakages into taxes and saving described above. There is moreover an additional leakage: some spending will be on goods and services produced outside of the country (imports).

The reason GDE (gross domestic expenditure), if measured correctly, should nonetheless be equal to GDI and GDP, subject to small statistical discrepancies, is that there are, as pioneers of national income accounting came to understand, three sources of spending that don't emanate directly from households. These are traditionally considered injections, and the sum of these injections should just match the sum of the leakages. Why? Each leakage category has a corresponding injection category. In the case of taxes, it is government spending on goods and services. In the case of imports, it is exports, and in the case of saving, it is investment (understood in the macroeconomic sense as acquisition of new structures, equipment or net accumulation of inventories). Nevertheless, there is no guarantee that each of these pairs will balance; indeed it is likely they will not. There can, for example, be a government deficit or surplus or a current account deficit or surplus, and private saving might fall short of or exceed private investment. But the sum of the injections must just equal or balance the sum of the leakages.

How do we know this? Because the gross income flowing to individuals must ultimately resolve itself into one of three bins: consumption, saving, or net taxes: $Y = C + S + T$. And on the output side, breaking down the aggregate by type of expenditure, we know that the total will equal the sum of consumption, investment, government spending, and net exports: $Y = C + I + G + X - M$. Setting the two right hand sides equal to each other, and subtracting C from each side, we have $S + T = I + G + X - M$. Rearranging by adding imports to both sides, we have $S + T + M = G + I + X$ or, in a useful rearrangement, $S = I + G - T + X - M$: private domestic saving must finance the sum of gross private domestic investment, the government deficit, and the current account surplus, which represents the net acquisition of foreign assets.

We can all apparently breathe a sigh of relief. In what seems like an affirmation of Say's Law, supply does indeed seem to create its own demand.

Not so fast, however. It turns out that there is a bit of a trick involved in guaranteeing this balance, and that is to consider any net acquisition of inventories, whether intended or otherwise, as part of gross expenditure as well as gross private domestic investment. Firms are viewed as purchasing additional inventories on their own account, whether or not their acquisition was planned. During the 1930s, the decade during which Kuznets was developing the logic of the national income and product accounting system, Keynes was working on the *General Theory*. By distinguishing between inventories voluntarily or involuntarily acquired, and assuming some sluggishness in price adjustment, one could develop a coherent explanation of how sizable output gaps (a difference between actual and potential output) might persist. Most nineteenth century and earlier economists were not

excessively concerned with output gaps (Malthus was an exception) and whether or not there might be a deficiency of aggregate demand predisposing toward recession, even though at every moment of time expenditure would match output and income. We cannot claim that they understood the detailed logic of why the three different national income and product accounting approaches should sum to the same annual magnitudes. Precursors to moderns did, however, stumble upon each of the three estimating approaches used by modern national income and product accountants.

The expenditure approach measures final expenditure flows, including spending by households on goods and services (personal consumption expenditure, or PCE), by businesses on plant and equipment plus any inventory investment (gross private domestic investment, or GPDI), by governments on goods and services (G), and by foreigners on a country's exports less what is spent on imports (net exports). This approach facilitates the calculation, for example, of the share of output appropriated by government at different levels, or the share of consumption in gross domestic product.

Part II: Historical Antecedents

Kendrick makes the point that antecedents for each of these three approaches (output, income, and expenditure) can be found in the pre-World War I period. The three approaches were refined during the second quarter of the twentieth century. In the process, understanding of their interrelationship improved, permitting a coherent explanation of why, in principle, each should total to the same magnitude. To review: why should total income, for example, necessarily equal total output? Because for each economic organization, it is out of the flow of value added that gross income flows to owners of both labor and capital originate. If income generated equals value added for each economic unit, then in principle it should be true in the aggregate. Why should aggregate expenditure (including any change in inventories, considered spending by the firm that accumulates them) equal total output, given the leakages out of gross income that flow to taxes, saving, and imports? Because the sum of these leakages will be matched by the sum of three categories of spending that don't originate directly in domestic households: spending by governments on goods and services, investment spending by businesses, and exports.

Those studying economic output prior to the twentieth century did not understand all of this, but they understood enough, at least intuitively, to go about making estimates of output and income. Kendrick, following Studenski, counts 13 countries for which measures of economic output had been calculated prior to 1920. Seventeenth century pioneers in England, especially William Petty and Gregory King, approached the problem from the income side, constructing social tables and using an approach that has come to be known as *Political Arithmetic*, and which we will see resurrected in work by Lindert and Williamson (2016). King and Petty's work provides a useful contrast with much of the work on the nineteenth century in the United States, which is built up from the production side.

Petty (1691) published estimates for Britain in 1665 of what we would now call national income. He estimated income from land at £8 million, and from other personal estate at £7 million, with total income, based on an informed guess about population and average income, equaling £40 million. With the "Annual Proceed of the Stock or Wealth of the Nation," consisting of rent, interest, and profits, at £15 million, he attributed the residual of £25 million to the "Annual Value of the Labor of the People."

Several decades later (1696), Gregory King calculated national income by dividing the population into 26 occupations or classes, estimating the number of families in each grouping, the average number of persons in each type of family, and the average per capita income for each type of household. He multiplied within each category, aggregated, and then added in an estimate of the revenues of the Crown. He did this first for 1688, and, based on estimates of spending, was able to calculate the gap between income and consumption, which yielded saving or capital accumulation.

He subsequently constructed a time series through 1698, using it to cast light on the capital consumption engendered by military conflicts with France. He used similar techniques to estimate income for France and Holland in 1688 and 1695, and the relative burden of the wars in each of the three countries. Unfortunately, his work, although circulated privately at the end of the seventeenth century, and utilized by Adam Smith in the *Wealth of Nations* (1937), was not available publicly until 1802. To the degree that their income estimates were proxies for output, neither Petty nor King denigrated the output of final services. They can thus be considered as falling into the "comprehensive" camp in terms of their conception of output.

The same cannot be said of the French Physiocrats (Quesnay 1972; Gide 1948), who believed that only agriculture was capable of producing a net product. The Physiocrats reflect a unique and unusual variant of the material product approach, focusing on income and product within agriculture, which was of course the bulk of the French economy. Since they studied intersectoral flows, the Physiocrats can be interpreted as foreshadowing Leontief's work on input-output matrices. They also contributed an understanding of the necessity of reserving a portion of a country's gross product (in their case, grain) for the replenishment, and ultimate expansion of a country's physical capital stock. That understanding would ultimately be reflected in the distinction between gross and net product or income.

Adam Smith should also be placed in the material product camp. Acknowledging his views, in particular his distinction between productive and unproductive labor, also helps us understand why Karl Marx was a classical economist, and how Smith, in spite of his opposition to Mercantilism and association with laissez faire economics, can be said to have foreshadowed the accounting system of the former Soviet Union more so than that of the United States. Smith believed that productive labor fixed itself in vendible material products, whether structures, equipment, consumer durables or nondurables. Labor generating services for final consumers was for Smith unproductive, a view that was shared more or less by Ricardo and Mill and eventually by Marx.

One might try to explain Smith's biases as based on his desire to deepen capital (increase the physical capital to labor ratio) through saving and accumulation, combined with the obvious fact that only goods can be accumulated. Services, as he noted, perish in the instant of their own creation. It does not follow, however, that favoring goods production at the expense of services will necessarily result in higher levels of capital accumulation (and, presumably, welfare), since goods can also be consumed.

Smith is revered today for his rejection of Mercantilist views. Mercantilism stressed that a country desirous of aggrandizing its national power should aim to accumulate precious metals by running export surpluses. Smith, in contrast, argued that the strength of a nation's economy should be measured not by its holdings of precious metals but by its stocks of productive resources: labor, capital, and land, and the flows of output and income they enable. We should also, however, acknowledge that Smith's emphasis on the distinction between productive and unproductive labor, a leitmotiv of classical economics up through and including Ricardo, Mill, and Marx, represented a detour, with detrimental consequences for countries committed to it. Thus, for example, the material product system of national accounting can be plausibly blamed in part for stinting wholesale and retail distribution sectors within socialist economies, resulting in enormous waste.

Marx built on Smith's distinction between productive and unproductive labor, and also used it as a vehicle for focusing on the desirability of accumulating physical capital. Among the classical economists he had distinctive views about the objective of production in a capitalist economy, which was, he argued, to produce a net income (or surplus value) only for the capitalist class. In his view, nonwage income (rent, profits, dividends, interest) did not represent a legitimate return for the employment of a productive factor, but the extraction of surplus value made possible via the employment of wage labor. If Ricardo viewed landlords as living at the expense of the rest of society, for Marx it was industrial capitalists who were the principal parasites.

Marshall represented the decisive turning away from variants of the material product approach in Anglo-American economic thought, refocusing on the comprehensive approach to product and income accounting reflected in Petty and King, and emphasizing that the objective of an economic system was ultimately to satisfy household wants, and that the accumulation of physical capital should be understood only as a means to that end.

Part III: US Estimates of Output and Income Prior to the Second Half of the Twentieth Century

Samuel Blodget (1806) is generally credited with producing the first estimate of US national income. Using methods echoing those employed more than a century earlier by Gregory King, and more recently by Lindert and Williamson, he divided the US employed persons into classes, in his case seven, estimated annual per capita income for each, multiplied, and aggregated (see Blodget 1806; Rhode and Sutch 2006).

The 1840 US census was the first to ask a series of detailed questions providing a more solid basis for income and output estimation from the production side. George Tucker (1843) used that census to estimate commodity output in the aggregate and by state. He updated his work in 1855 based on data from the 1850 census. Ezra Seaman (1868) published similar estimates based on the 1840 and 1850 censuses. For the years 1880 and 1890, Charles Spahr (1896) developed more comprehensive estimates, and included calculations of the size distribution of incomes in the 2 years, which he used to draw conclusions about inequality trends. Wilford King (1915), using Spahr's framework, extended estimates to 1910. He concluded that labor's share of income was rising, leading him to doubt that inequality was increasing, although the former is not necessarily evidence against the latter proposition.

Simon Kuznets' work during the early 1930s helped define the architecture of a more fully developed, logically consistent accounting system for national output and income. Not all of the decisions he made about what to include and what to exclude, however, were ultimately incorporated into what became the now internationally accepted approach.

Here is the history. As the Depression worsened in 1931, government economists complained that estimates of output and income from the National Bureau of Economic Research appeared many months or years after the fact and were consequently of little use for business cycle forecasting or policy analysis. The National Industrial Conference Board published somewhat more timely estimates, prepared under the direction of Robert F. Martin, but they still were not available soon enough to be of real assistance to policy makers.

In February of 1932, officials in the Department of Commerce's Bureau of Foreign and Domestic Commerce joined forces with individuals working with Senator Robert LaFollette, a progressive senator from Wisconsin. These discussions resulted, on June 8, 1932, in the introduction of a Senate resolution calling for the production of annual estimates of national product and income for the years 1929, 1930, and 1931. The work was to be conducted within the Bureau of Foreign and Domestic Commerce and was begun under the direction of J. Frederick Dewhurst. When, by November of 1932, it became clear that Dewhurst and his limited staff would not be able to carry forward the project, an agreement was reached with the NBER for Kuznets to take over. Kuznets had been working on estimates of national income for the NBER since 1929 and was in the process of developing improved procedures more explicit about definitions and more careful about citing original data sources. He transitioned to the government in January of 1933, and on January 4, 1934, roughly a year later, delivered his report to the Senate. *National Income, 1929–1932* detailed two measures of national income, one excluding and one including retained business earnings. Within 8 months, approximately 4,500 copies had been sold (Carson 1975, p. 159). For a Senate document, it was a bestseller. Almost immediately, the Commerce Department took steps to assume responsibility for maintaining these estimates on an ongoing basis.

In these estimates, Kuznets excluded some categories of income that King had included. Among these were service flows from consumer durables, the value of

services provided within the household economy, earnings from illegal employment or the informal economy, capital gains, and relief and charity payments. For the first three of these, estimation was difficult and including them risked introducing a good deal of noise. Leaving them out might provide a less conceptually satisfying measure of levels but enable a more reliable calculation of growth rates.

Nonmarket services provided within the context of the household economy services were one of these categories, although Kuznets noted that in the circumstances of the Great Depression, their exclusion did not necessarily enable changes in national income to be a better proxy for changes in welfare. The narrower measure he provided, which excluded services provided within the household, fell sharply between 1929 and 1932, as the unmeasured household sector grew to take up "some of the slack imposed by the shrinkage of the market economy" (Kuznets 1934, p. 4). So the measures Kuznets reported suggested a decline in welfare more extreme than what probably occurred. Although the estimates in his report spanned only a 4-year period, Kuznets was also concerned that the practice of excluding the value of household production could bias the welfare implications of long run increases in per capita output, since the importance of such production tends to decline with the process of economic development. For the same reasons, cross-national comparisons of regions at different stages of development might be jeopardized, to the degree that output per capita measures were interpreted as proxies for welfare.

Kuznets' reasons for excluding capital gains were varied. On the one hand, he argued that including them would represent double counting, since it would reflect "both changes in national income and its capitalization." He also says it would "distort" the calculation of national income (1934, p. 5). The most compelling argument for excluding capital gains is that they do not represent value added as a result of current period production. To include them as part of income disrupts the posited equality between aggregate income and product.

In common with his treatment of the service flow from consumer durables, Kuznets also excluded the value of owner occupied real estate. His view was that "... there is some doubt as to the propriety of including this item, since the ownership of a home combined with its possession does not constitute a participation by the proprietor in the economic activity of the nation in the same recognized fashion as does his work for wages, profit or salary" (1934, p. 12) Ultimately, the Department of Commerce thought otherwise, treating homeowners as in the business of producing housing rental services, whether they chose to consume them or not. Accepted practice today is to make an imputation for the service flow from owner occupied housing, incrementing the aggregates for product, income, and expenditure. An argument for doing this is that it would make little sense for the GDP growth rate to change simply because of a change in the housing tenure rate, which could be the consequence of accepting Kuznets' approach.

His largest deviation from current practice was to treat government spending on goods and service as intermediate goods, arguing that it would be double counting to include in measures of output and income both government spending and the private sector final output it facilitated. Gilbert et al. (1948, pp. 182–183) pointed out, however, that Kuznets was using the concept of intermediate goods

quite differently when applying it to government public goods like national defense, justice systems, fire and police protection than was the case when it was applied to the purchased materials and (nonwage) services deducted from an entity's gross sales to obtain value added. In the latter case, payments to suppliers were associated with the delivery of specific goods or services. To the degree that taxes support a judicial system, for example, it seems a stretch to call this a fee for service, a point that takes on particular salience in the event we end up being prosecuted.

Kuznets, however, saw little difference, but on this he was opposed by many economists, and the conventions adopted by the Department of Commerce reflected these objections. Goods and services produced by government directly, as well as government purchases of goods and services, were all considered part of final output, and after 1942 were included in measures of gross national product and expenditure as well as income.

Kuznets' position was consistent with his continuing concern that we not assume that national output and income measures were necessarily good proxies for welfare. In particular, he did not see trillions spent on the military as satisfying human needs as directly or in the same way as spending on (and production) of food, clothing, or shelter. Similarly, and perhaps less controversially, he observed that "occupational expenses" such as commuting, although counted as part of consumption, really reflected an intermediate input into the production of goods and income that did satisfy human needs. The commuting itself did not directly contribute to welfare but only indirectly through its facilitation of our ability to earn income.

Kuznets stressed several times that since value added was measured at market prices, the aggregate depended not only on the vector of output but also on relative prices, which could be influenced by the distribution of income. For that reason as well, he said, per capita output should not be interpreted as a measure of welfare. All of these concerns are well and good and certainly valid. But it was probably unreasonable to expect economists and others to refrain from calculating per capita output over time or across different countries and make inferences from these numbers about material welfare.

Kuznets' point about the possible influence of the distribution of income on relative prices, and therefore on the value of production aggregates, received little emphasis in subsequent decades. Perhaps economists believe the impact of variation in income distribution on market prices to be relatively minor. On the other hand, it is often noted that, from the standpoint of assessing how well a society meets human needs, it matters (on the income side) how unequally that product is distributed among households. For example, adult height, a reflection of consumption/nutrition through adolescence, is strongly correlated cross-nationally and over time with the log of per capita income (Steckel 1995). But for a given per capita income, a higher Gini coefficient (a measure of inequality) is associated with lower average height. Kuznets' concern is related but not exactly the same, since it involves the impact of inequality on the estimation of the aggregate itself, which appears in the numerator of an output per capita measure.

Finally, Kuznets' estimates were of nominal income. In the 1934 Senate report, he made no attempt to convert nominal estimates to measures of real output

and income, attributing his reticence to the absence of an appropriate deflator that would cover both goods *and* services purchased by households. He called attention to the poor data for household expenditures on services but did not stress a need for a broader deflator that would also cover investment goods (structures and equipment) and government goods, which as noted, he chose to treat as intermediate. Let's assume a CPI or PCE deflator is the right price index with which to deflate consumption spending, which bears the most direct connection to human welfare, and that we were to follow Kuznets in treating government purchases as intermediate goods. In measuring changes in real output, we would at least for some purposes still want to use a broader deflator that also covered the roughly one-sixth of output that typically goes to producing investment goods (structures and equipment). Robert F. Martin's 1939 publication included estimates of real output going back to 1799 using either the CPI or a broader price index as a deflator (1939, Table 1, pp. 6–7). By 1937, as evident in *National Income and Capital Formation*, Kuznets was including estimates of real output in 1929 prices.

Interestingly, the term gross national product appears to have originated with Clark Warburton (1934), not Kuznets. The main difference with national income is that the gross measure included that portion of output devoted to maintaining the physical capital stock and thus compensating for depreciation. It is often argued that the preference for gross rather than net measures allows a more reliable estimate of growth rates, because calculating economic depreciation is difficult and often somewhat arbitrary – as much art as science. By 1937, Kuznets had adopted Warburton's approach and terminology (Kuznets 1937).

Kuznets' 1934 Senate report, although not the final word on procedures for calculating national income and product, was a milestone in their development. It also has some value for historical research, since it included a variety of observations about the differential impact on the income side of the worst years of the Depression. He found, for example, that income to property holders (in the form of interest and dividends) held up much better than labor income or income to entrepreneurs (p. 14), and that, although both income to capital and income to labor declined, labor's share dropped (p. 41). Over and over again, he reported that, for those industries where this could be distinguished, salary incomes between 1929 and 1932 declined less in percentage terms than did the income of those receiving wage incomes. Entrepreneurial income (what we call today income of proprietors) dropped sharply because it was heavily dominated by farmers and those working in construction. The main cause of the decline in proprietors' income was the drop in grain prices (p. 49), although surely the collapse in construction spending didn't help. The effects on income were aggravated because neither group tended to exit when incomes dropped (p. 33).

Some occupations or sectors did very well during the Depression in terms of real incomes. If you kept your job in government at any level between 1929 and 1932, your real income went up, as did those working in private higher education. Between 1929 and 1932, the number employed in private higher education as well as their per capita compensation went up quite substantially in real terms (p. 148). More generally, if you were a salaried worker and kept your job in the

Depression, your standard of living improved. Overall, Kuznets's report documented the reality that those on the lower rungs of the income distribution suffered disproportionately: "The Depression seems to have put its greatest burden upon those who, in view of their already low position on the economic scale, could least afford to lose" (p. 19).

Other miscellaneous notes: Mining, manufacturing, and construction suffered the greatest employment losses (p. 23). Finance had the highest average per capita compensation (p. 28). Technical change in steam railroads was "a thing of the past" (pp. 86–87). Motor transport (trucking) felt the effects of the depression much less severely than steam railroads. Interest payments declined hardly at all, in contrast to dividends. Defaults on mortgages (and cessation of interest payments) were much larger than defaults on corporate debt (p. 120).

Finally, Kuznets included an interesting discussion of cyclical effects on productivity within manufacturing. In food and tobacco, output per worker continued to increase between 1929 and 1932. The same was true in chemicals and petroleum refining through 1931 (p. 72). In most other industries, particularly industries where the quantity of output fell absolutely, labor productivity also declined, perhaps reflecting the impact of initial labor hoarding.

Kuznets followed his Senate report and 1937 publication with work extending his estimates first to 1919 (*National Income and its Composition*, published in 1941), and ultimately to 1869 (*National Product Since 1869*, published in 1946). As scholars worked out the logic and debated the conventions of national income and product accounting, they also began to work to extend their more systematized understanding of methods to construct income and output aggregates in earlier periods. In the process, they both critiqued and built upon earlier efforts. Because the 1840 and subsequent US censuses enabled calculations of value added for the material product portion of output (commodity output), the estimating technique for someone interested in a comprehensive measure of output required making informed guesses based on data from later periods about the ratio of services output to goods output and using this to "mark up" the commodity totals.

This was the basic approach followed by Kuznets in constructing annual estimates back to 1869. His pre-1919 estimates were based on commodity (goods) data from Shaw (1947), which Kuznets marked up by assumed margins for transportation and distribution. The size of those margins, and more generally the question of whether service sector (noncommodity) output varied roughly one for one with goods output, lay at the heart of Christina Romer's subsequent questioning of how much more moderate had been post-World War II business cycles relative to those that preceded them (Romer 1989, 1994).

For 1919–1929, however, Kuznets was able to construct his estimates from the income side, rather than the incomplete data on the product side, and his estimates for these years are generally assumed to be of higher quality than the pre-1919 estimates. Romer also endorsed income side estimates for 1909–1918 contained in an appendix to Kuznets (1961), although Kuznets believed these numbers to be less accurate than those for 1919–1929, and ultimately judged his product side estimates for 1909–1918 to be superior (see Weir 1986, p. 355).

John Kendrick (1961) made adjustments to the treatment of government expenditure in Kuznets' GNP series to make them more comparable to the annual series from 1929 onward maintained by the BEA. As noted, Kuznets' practice had been to treat government expenditure as an intermediate good, not part of final product. Kendrick's have since become the standard series referenced by students of the 1920s. Romer's annual GNP series for 1919–1929 are Kuznets' income side estimates with Kendrick's adjustments and some other minor adjustments.

Most of Romer's revisionism applies to the pre-1909 data. Her big differences with Kuznets involved how much pre-1909 GNP was likely to have varied with changes in commodity output. Kuznets used freehand regression to estimate the elasticity of GNP with respect to goods production using data for the years 1909–1938 (Kuznets 1961, pp. 536–37). Based on data from these years, he concluded that the elasticity approached one and used this to backcast a GNP series using goods data for the earlier years. Romer reestimated the elasticity using data from 1909 to 1985 but excluding the years of the Great Depression and the Second World War (1929–1946). She made a number of other minor changes – using log differences rather than ratios, allowing the elasticity to vary over time, and using what she described as "normal" rather than peak years to establish trend from which deviations might be calculated.

Based on these regressions, she found that the elasticity was not in fact very time sensitive: "the time varying coefficient measuring the sensitivity of GNP to commodity output fell from 0.583 in 1909 to 0.527 in 1985" (Romer 1989, p. 20). The more important aspect of her results was not the small difference in these two numbers but rather their moderate size. Kuznets had concluded that GNP varied almost one for one with commodity output. Romer argued that the elasticity was closer to 0.5 or 0.6. Thus her estimates of pre-1909 GNP are much less volatile than Kuznets', which is what underpins her conclusion that the pre-World War I business cycle was not markedly more severe than the post-World War II cycle.

Romer justified excluding 1929–1946 from her regression on the grounds that we could expect the elasticity of GNP with respect to commodity output to have been unusually large during these years because the fluctuations in both series were so substantial. She maintained that it would be inappropriate to extrapolate backwards from this "abnormal" period to more "normal" years between 1869 and 1908, and she attributed Kuznets' high elasticity estimate in part to his inclusion of the years 1929–1938 in his estimating regression. Weir (1986, p. 355) questioned Romer's exclusion of 1929–1946, arguing that there was little evidence of a structural break during the depression years. Nevertheless, Romer's argument about the relative severity of the pre-World War I business cycle, and her means of reaching that conclusion, have subsequently been widely accepted.

It should be kept in mind that Kuznets was adamant that although he thought his annual estimates back to 1869 were useful for calculating trend growth rates, he did not think they were sufficiently accurate to form the basis for the exploration of cyclical variation (Kuznets 1961; Rhode and Sutch 2006). Robert Gallman, Kuznets' student, who extended estimates to 1834, took the same position (Gallman 2000). Kuznets, along with Gallman, also did pioneering work estimating the growth

of the US physical capital stock. Gallman, extended both series back to 1834, continuing Kuznets' tradition of careful attention to detail, crosschecking of calculations, and documentation of sources and data transformations. With respect to the capital stock, Gallman emphasized the dominance of structures in both the capital stock and net investment flows, a theme emphasized in Field (1985). Gallman also considered investment in land clearing as equivalent to the creation of a reproducible tangible asset, thus a capital good, and emphasized how empirically important it was in the nineteenth century.

For the 1909 period and earlier, Kuznets and Gallman did not have access to better data than did Tucker, Seaman, Spahr, or Wilfred King. Nor, for the most part, could they use electronic spreadsheets or other data processing conveniences, which might have made their job easier. What they did have was a more solid understanding of the logic of national income and product accounting, an essential starting point for those wishing to do research in this area.

Extending the aggregates much before the 1840s, however, continued to strike these scholars as daunting. The first serious effort to venture further back in time is reflected in estimates published by Robert F. Martin in 1939. Martin provided decadal numbers starting in 1799 and then annual data from 1900 through 1938. He concluded that real per capita income had fallen during the first three decades of the nineteenth century, beginning to rise again only in the 1840s (Martin 1939, Table 3, pp. 14–15). Because of the absence of reliable production data for commodity output prior to the 1840 census, the years before this are commonly referred to as a statistical dark age.

Using an approach suggested by Kuznets (1952), who had been critical of Martin's estimates, Paul David (1967) attempted to shed light through what he described as controlled conjectures. He took aim at what he perceived as a consensus, based on Rostow's work (1960) and apparently supported by Martin's estimates, that there had been a marked acceleration in the rate of growth in output per head sometimes between 1800 and 1840. This would have been consistent with Rostow's claim that a takeoff typically accompanied entry into sustained economic growth. David pushed for a more gradualist perspective on the entire 1790–1860 period, emphasizing instead a post-Civil War acceleration associated with, among other things, a rise in the national saving rate.

He began by observing that growth in output per head would equal the sum of growth of labor force participation and growth in output per worker, and the advance in the former was relatively modest (about 0.3% per-year). On the other hand, there were, across the antebellum period, major shifts in the sectoral shares of agriculture and the nonagricultural sectors. Following Kuznets, his conjectures were driven by data from later in the nineteenth century suggesting that value added per worker in nonagricultural sectors was twice (or more) what it was in agriculture. David assumed that growth within each sector (and thus the average for the economy) could be proxied by the rate of advance within agriculture. Using these moving parts, he assembled an engine to "retrodict" growth in output per hour, and in combination with the data on modest increases in participation rates, output per head. His conclusion: growth in output per head between 1790 and

1860 at about 1.3% per year. Having begun by endorsing Kuznets' criticisms of Martin's estimates, and aiming to soften Rostow's takeoff, David's numbers nevertheless still showed acceleration between the first two decades of the century and decades three and four (0.28% per year between 1800 and 1820 vs. 2.0% between 1820 and 1840), although this was earlier than Martin or Rostow had suggested.

When incorporated into a similar Kuznetsian framework, Thomas Weiss's revisions to Lebergott's labor force data suggested somewhat higher levels of output per capita at the start of the nineteenth century, slower growth between 1820 and 1840, and consequently somewhat lower overall growth rates over the first six decades of the century (Weiss 1992, Table 1.2, p. 27). One should be careful in concluding that such revisions necessarily reflected poorer economic performance. If US residents in the first decades of the century enjoyed higher levels of output per capita than was previously suggested, this is not necessarily a less rosy picture, simply because the growth rate was lower.

In an unpublished working paper written decades after his original contribution, David articulated some second thoughts (David 2005), expressing reservations about his earlier uncritical acceptance of Kuznets' reading of relative productivity levels in the agricultural and nonagricultural sectors. David now argued that the growing share of labor outside of agriculture contributed to rising output per head not because output per hour was higher outside of agriculture, but because people worked so many more hours per year once they left agriculture. He still pressed for a more gradualist reading of trends in output per head in the first six decades of the nineteenth century than had been suggested by Martin or Rostow but was now using a much modified retrodictive engine to arrive at these results. If David's revisionism about relative productivity levels are to be taken seriously, they pose issues for much other work.

The seventeenth and eighteenth centuries, which include the colonial and revolutionary periods in US economic history, have in recent decades attracted a remarkably wide range of inquiries aimed at estimating levels and growth rates of output and/or the standard of living across this long period. Both economists and historians have contributed. The energy and range of interest in these explorations may partly reflect the relative paucity of data, and thus the premium placed on creative inferences from that which is available. In the absence of comprehensive data on commodity output, the starting point for most nineteenth century estimates, scholars have used different means to extrapolate output per capita levels and rates of growth. Alice Hanson Jones, for example, whose research had built up estimates of wealth using probate records, assumed a wealth to income ratio to estimate income (1980). Others have tried to infer income or product from data on imports (for example, Egnal 1998, but see Mancall and Weiss 1999). Steckel (1995, 2006) drew inferences about consumption levels from height data.

Rosenbloom and Weiss (2014) provide a useful overview of research on the colonial period and the different data sources and methods of drawing inference from them, along with a comprehensive set of references. Their main agenda, however, is to estimate product per capita and its growth in the Mid-Atlantic Colonies (Pennsylvania, New Jersey, New York, and Delaware), using the

framework that David and Kuznets exploited: inferences about the respective growth rates of output per person in agriculture and nonagriculture, combined with estimates of shifts between the sectors and changes in labor force participation rates. Their research is echoed in the Lindert-Williamson conclusion that a significant retrogression in economic growth took place during the revolutionary and Articles of Confederation periods (1775–1790) (see below). It is also notable for acknowledging the important role the service flow from the housing capital stock (whether owned or rented) makes to aggregate output, and thus the contribution of the accumulation of residential housing to increased actual and potential output. This treatment is a valuable counter to those who dismiss residential capital as nonproductive. Such capital is indeed unusual in comparison to that employed in agriculture, manufacturing, or transportation because it contributes to aggregate product largely without the cooperation of labor. But historically, as is true today, it makes a significant contribution to output and consumption. They conclude that growth in the Atlantic colonies between 1720 and 1800 was modest. Using a similar approach, Mancall et al. (2004) estimated output growth in the lower South, with a similar conclusion, and Mancall and Weiss (1999) develop the no growth argument for the entire colonial period.

Growth, however, is not all that matters. Levels do as well, which brings us to Lindert and Williamson (2016). Their book attempts a grand synthesis, providing an overview of, and new perspectives on, output growth as well as inequality trends from the seventeenth through the twentieth centuries. I focus here on their contribution to our knowledge about historical trends in output, particularly in the period up to 1800. Kuznets and others, including David and Weiss, had built most of their estimates for the period prior to 1919 from the production side. Lindert and Williamson reverted to the political arithmetic tradition associated with Petty and Gregory King (as well as Blodget). They built income side estimates by dividing the population into groups, searching for information on their respective labor and property earnings, and constructing "social tables," as had Petty and Gregory King. Aggregating estimates of free labor earnings, property incomes, and (up to 1860) slaves' retained earnings (the costs of their subsistence), Lindert and Williamson constructed five social tables for the years 1774, 1800, 1850, 1860, and 1870. Their work is a vivid illustration of the challenges involved in building estimates of the aggregate from the income as opposed to production (value added) side.

Their most radical conclusion is that "the US had reached world leadership long before the founding fathers constructed their new republic" (2016, p. 2). It did not grow very fast during the colonial period, but the level of per capita output throughout the epoch was high (this is consistent with data on heights). They find that America's per capita income exceeded Britain's in the colonial period, fell behind during the years of the Revolutionary War and the Articles of Confederation (1775–1790), when per capita income may have declined by 30%. With the resumption of growth after the adoption of the Constitution, the country had recovered the lead over Great Britain by 1860, although it lost it during the regression associated with the Civil War and then again during the Great Depression of the 1930s.

Their overall conclusion is this: Maddison was quite wrong to argue that it was not until the start of the twentieth century that US income per capita overtook Britain's (2016, p. 9). Lindert and Williamson point out that transatlantic migration was overwhelmingly from Britain to the United States rather than vice versa, US population growth was very rapid, with women enjoying the highest fertility in the world and children experiencing the highest survival rates in the world.

Conclusion

As the modern apparatus for estimating national output, income and expenditure was developed and rationalized, cliometricians, new economic historians, and government statisticians have refined our understanding of the historical record of US economic growth. By exploring the genesis of modern national accounting systems, their main principles and conventions, and their application to historical data, this chapter has emphasized opportunities to improve our understanding of the past, not always or necessarily by using previously unavailable data but also by innovating in using known data sources and developing new ways to draw inferences from them.

Cross-References

- ▶ Cliometrics and the Great Depression
- ▶ Cliometrics of Growth
- ▶ GDP and Convergence in Modern Times
- ▶ The Golden Age of European Economic Growth
- ▶ The Great Depression in the United States

References

Blodget S (1964) Economica: a statistical manual for the United States. Augustus M. Kelley, New York. Reprint of 1806 edition

Bolt J, van Zanden JL (2014) The Maddison project: collaborative research on historical National Accounts. Econ Hist Rev 67:627–651

Bolt J, Inklaar R, de Jong H, van Zander JL (2018) Rebasing Maddison: the shape of long run economic development [Internet]. Updated 25 Jan 2018; Cited 28 Jan 2018. Available from https://voxeu.org/article/rebasing-maddison

Carson C (1975) The history of the United States national and product accounts: the development of an analytical tool. Rev Income Wealth 21:153–181

David PA (1967) New light on a statistical dark age: U.S. Real product growth before 1840. Am Econ Rev 57:294–306

David PA (2005) Real income and economic welfare growth in the early republic. Or, another try at getting the American story straight. Working Paper, Stanford University, December 9, 2005

Egnal M (1998) New World economics. Oxford University Press, Oxford

Field AJ (1985) On the unimportance of machinery. Explor Econ Hist 22:378–401
Gallman RE (2000) Economic growth and structural change in the long nineteenth century. In: Engerman SL, Gallman RE (eds) The Cambridge economic history of the United States: Vol. 2: the long nineteenth century. Cambridge University Press, Cambridge, pp 1–55
Gide C (1948) A history of economic doctrines from the time of the Physiocrats to the present day. DC Heath, New York
Gilbert M, Jaszi G, Denison E, Schwartz C (1948) Objective of national income measurement: a reply to professor Kuznets. Rev Econ Stat 30 (August):151–195
Jones AH (1980) Wealth of a nation to be. Columbia University Press, New York
Kendrick J (1961) Productivity trends in the United States. Princeton University Press, Princeton
Kendrick J (1970) The historical development of national income accounts. Hist Polit Econ 2:284–315
King W (1915) The wealth and income of the people of the United States. Macmillan, New York
King G (1696) Natural and political observations and conclusions upon the state and condition of England. The Johns Hopkins Press, Baltimore, 1936
Kuznets S (1934) National income, 1929–1932. Published as U.S. Congress, Senate, S. Doc 124, 73rd Congress, 2nd session
Kuznets S (1937) National income and capital formation, 1919–1935: A preliminary report. National Bureau of Economic Research, New York
Kuznets S (1941) National income and its composition, 1919–1938. Assisted by Lillian Epstein and Elizabeth Jenks. 2 vols. National Bureau of Economic Research, New York
Kuznets S (1946) National Product since 1869. Assisted by Lillian Epstein and Elizabeth Jenks. National Bureau of Economic Research, New York
Kuznets S (1952) National income estimates for the United States Prior to 1870. J Econ Hist 12(Spring):115–130
Kuznets S (1961) Capital in the American Economy: its formation and financing. Princeton University Press, Princeton
Lindert P, Williamson JG (2016) Unequal gains: American growth and inequality since 1700. Princeton University Press, Princeton
Maddison Historical Statistics (2018) Available at https://www.rug.nl/ggdc/historicaldevelopment/maddison
Mancall M, Weiss T (1999) Was economic growth likely in colonial British North America? J Econ Hist 59:17–40
Mancall M, Rosenbloom J, Weiss T (2004) Conjectural estimates of economic growth in the Lower South, 1720 to 1800. In: Guinnane T, Sundstrom W, Whatley W (eds) History matters: essays on economic growth, technology, and demographic change. Stanford University Press, Stanford, pp 389–424
Martin R (1939) National Income in the United States, 1799–1938. National Industrial Conference Board Studies no. 241. National Industrial Conference Board, New York
Petty W (1691) Political Arithmetick: or, a discourse concerning the extent and value of lands, people, buildings. R. Clavel, London
Quesnay F (1972) Tableau Economique. A. M. Kelley, New York
Rhode P, Sutch R (2006) Estimates of national product before 1929. In: Carter S et al (eds) Historical statistics of the United States: earliest times to the present: millennial edition, vol III. Cambridge University Press, Cambridge, pp 3-12–3-20
Romer CD (1989) The prewar business cycle reconsidered: new estimates of gross national product, 1869–1908. J Polit Econ 97(February):1–37
Romer CD (1994) Remeasuring business cycles. J Econ Hist 54(September):573–609
Rosenbloom J, Weiss T (2014) Economic growth in the mid-Atlantic region: conjectural estimates for 1720 to 1800. Explor Econ Hist 51(January):41–59
Rostow W (1960) The stages of economic growth: a non-communist manifesto. Cambridge University Press, Cambridge

Seaman E (1868) Essays on the progress of nations, in civilization, productive industry, wealth and population. C. Scribners, New York
Shaw W (1947) Value of commodity output since 1869. National Bureau of Economic Research, New York
Smith A (1937) The wealth of nations. Modern Library, New York
Spahr C (1896) An essay on the present distribution of wealth in the United States. Crowell, New York
Steckel R (1995) Stature and the standard of living. J Econ Lit 33:1903–1940
Steckel R (2006) Health, nutrition, and physical well-being. In: Carter S et al (eds) Historical statistics of the United States: earliest times to the present: millennial edition. Cambridge University Press, Cambridge
Studenski P (1958) The income of nations. New York University Press, New York
Sutch R (2006) National income and product. In: Carter S et al (eds) Historical statistics of the United States: earliest times to the present: millennial edition, vol III. Cambridge University Press, Cambridge, pp 3-3–3-12
Tucker G (1843) Progress of the United States in population and wealth in fifty years, as exhibited in the decennial census. Little and Brown, Boston
United States Department of Commerce, Bureau of Economics Analysis (2017) Concept and methods of the U.S. National Income and Product Accounts (November). Available at https://www.bea.gov/national/pdf/all-chapters.pdf
Warburton C (1934) Value of the gross national product and its components, 1919–1929. J Am Stat Assoc 29(December): 383–388
Weir D (1986) The reliability of historical macroeconomic data for comparing cyclical stability. J Econ Hist 46:353–365
Weiss T (1992) U. S. labor force estimates and economic growth, 1800–1860. In: Gallman RE, Wallis JJ (eds) American economic growth and standards of living before the civil war. University of Chicago Press for the NBER, Chicago, pp 19–78

The Census of Manufactures: An Overview

Chris Vickers and Nicolas L. Ziebarth

Contents

Abstract

This chapter provides an overview of the Census of Manufactures (COM) from the early nineteenth century through the late twentieth century. The focus is on research that uses the original establishment-level schedules. After summarizing how the COM has changed over time in terms of content and quality, the chapter demonstrates the usefulness of the COM in studying a variety of economic questions. These questions include the sources of long-run productivity growth, the causes and consequences of business cycles, and changes in the

C. Vickers
Department of Economics, Auburn University, Auburn, AL, USA
e-mail: czvickers@gmail.com

N. L. Ziebarth (✉)
Department of Economics, Auburn University and NBER, Auburn, AL, USA
e-mail: nlz0003@auburn.edu; nicolas.lehmannziebarth@gmail.com

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_66

income distribution. It then highlights the value of collecting additional establishment-level data from other sources.

Keywords
Great Depression · Manufacturing · Business Cycles · Establishment data

Introduction

Manufacturing has played a crucial role in the development of the American economy. While its fraction of employment has declined in recent decades, it has maintained a stable share of real GDP over the past 50 years, as the growth in computers and electronics balanced falls in other sectors (Baily and Bosworth 2014). It remains perhaps the central sector in the imagination of the public and policy makers. It is unsurprising then that the US federal government has carried out a regular Census of Manufactures (COM) since 1810, only 21 years after the constitution came into effect. The data from these censuses have informed numerous policy debates as far back as disputes over nineteenth century trade policy. Moreover, they have been used extensively by academics to understand the American economy. In light of this interest, we provide an overview of the COM through the years, with particular focus on the insights gleaned from the *establishment*-level schedules.

This chapter begins by discussing the focus on manufacturing, and having done so, why we then center on establishment-level data rather than published tables. One reason to focus on manufacturing is simply its importance in the American economy since at least the middle of the nineteenth century. Employment counts and share of value added in GDP underestimate the centrality of manufacturing in the economy. One reason for this is that manufacturing has historically been at the heart of technological change. While technological change has swept through agriculture and services as well over the last two centuries, those changes were dependent on the striking developments in the production of manufactured goods embodying these technological changes.

A second, related motivation for studying manufacturing is the crucial role it has played in the history of economic thought, going back to Adam Smith's famous pin factory. Intuitions about, for example, returns to scale are more often than not intertwined with how a manufacturing establishment works. A factory seems more amenable to economic theorizing compared to, for example, a bank. It generally has more clearly defined inputs and outputs as well as a clearer process linking the two. Understanding how production occurs at the level of the factory then provides important insights into these economic concepts. How production of goods occurs is particularly important when one thinks about the major technological innovations of the past, such as electricity. To fully reap the value of these innovations required that factories be completely rethought and reorganized (David 1991).

Having decided to study manufacturing, why work with establishment-level data rather than the published volumes, which, in many cases, provide a remarkable level

of detail? The tables often tabulate many important variables as well as provide relatively fine levels of disaggregation, and they have the large advantage of being far easier to collect. One obvious reason for going further down to the establishment level is it simply provides more data. The published volumes do not contain every single possible cross-tabulation or level of geographic or industry disaggregation. For example, the published volumes in many cases do not provide information for a particular industry by town. Other times the Census Bureau, for reasons not known today, collected a particular variable and never tabulated it. This may not be a truly compelling motivation for why all the schedules need to be collected, as estimates of these cross-tabulations could be obtained with a random sample of the subpopulation of interest. Of course, collecting all of the records for this subpopulation may allow for greater precision in the estimates, but this is not really an *economic* reason to do it. For example, questions about the length of the workweek were asked during the Great Depression but do not appear in the published reports. One could estimate this variable over time, conditional on covariates such as industry and geography, without fully transcribing all of the schedules.

This chapter emphasizes the importance of the establishment-level returns themselves to answer questions that are central to economic history and simply cannot be addressed with the published volumes. For us, the essential reason to collect these establishment-level records is to exploit the *establishment*-level variation. In introductory microeconomics, students are introduced to the concept of a representative establishment (or consumer) and that abstraction, while tremendously useful, wipes away all of the differences that define a particular establishment. It is not that economists thought there were no differences between establishments but questioned whether the differences mattered. The value of establishment-level data, in the end, is an empirical question.

There are a variety of reasons why these differences might matter. First, this variation is central to answering questions that might properly be classified as ones of industrial organization. For example, how do establishments within an industry compete with one another? Second, differences across establishments are useful for identifying causal macroeconomic relationships a popular use of establishment-level data recently. For example, Fuchs-Schuendeln and Hassan (2016) discuss the use of natural experiments to identify parameters, such as the fiscal multiplier, or to identify causal factors in economic growth. Third, those differences are useful for answering questions precisely about the *differences*. In thinking about earnings inequality, a natural question is how much wage dispersion occurs due to differences between establishments. In addition, understanding the sources of these differences can inform other critical questions. For example, evidence on price changes at the firm level has become an important input into models of the effects of demand shocks and monetary policy (Nakamura and Steinsson 2013).

This chapter proceeds as follows. First, there is a short history of the earliest COMs in the first part of the nineteenth century focusing on its creation and evolution up to the first "real" COM in 1850. The chapter then discusses in more detail three separate "eras" of the COM: (1) the late nineteenth century, (2) the Great

Depression, and (3) the modern era. This discussion includes the leading samples collected from these eras as well as the main questions that have been addressed. Throughout the focus is on how the *establishment-level* returns have been utilized. There are additional papers based on the published volumes that are not reviewed here.

Early Nineteenth Century COMs

The history of the COM stretches back to 1810 when it was taken in conjunction with the third Population Census. In fact, James Madison proposed that the 1790 Census include occupational statistics more than 20 years before that (Fishbein 1973). Thomas Jefferson requested similar questions be added to the 1800 Population Census. Congress rejected these proposals, but the proposals show the keen interest in collecting information on economic variables going all the way back to the founders. While the Census of Agriculture covered a much more economically important sector in the nineteenth century, it was not first taken until 1840. In some ways, the existence and persistence of the COM is remarkable. The Constitution mandates that the Population Census be undertaken for apportionment purposes each decade, but the COM has no such authority. Yet throughout the nineteenth century, there was wide support for the COM, with many politicians seeing the need for accurate data on this important sector. Fishbein (1973) discusses how the "dumping of European goods on the American market after the Napoleonic Wars and the depression of 1819 led to increasing demands for [just this type of] data" and was a main motivator for President Madison's request to Congress to authorize the 1820 Census. The existence of the COM is even more striking when one considers the stinginess of early Congresses in using funds from the public purse.

At the same time, given this lack of a clear constitutional mandate or permanent bureaucracy, it is perhaps not surprising that the completeness and quality of these earlier COMs was spotty and inconsistent from census to census. The first one in 1810 offered enumerators little by way of training and, remarkably, no actual questionnaire. This COM appears to have experienced resistance on the part of business owners in providing information, and there was no penalty for refusing to answer the questions. Fishbein (1973) goes so far as to say that because of so many inaccuracies, this census' "usefulness for research may be seriously questioned." These problems were evident contemporaneously, and numerous writers including Tench Coxe, who was tasked with tabulating returns, highlighted the myriad problems, including underreporting, misreporting, and transcription errors (Coxe 1814).

Adam Seybert, a congressman from Philadelphia, pushed the next COM in 1820. In responding to the deficiencies of the 1810 COM, Seybert successfully argued that the next one be based on an explicit questionnaire and that the enumerators receive clear instructions on the type of businesses to be enumerated. Unfortunately, these modifications did not change the overall outcome of the COM, which was still grossly inaccurate to the point where a journal of the time, the *Niles Register*, stated that "it will have been better if the subject of the inquiry had altogether been

omitted." (Fishbein 1973). There was no COM taken in conjunction with the 1830 Population Census. The 1840 COM was simplified in an important way by removing the qualitative questions. For example, the 1820 COM asked for "general remarks concerning the establishment, as to its actual and past condition, the demand for, and the sale of its manufacture." The COM no longer asked the establishment or owner's name, a change intended to increase response rates. Unfortunately, like the one before it, the 1840 response rate was terrible for the same reason as before. While effort was put into formulating the questionnaire, there was basically no effort put into organizing the actual data collection. The result was the same, with Frank Bowen, a leading economist at the time, criticizing the government for even attempting the census in the first place (Fishbein 1973).

To the best of the present authors' knowledge, there is no major collection of the establishment-level records for the COMs from the first half of the nineteenth century. Fishbein (1973) states that it had been believed that most of the 1810 schedules were lost in the War of 1812 when the British burned part of Washington. Some schedules did survive, as they were written on the population schedules, which were not all in Washington at the time. Sokoloff (1982), in his dissertation, collected a sample of these records from 1820 for the Northeast. Subsequent work by Sokoloff (1984, 1986) explored the productivity differences across establishments based on these schedules. The former work documented that smaller establishments were more productive, a seemingly puzzling finding of decreasing returns to scale.

Late Ninetieth Century COMs

The COM in 1850 was the first of sufficient quality for modern econometric study. The Census Act of 1850 provided for a comprehensive enumeration of manufacturing establishments. After this, the COM was taken decennially in hand with the Population Census through 1900. It switched to a quinquennial schedule with 1905, the first census taken separately from the Population Census, and then to a biennial cycle starting in 1919. This section focuses on the COMs from 1850, 1860, 1870, and 1880. These censuses provide insights into a key period in American economic history, with the demise of slavery and the rise of manufacturing.

Congress was cognizant of the data limitations of earlier COMs and instituted a number of reforms (Fishbein 1973). They established a Census Board to prepare schedules, and this body consulted with leading statisticians. Enumerators were paid (though inadequately) for each firm reported, and penalties were instituted for noncompliance. Further reforms in the tabulation process reduced tabulation errors in the published reports. These reforms were kept in place and further modifications and improvements made. By 1880, experts were hired to gather data and special schedules prepared for particularly important industries.

These censuses asked questions about output broken down by type, quantity, and value, as well as inputs in the form of labor, capital, and raw materials. As Atack and Bateman (1999) point out, the quality of these data is directly related to the

Table 1 Comparison of variables available

Century	Physical output	Value of capital	Hours	Employment breakdown
19th	Yes	Yes	No (half time)	Men, women, children
Early 20th	Yes	No (some physical)	No (shift length)	White, blue collar
Late 21st	Yes	Yes	Yes	Production, nonproduction

Notes: nineteenth century includes 1850, 1860, 1870, and 1880 COMs. Early twentieth century includes 1929, 1931, 1933, and 1935 COMs. Late twenty-first century includes all COMs after 1962

instructions given to the enumerators. Relative to the COMs from earlier in the nineteenth century, census officials made an effort to write clear questions for establishments and provide clear instructions to the enumerators on how to ask the questions. Even then, refinements and modifications were made to the questions to simplify the process for establishments and enumerators.

Table 1 summarizes the availability of certain variables throughout the years. For the nineteenth century, the Census instrument contains remarkably detailed information. It includes information on physical output, a limitation in some modern establishment-level datasets; capital, a limitation in the 1930s Censuses; and hours worked. At the same time, an important point emphasized by Atack and Bateman (1999) is the changes in wording of certain questions from census to census. For example, the question for "hands employed" evolved from 1850 to 1860, with much more precision in the instructions to the enumerator. In particular, rather than letting the number reflect an average over the whole year or selecting a particular day when "an average number was employed" as it was in 1850, enumerators were required to report on the "average number employed throughout the year" starting in 1860.

As is shown in the discussion of the 1930s censuses, it was not always the case that the clarity of the questions asked in COMs improved over time. Part of the problem is that there was not a permanent Census Bureau before 1902, which led to inconsistencies across censuses. For example, in 1880, "no specific instructions were given to the enumerators collecting manufacturing data beyond the questions on the forms." (Atack and Bateman 1999). There are also the perennial questions surrounding the quality of the enumeration without access to a master list of the establishments to be enumerated. This problem is compounded for the nineteenth century censuses, for which no central repository exists. Many of the schedules for particular states have been returned to their respective state archives. The National Archives in Washington, DC has a very large collection, but the poor quality of the microfilm and the fact that the forms were hand written, not typed, makes it difficult to read the text in many cases.

Atack-Bateman-Weiss Sample

The sample from the nineteenth century COMs that has served as the basis for all work at the establishment-level was originally collected by Jeremy Atack,

Fred Bateman, and Thomas Weiss (ABW) . Fred Bateman and Thomas Weiss originally collected the data from 1850 to 1870. Jeremy Atack and Fred Bateman subsequently added the 1880 data and extended the 1850–1870 series to construct nationally representative samples. These are available from ICPSR as Atack and Bateman (2004) and Atack et al. (2006). As discussed in more detail in the paper by Atack and Bateman (1999), this dataset consists of representative samples from the 1850, 1860, 1870, and 1880 COMs. "The sampling scheme was based on the number of establishments reported in each state in the summary census statistics. The goal was a sample size of between two hundred and three hundred firms from each" (Atack and Bateman 1999, p. 183).

The sample was constructed when digital storage was much more expensive than today. This led the creators of the sample to make a number of decisions on what to actually record. For example, the sample does not record the name or other identifying information of the establishment beyond the town in which it is located. This makes it impossible to link an establishment over time to study, say, the turnover of establishments. Of course, given the overall desire to construct a representative sample with "adequate microlevel diversity" (Atack and Bateman 1999, p. 183), the probability of actually being able to construct a link even for an establishment that was able to survive 10 years would be low.

Table 2 summarizes the sample, reporting only industries that have more than 50 establishments in one of the years. As is clear, the sample covers a wide variety of industries, from less technologically advanced industries like carpentry to "high tech" ones of the time like "iron forges and steel." These differences in the level of technology are reflected in differences in the scale of the operation. The average size of an "iron forges and steel" establishment is 39 times larger in terms of installed capital and about 13 times larger in terms of employment. One interesting note about the employment variable here is that it is derived from the sum of employment for men, women, and children, separately. These questions provide fascinating insights into the segregation of sexes (and children) within and across certain industries.

Value for Understanding the Development of the American Economy

Even with the limitations in terms of the size of the sample constructed, questions asked, and worries over the quality of the returns, the ABW sample has proven enormously valuable for understanding the long-run development of the American economy. One paper by Ziebarth (2013a) using this sample specifically contrasts nineteenth century America with two developing economies today, China and India, in terms of the allocation of resources between establishments. He finds similar levels of misallocation using the accounting framework of Hsieh and Klenow (2009) across these three countries. This is particularly striking given the belief that the USA had (and still has) better economic institutions than these countries and that these institutions are central to an efficient allocation of capital.

Others have used these schedules to trace the process of capital deepening and the diffusion of key technologies. Atack et al. (2005) focus on the role of establishment size and its relationship to capital per unit of labor. In light of the growth in the

Table 2 Summary statistics of the nineteenth century sample

Industry	Establishments	Log employees	Capital
Agricultural services	118	0.076	34.808
Carpentry	1,037	0.059	20.995
Meatpacking	359	0.081	269.375
Dairies	161	0.038	21.433
Flour milling	2,197	0.024	67.794
Bakeries	458	0.044	27.731
Beverages	309	0.054	128.172
Cigars	483	0.127	25.451
Broadwoven woolens	210	0.342	295.551
Yarn	106	0.247	273.153
Men's clothing	794	0.323	91.767
Millinery	242	0.177	56.393
Sawmills	3,108	0.053	51.779
Millwork	387	0.068	47.254
Wooden containers	555	0.059	23.236
Wood furniture	713	0.097	59.025
Paper mills	145	0.247	289.273
Newspapers	141	0.133	113.924
Book publishing	155	0.224	105.175
Organic chemicals	129	0.136	64.158
Nitrogenous chemicals	96	0.120	197.511
Leather tanning	800	0.069	110.695
Boots and shoes	2,289	0.089	22.491
Saddlery and harness	779	0.061	37.094
Brick and tile	411	0.118	41.069
Iron forges and steel	153	0.645	776.240
Iron castings	197	0.264	199.750
Nonferrous metals	367	0.053	68.027
Edge tools	160	0.223	148.696
Sheet metal	119	0.383	160.280
Miscellaneous fabricated metal products	186	0.166	97.356
Steam engines	256	0.306	241.504
Agricultural implements	460	0.125	130.673
Carriages	112	0.019	7.899
Wagons and carriages	837	0.082	46.720
Jewelry	172	0.119	115.159
Blacksmithing	2,325	0.022	7.559
Conglomerate enterprises	299	0.327	371.287

Notes: All statistics are calculated over the four census years. Establishments is the total number of establishments, Employees is the average number of log employees summing men, women, and children. Capital is in units of 100. The statistics are unweighted. We only report statistics for industries with more than 100 establishments over the 4 years

average establishment size during this period, these authors interpret this relationship as a shift from the artisan shop to the factory system. Atack et al. (2008) study the relationship between steam power, establishment size, and labor productivity, while Atack and Bateman (2008) examine establishment size and profitability. Using these records along with information on the rollout of the railroad network, Atack et al. (2011) link the transportation network to the growth in the average size of a manufacturing establishment. All of these works depend essentially on the information present in the establishment-level schedules that does not appear in the Census published tables.

Another major application of these data has been to return to the puzzle of a *negative* association between size and productivity, a relationship that is completely reversed in modern data. This was documented going back to the early nineteenth century by Sokoloff (1984). He argues that this "small firm effect" was due to under-reporting of the so-called "entrepreneurial labor" input. Sokoloff attempts to "solve" this problem by adjusting labor inputs upwards for the smallest firms. With this correction, he finds strong evidence for increasing returns to scale, which he takes as evidence for the returns to division of labor. Margo (2015) focuses on the adjustment made by Sokoloff and argues that there is no basis for it and, hence, the puzzle remains.

Great Depression COMs

While there were COMs taken between 1880 and 1929, as far as the present authors know, the underlying schedules have been lost due to accidents such as the fire that destroyed much of the 1890 Population Census, bureaucratic neglect, or active destruction to conserve space at the National Archives. The next COM with available establishment-level schedules is from 1929, the first of the Depression COMs. The COM had switched to a biennial schedule starting in 1919 after working on a quinquennial schedule between 1904 and 1919. For reasons that remain unclear to this day, this COM as well as the subsequent ones in 1931, 1933, and 1935, covering the first half of the Great Depression, were preserved. The physical schedules at the National Archives are organized by industry and year so it is much easier to collect industry-level samples rather than, say, state-level ones, which require going through all of the industries to find, in most cases, just a few establishments for a given industry in the state of interest. In fact, given the order on the returns, constructing a random sample is somewhat complicated with a random number generator. For example, consider the following procedure for constructing a random sample: simply sample every fifth establishment. Though sensible, this will not generate a random sample because it will under represent state-industry pairs with only a small number of establishments. In particular, think about a case where a particular state-industry pair had less than five establishments. Then the probability under this sampling strategy of having strictly greater than 1 of these establishments

is 0, whereas under a simple random sampling scheme, this probability would be strictly positive.

The schedules provide a wealth of detail about key variables on both the output and input sides of an establishment. The schedules contain information about sources of revenue by product, including not only revenue but physical output, product by product. For inputs, there is a set of questions about labor use including total wage and salary bill, number of wage earners employed at a *monthly* frequency, and average shift length for the hourly employees. The schedules ask about the number of days the establishment was in operation and the cost of intermediate goods used, allowing for a value added calculation. Finally, the intermediate goods category includes not just products used directly in the process, like timber in making furniture, but also electricity and other fuels for generating power. The intermediate goods used are, in some cases, broken down into quantity and value used in production. See Table 1 for the availability of various questions.

One issue to keep in mind when working with these records is the year-to-year variation in the wording of the questions asked. For example, the most basic question on revenue was modified in 1929. For all preceding years and all years hence, the question asked the "economist's version" inquiring about the value of output *produced that year*. In 1929, they instead asked a question about the value of products *sold that year*. This could in principle make a big difference depending on changes in inventories, particularly as it interacts with the peak in the business cycle. The COM after 1929 went back to the production-based question since the returns were found to be "unsatisfactory" as reported by the Census. At the same time, the Census stated that the switch did not invalidate comparisons of 1929 to other years. The present authors' suspicion is that establishments simply ignored the different phrasing of the question and continued to report production totals. This is based on the evidence from the cement industry where the output totals on the forms have been compared to outside sources. The differences, which are small to begin with, do not appear to be systematically larger in 1929 (Chicu et al. 2013).Turning to labor inputs, the COM provides some breakdown into white- and blue-collar labor inputs as well as information on the intensive margin of blue collar labor with information on hours worked. It is important to stress that these variables must be approached with great care because of the variation in questions across years. First, the 1931 COM does not include white-collar workers at all. For blue-collar workers, there are two different choices for the employment number. In each of the years 1929, 1933, and 1935, the census asks for the total number of wage earners on a particular date in December, with a further breakdown in 1929 between men and women. It additionally asks for the number of wage earners employed in each month. Note that "blue-collar" here does not refer to "unskilled" labor. The census forms in 1929 and 1933 ask establishments to include "skilled and unskilled workers of all classes, including engineers, firemen, watchmen, packers, etc." along with "foremen and overseers in minor positions who perform work similar to that done by the employees under their supervision" (Bureau of the Census 1932, 1936). The question for this group in 1935 was somewhat different, asking respondents to include "all time and piece workers employed in the establishment," not including

employees included in the other enumerated categories of officers, managers, clerks, and "technical employees" (Bureau of the Census 1938).

For white-collar workers, the variation across years is more complex, even setting aside the omission of information for 1931. As for blue-collar workers, the employment counts were based on a particular date in December for 1929, 1933, and 1935. Each census asks about the amount paid to officers during the year and included a count of, but not income to, proprietors. For the remaining white-collar workers, in 1929, the census includes the number of "managers, superintendents, and other responsible administrative employees; foremen and overseers who devote all or the greater part of their time to supervisory duties; clerks, stenographers, bookkeepers, and other clerical employees on salary," as well as the total amount that this group was paid (Bureau of the Census 1932). In 1933, managers and clerks are reported separately, along with their total wage bill. There is no mention there of foremen. In both 1929 and 1933, the census specifically includes foremen in "minor positions who perform work similar to that done by the employees under their supervision" as wage earners (Bureau of the Census 1932, 1936). For 1935, the same three white-collar categories of officers, managers, and clerks are reported, with the total number of clerks being reported for four separate months. For this year, there is also an entry for the number of "technical employees" including "trained technicians, such as chemists, electrical and mechanical engineers, designers, etc.," which is not asked in any other year, along with the income they are paid (Bureau of the Census 1938).

The biggest missing piece of information on the schedules regards the book value of the capital stock employed, a key production factor in any industry. Furthermore, there is no information about investment made by the establishment with which it might be possible to infer something about the capital stock. That said, there are some questions for various industries on the "quantity" of capital employed, particularly in 1929 and 1935. For example, in the ice industry, the schedule asks the establishments to report the number of compressors and the horsepower of those compressors used to make ice. Similarly, the schedules for the bakery and timber industry contain information respectively about the number of ovens and capacity of saws. Theoretically, one might back out implied investment for later years by comparing, say, the number of compressors at a particular ice establishment over time by assuming some depreciation rate. The practical difficulty with this is that the intermediate years of 1931 and 1933 have, in general, even less information on the capital stock.

To get a sense of the richness of these schedules, it is useful to compare them first to the modern establishment-level datasets including (1) the modern COM and (2) the ASM. Both the ASM/COM have similar information on employment, hours worked, wages broken down by blue- and white-collar workers, materials and energy inputs, and revenue. However, they dominate the 1930s COM with additional information on inventories, investment, and age along with establishment and firm identifiers. Each of these modern sources has their own unique drawbacks as well. The modern COM is taken every 5 years making it ill-suited for research on business cycles. The ASM, on the other hand, is at an annual frequency but lacks

the comprehensiveness of a Census, though it is representative. The bigger limitation of the ASM is the lack of information regarding prices and physical quantities of products, as establishments report only total revenue by product. The modern COM classifies the labor force into production and nonproduction workers, quite similar though not identical to the 1930s COM. All of the salaried workers would fall into the category of nonproduction workers. Some hourly blue-collar workers, such as janitors, are classified as nonproduction workers in the modern taxonomy.

There are a number of questions that can be asked about the quality of these data. Perhaps most prominently: How complete are the returns? This is really a two part question: (1) How well did the enumerators canvas all the establishments? (2) How well did the archives keep all of the schedules collected? Ziebarth (2015) contains much greater detail on these questions. On the first question, Chicu et al. (2013) compare the set of Portland cement establishments enumerated by the COM to the *Pit and Quarry Handbook*, which, among other things, contains a directory of all cement establishments. They find almost perfect concordance between the two. For the second, there are cases where particular states for individual industries are completely missing. For example, the 1931 returns from Texas in the manufactured ice industry have not been located at the National Archives.

While completeness is crucial, nearly as important is the quality of the data reported on the forms. Though there were penalties for not filling out the forms (whether they were actually enforced is a separate question), these were self-reported numbers with no checking against administrative data. By examining the schedules themselves, it is clear there was editing of the returns. Numbers are crossed off and replaced with other numbers. Sometimes it is clear that this was fixing computation errors on the part of the establishment. For example, there is a total costs field, which should just be a sum of three other components. In some cases, the establishment simply made a mistake in totaling those costs and some clerk at the Census Bureau fixed it. In other cases, for inscrutable reasons, a number is crossed out and replaced with a totally different one with no clear precedent from the form.

Probably the best one can hope for is that measurement error takes a classical form. A priori, this seems unlikely. At least today, the largest firms have account managers assigned to them whose job is to ensure a good response from those firms. According to the official documentation for the 2007 Economic Census (Gauthier 2007, p. 49):

> To identify companies requiring telephone follow-up calls after the response deadline, the NPC periodically generated and updated a list of delinquent companies (sorted by size of payroll). The Census Bureau contacted nonrespondent companies in descending payroll order (the largest companies first) and attempted to speak with the person the company had designated as its census contact.

This sort of measurement error process would be consistent with measurement error being zero on average across all establishment sizes assuming positive and negative misreports are equally likely to be identified. However, it would generate

heteroskedasticity in the measurement error as a function of the establishment size. To formalize this worry, consider the following econometric model:

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i$$
$$x_i = x_i^* + \eta_i$$

where $\mathrm{E}\left[\eta_i \middle| x_i^*\right] = 0$, $Var(\eta_i) = \sigma\left(x_i^*\right)$, and only (x_i, y_i) is observed. The difference with the classical measurement error case is that the variance of the measurement error depends on x_i^*. Assuming that $\sigma\left(x_i^*\right)$ is decreasing, then consider running the regression restricting attention to observations with $x > \bar{x}$. Then, under some conditions, the OLS estimator converges to $\beta_1\left(1 - \frac{Var\left(\eta \middle| x > \bar{x}\right)}{Var\left(x \middle| x > \bar{x}\right)}\right)$. If $Var(\eta | x > \bar{x})$ is assumed to be decreasing, the classical measurement error bounds become tighter as $\bar{x}$ gets larger. Without making more assumptions like $Var(\eta | x > \bar{x})$ goes to zero or that it decreases at certain rates as $\bar{x}$ increases, this is the best that can be done. This setup also generates an interesting bias versus variance tradeoff. By conditioning on values of x big enough, this reduces the bias but also reduces the effective sample size and the variance of the estimate. The same thing would be true if $\sigma(x^*)$ were decreasing. Then one could condition on values small enough with the same resulting tradeoff.

The question then is how big a problem this could be. To provide some empirical evidence, this present chapter studies a dataset generated by randomly sampling 25 establishments from each of 26 industries for which the authors had pictures of the original schedules. The test for the presence of editing is whether the revenue variable showed markings of edits done by the Census Bureau on the schedules. The idea is that if there is differential measurement error along these lines, there should be more markings for larger establishments. The size of the revisions in terms of the percentage of the revenue numbers by size when changes are made was also recorded.

The first regression is of the probability of a correction on size, and whether or not a form was "nonstandard." Occasionally firms received a "general form" rather than either the normal form for 1929 or the industry-specific form. This form contained considerably less information than the others. The results are in the first column of Table 3. The size of an establishment as measured by the decile of an establishment's final revenue does not strongly predict whether the figure for total revenue on the schedule form was corrected by the census.

The second regression examines if these characteristics determined whether the regression was upward, conditional on there being a revision. A third regression examines the magnitude of the change. These regressions are in columns two and three of Table 3. Size is not a predictor of these variables either. Note that the sample size is smaller here for two reasons. First, these regressions condition on there being a change in reported revenue. Second, in many cases, the original value of the revenue variable could not be deciphered, as the scratched out value was illegible.

Table 3 Predictors of edits to revenue variable

	Corrected	Revenue revised up?	Log change
Nonstandard form?	−0.08	0.85***	−0.13
	(0.05)	(0.13)	(0.31)
Decile by size of final revenue	0.01	0.03	0.02
	(0.01)	(0.03)	(0.05)
N	627	90	71
Adj. R-Squared	0.069	0.158	0.113

Notes: These data are from the 1929 COM. All regressions include industry fixed effects and standard errors are robust. The last two regressions are conditional on a change in the establishment's revenue in the first place. The nonstandard form indicator is for establishments that received a form other than the industry-specific form, so-called "Form B"
***1% significance level

One conclusion from these results is that perhaps measurement error with conditional heteroskedasticity is not much of a problem.

Bresnahan-Raff Sample

The first sample from these returns comes from the work of Timothy Bresnahan and Daniel Raff. They focused on the complete returns from particular industries. The two that they focused on in their published work were the automobile (Bresnahan and Raff 1991) and blast furnaces (Bertin et al. 1996) industries. They published the data for the cotton goods and automobile industries as Bresnahan and Raff (2011). Subsequent revisions to the cotton goods data in collaboration with Changkeun Lee and Margaret Levenstein was published as Raff et al. (2015a, 2015b). In addition, Bresnahan and Raff collected the returns from the cork, matches, soap, oil refining, tires, glass, steel, and cigarettes industries. The present authors were able to obtain their original Excel spreadsheets from Margaret Levenstein at ICPSR. One issue with those spreadsheets is that for some industries, such as the glass industry, there are variables missing.[1] For the sample discussed below, these missing variables have been reentered.

Vickers-Ziebarth Sample

The Vickers-Ziebarth sample was built over a time for a variety of projects, both by the present authors and collaborators as in Chicu et al. (2013), Vickers and Ziebarth

[1]One theory is that the original spreadsheets were transferred to Excel format, but earlier versions of Excel limited the number of columns a spreadsheet could have, thus cutting off a number of variables in industries with many products listed on the schedules.

Table 4 Summary statistics of the 1930s sample

Industry	Establishments	Log employees	Incorporated	Durable
Beverages	14,907	1.25	43.7	0
Ice cream	10,105	1.38	54.0	0
Ice, manufactured	13,242	1.47	78.9	0
Macaroni	1,269	2.09	49.3	0
Malt	131	3.04	96.8	0
Sugar, cane	279	4.45	71.9	0
Sugar, refining	77	6.40	88.3	0
Cotton goods	4,483	5.03	91.7	0
Linoleum	23	6.15	100	1
Matches	82	4.87	93.8	0
Planing mills	12,582	2.27	66.9	1
Bone black	223	3.13	98.1	0
Soap	1,004	2.38	80.9	0
Petroleum refining	1,547	4.05	94.9	0
Rubber tires	224	5.40	95.2	1
Cement	638	4.56	98.6	1
Concrete products	5,733	1.51	57.5	1
Glass	923	4.95	94	1
Blast furnaces	329	4.97	100	1
Steel works	1,720	5.62	98.5	1
Agricultural implements	916	3.17	80.4	1
Aircraft and parts	379	3.30	92.0	1
Motor vehicles	627	4.99	93.9	1
Cigars and cigarettes	145	3.97	77.2	0
Radio equipment	786	3.91	86.9	1

Notes: This table is taken from Benguria et al. (2017). All statistics are calculated over the four census years. Establishments is the total number of establishments, log employees is the average number of log employees across establishments. Incorporated is the percentage of establishments that are incorporated. Durable is whether we coded an industry's product as durable

(2014), and Ziebarth (2013b), and other such as Lee (2014) and Morin (2016). The initial set of industries included manufactured ice, cement, and macaroni. Over time, the sample came to include the original Bresnahan and Raff sample as well as the set of industries they collected but which to our knowledge are not used in published research. Table 4 shows some basic summary statistics of the combined sample. This shows the varied nature of the sample with many different types of industries covered. It has "high tech" industries such as aircraft and radios. It includes durables such as cement and steel and nondurables such as ice cream or manufactured ice. In addition, there are differences in whether the industries are mainly consumer oriented, like beverages, versus business oriented, such as planing mills. In addition, as the table shows, there is considerable variation across the industries in the relative importance of establishments that are part of multiplant (MP) firms; that is, firms comprising two or more establishments. The fraction of

Table 5 Percent of national manufacturing total of 1930s sample

Year	1929	1931	1933	1935
Establishments	11.3	10.5	9.91	9.48
Total wages	20.5	18.2	19.0	20.7
Value of product	20.5	18.4	21.0	18.8

Notes: This table is taken from Benguria et al. (2017). All of the national totals are from the 1935 published report on the Census of Manufactures, which reported the previous years totals as well

establishments in an industry which are part of an MP firm ranges from 0% in macaroni to 72% in rubber tires. When looking at the share of industry employment or revenue rather than number of establishments, the number from MP firms can be even higher: in soap, 95% of employment and 96% of revenue comes from MP firms.

The dataset includes harmonized variable names as well as establishment identifiers and, in a select set of industries, firm identifiers. Besides the cement industry, where directories from the Cement Institute were employed, we manually link establishments into their parent firm using the name of the parent company. If firms change their names over time, this can be a difficult process. Almost surely, the errors in the process of matching lead to an underestimate of the fraction of establishments that are part of MP firms. This is because it is much more likely to miss a legitimate match than it is to mistakenly make a match. This will generate the usual attenuation bias.

Table 5 shows the coverage of this dataset in terms of the fraction of manufacturing establishments as well as total manufacturing revenue and wages. In each of the four censuses, the data includes about 10% of all manufacturing establishments. Because our sample is skewed toward more large-scale industries, it ends up covering nearly twice as much of manufacturing revenue and wages. A separate question is geographic coverage of our sample. This is difficult to do because the published industry-county totals are still in the process of being collected. However, the 1930 Population Census asked a question about a person's industry of employment, allowing us to compare our total employment numbers to it. Figure 1 then shows the fraction of 1930 manufacturing employment represented in the sample by county. Obviously, one should not expect 100% coverage since we only have a select set of industries, but we do have reasonably good geographic coverage.

Value for Understanding Business Cycles

These schedules have proven immensely useful in understanding the contours of the Depression as well as the effects of numerous New Deal policies. The earliest work here by Bresnahan and Raff (1991) focused on the "cleansing" nature of the Depression. Focusing on the automobile industry, they argued that the least productive establishments exited at a higher rate. The least productive establishments

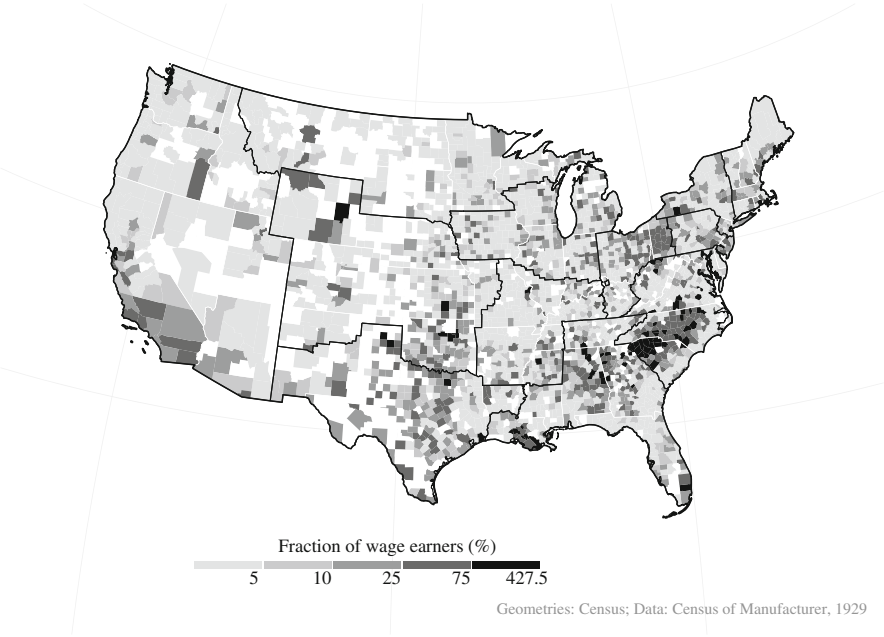


Fig. 1 Geographic coverage of 1930s sample. (Notes: This figure is taken from Benguria et al. (2017). The fraction of wage earners is the total number of wage earners in our sample in 1929 relative to the total number of manufacturing wage earners reported in the 1930 Population Census from Haines (2010). Totals may sum to over 100% if counties employ more wage earners than live in the county or if there were large declines in manufacturing employment between 1929 and 1930)

tended to be those that were craft producers. Subsequent work by Lee (2014) has challenged their interpretation that productivity was the most important selection criterion in the automobile industry. Scott and Ziebarth (2015), in a study of the radio industry, echoed Lee's point that selection seemed to be related to sheer size of the establishment rather than productivity per se. Bresnahan and Raff's original work showed the immense value of the establishment level returns. The questions they addressed were simply not answerable with the published volumes. Raff (1998) emphasized many of these same points.

Besides this work on questions of industrial organization, the establishment-level schedules have been useful in understanding productivity dynamics during the Depression. There is a marked decline in productivity from 1929 to 1933, which is much larger than one would expect given the decline in output (Ohanian 2001). Bertin et al. (1996) attempt to identify the role of short-run increasing returns to scale through a study of the blast furnaces industry. Ziebarth (2017) approaches the puzzle from a different direction by focusing on the role of misallocation of resources across heterogeneous establishments. This dovetails with subsequent work by Loualiche et al. (2017) that studies the role of firms' internal capital markets in the process of resource allocation.

Finally, a set of papers have used the establishment-level returns to better identify the causal effects of macroeconomic shocks. For example, Ziebarth (2013b) uses variation within the state of Mississippi in the extent of local bank failures, building on a natural experiment first outlined by Richardson and Troost (2009). He finds that bank failures had a negative effect on output and employment consistent with some form of nonmonetary effects of the banking crisis. Mathy and Ziebarth (2017) focus on the border of Louisiana and Mississippi to identify the effects of the political uncertainty induced by the sometimes erratic actions of Huey Long. They are not able to identify any deleterious effects, at least through this uncertainty channel.

Modern COMs

The COM continued in 1937 and 1939, was suspended during World War II, and resumed again in 1947. The timing of the COM was sporadic then for a decade with surveys done in 1954, 1958, and 1963. Finally, in 1967, the COM switched back to a quinquennial schedule that still persists today. In addition, the Census conducts an annual survey, the ASM (Nucci 1998). Longitudinally linked microdata are available through the Census' network of Research Data Centers for 1963, 1967, and 1972–2015 (United States Census Bureau 2015). Note that getting access to these data requires obtaining a security clearance and physically going to one of these secure research data centers. It had been thought that much of the data for the 1950s and 1960s ASMs had been lost, but recent work at the Census has recovered these files, which "were stored on tapes readable only by an old, archaic, and rather frail Unisys Clearpath IX 4400 mainframe that was a descendant of some of the earliest mainframes ever in existence" (Becker and Grim 2011). As Becker and Grim note, a tremendous amount of work remains to actually transform these recovered files into datasets useable for researchers. This mainframe has also been used to recover information on imputation in recent COMs, including imputation flags for total employment, total value of shipments, total cost of materials, and the wages and hours for production workers (White 2014).

Background on the Modern COM Instrument

The COM instrument for this period of time has basically all of the questions from instruments from earlier COMs including information on output by quantity and value as well as information on investment. The drawback of the COM is that because it is taken only every 5 years, it is not that useful for addressing business cycle questions. Instead, researchers have to rely on the ASM, which as the name suggests is done annually. This timing is still somewhat of a limitation since business cycles do not often last longer than a year. More problematic is the fact that the ASM is a survey and asks a limited number of questions. The ASM is not a strictly random or even stratified survey. The largest establishments are always surveyed while

smaller establishments rotate in and out based on some probability of being sampled. This makes it difficult to study *within* establishment changes over time except for the largest establishments, though the randomly selected small establishments are surveyed for a period of time.

The additional problem of the limited set of variables is reflected in the fact that the ASM does not provide information on physical output, rather it only reports revenue. If establishments all set a constant markup of price over marginal cost, this would not be a problem, since relative comparisons of revenue would reveal relative differences in physical output. Under this pricing rule, establishments that are more productive should charge lower prices and, in the end, productivity based on *revenue* rather than physical output per unit of inputs should be equal across establishments. However, as shown by Foster et al. (2008), this is certainly not the case. In fact, using the COM where there is information on physical quantity and price, they find for a set of industries that sell homogeneous products that the dispersion of *physical* productivity is much greater than *revenue* productivity while at the same time, revenue productivity is much more predictive of an establishment's survival.

One clear benefit of using the modern COMs is much clearer documentation of the procedures used to collect the records. For example, while presumably there was some list of establishments to canvas for the 1930s and earlier COMs, such a list has not been located. This would be very useful in testing the joint hypothesis that the Census did a good job of canvasing the full set of establishments as well as whether the National Archives did a good job holding onto all of the establishments. For the modern COM, this is not a problem since the canvas list is based on IRS records of establishments in existence.

Research from the Modern COM

The modern COM has proved immensely important for a number of areas of inquiry. This section focuses on the use of these data in identifying the role played by employers in generating earnings inequality. For good reasons, labor economists going back to Mincer (1975) have tended to focus on the role of characteristics of employees in explaining wage dispersion. It has only been more recently in the work of Davis and Haltiwanger (1991) that economists have begun to explore employer-level determinants of earnings beyond simply industry. For example, they find evidence that the age of an establishment is an important predictor of pay with older establishments paying more. This work cannot address *employee* characteristics in determining wages since the Longitudinal Research Datafile they were working with did not have that information. Benguria et al. (2017) study similar questions for the Great Depression.

One way in which these records have proved immensely important is in uncovering facts that are completely obscured by aggregate numbers. A particular case of this is the distinction between net and gross job flows. Obviously, economists have known that an increase in net employment over a year does not mean that there were no job losses. Instead *gross* flows exceed net flows. However, economists did

not have any idea of how much bigger these gross flows were until the work of Davis and Haltiwanger (1990). They find that while net employment growth rates (in manufacturing) between 1947 and 1993 were on average around 0, the total of job destruction and creation rates was over 12%. Even this understates the size of these gross flows because Davis and Haltiwanger (1990) were not able to use matched employee-employer data. Therefore, they are unable to distinguish an establishment with no net growth because of neither job creation nor destruction, from another establishment where there was positive but equal amounts of job creation and destruction.

Understanding these jobs flows has also had impacts on how economists think about the sources of aggregate productivity growth. As noted earlier, there is a large dispersion in productivity at the establishment level, and hence, there is potential for aggregate productivity increases by reallocating workers from low to high productivity establishments. The question then is quantifying how these gross flows contribute to the reallocation of workers and production away from the least productive establishments to more productive ones. Baily et al. (1992), among many others, find that these reallocation effects, and net entry in particular, explain an important fraction of productivity growth.

Directions for Future Work

For a long time, economic historians had to pass over in silence crucial questions regarding the development of American manufacturing due to a lack of data or, more precisely, a lack of ability to collect data. By comparison, the present is a Golden Age, at least when it comes the ability to gather, store, and sift massive amounts of historical data that were always there but just out of reach. This is reflected in the growing literature drawing on the original COM schedules surveyed here and, more broadly, in the style of leading work in economic history. This work has attempted to summarize the incredibly valuable data drawn from the establishment-level returns of the COM. These schedules have provided great insights into a number of key questions in economic history from the development of manufacturing to the role of competition policy to the sources of wage inequality.

There appears to be no reason why this quantitative turn in economic history will slow down anytime soon. If anything, with the rise of algorithms that can extract information from relatively unstructured images, “Big Data” in economic history will only get bigger. So where does the profession go from here when it comes to the COM, in particular? There is the always the obvious: just more of what is there. This seems particularly compelling for the nineteenth century COMs where the size of the Atack, Bateman, and Weiss dataset really limits the extent to which the sample can be restricted along different dimensions while still maintaining enough power. This should be relatively simple since FamilySearch.org in conjunction with the National Archives has completed the first (and most time consuming in terms of work hours) step by digitizing the nineteenth century schedules for a number of

states and made them freely available. All that would be required is having the data from those images entered into a spreadsheet.[2]

While no doubt useful, more is not enough. Instead, there seems to be great scope for collecting other information from establishment-level datasets. Economic historians will never have datasets like the Census' Longitudinal Business Database that include matched employer and employee datasets, but there is little doubt that there remain numerous creative uses pairing the COM with other data sources. These pairings can help fill in gaps with the COM. For example, one drawback of the nineteenth century COMs is their low frequency. An establishment is, at best, observed every 10 years making it unlikely that any establishment will be observed in multiple COMs. One possible way to fill in this gap would be to use the records from Dun & Bradstreet, which was a credit rating agency at the time. The information beyond existence of an establishment in these volumes, which actually covers all businesses, not just manufacturing, are limited to variables for credit rating and an estimate of net worth. Still, being able to identify when an establishment exits and even these somewhat meager financial variables would be tremendously useful. Hansen and Ziebarth (2017) do this for the Great Depression.

Another way in which matching other sources would be useful is in providing information not covered by the COM. For example, the COM is sorely lacking when it comes to financial information on the establishments beyond a question about capital invested in the nineteenth century COMs. One possibility would be linking to sources such as Moody's Manuals that would cover the largest establishments with bonds issues in financial markets. These records have very detailed information on the maturity of bonds as well as the amount. For example, Benmelech et al. (2016) examine the effects on employment of those businesses that had the unfortunate occurrence of having a bond come due during the Depression.

The present authors are optimistic that even though economic historians have been working with these schedules for nearly 30 years now, the best is yet to come. Digital cameras, cheap data storage, and outsourced (and perhaps in the near future, automated) data entry have revolutionized economic history, and this looks likely to continue far into the future.

References

Atack J, Bateman F (1999) Nineteenth-century U.S. industrial development through the eyes of the census of manufactures: a new resource for historical research. Hist Methods 32:177–188

Atack J, Bateman F (2004) National sample from the 1880 census of manufacturing (ICPSR 9385). Inter-university Consortium for Political and Social Research [distributor]

Atack J and Bateman F (2008). "Profitability, firm size, and business organization in nineteenth-century U.S. Manufacturing" in Quantitative Economic History edited by Joshua L. Rosenbloom. Routledge: Abingdon, UK

[2]The website https://www.archives.gov/digitization/digitized-by-partners.html provides additional details on what is available.

Atack J, Bateman F, Margo RA (2005) Capital deepening in United States manufacturing, 1850–1880. Econ Hist Rev 58:586–595
Atack J, Bateman F, Weiss T (2006) National samples from the census of manufacturing: 1850, 1860, and 1870 (ICPSR 4048). Inter-university Consortium for Political and Social Research [distributor]
Atack J, Bateman F, Margo RA (2008) Steam power, establishment size, and labor productivity growth in nineteenth century American manufacturing. Explor Econ Hist 45(2):185–198
Atack J, Haines MR, Margo RA (2011) Railroads and the rise of the factory: evidence for the United States, 1850–1850. In: Rhode P, Rosenbloom J, Weiman D (eds) Economic evolution and revolutions in historical time. Stanford University Press, Stanford
Baily MN, Bosworth BP (2014) US manufacturing: understanding its past and its potential future. J Econ Perspect 28(1):3–26
Baily MN, Hulten C, Campbell D (1992) Productivity dynamics in manufacturing plants. Brookings papers on economic activity, microeconomics, pp 187–249
Becker RA, Grim C (2011) Newly recovered microdata on U.S. manufacturing plants from the 1950S and 1960S: some early glimpses. CES working paper 11–29
Benguria F, Vickers C, ZIebarth NL (2017) Earnings inequality in the Great Depression. Unpublished
Benmelech E, Frydman C, Papanikolaou D (2016) Financial frictions and employment during the Great Depression. NBER WP 23216
Bertin AL, Bresnahan T, Raff DMG (1996) Localized competition and the aggregation of plant-level increasing returns: blast furances, 1929–1935. J Polit Econ 104:241–266
Bresnahan T, Raff DMG (1991) Intra-industry heterogeneity and the Great Depression: the America motor vehicles industry, 1929–1935. J Econ Hist 51:317–331
Bresnahan TF, Raff DMG (2011) Census of manufactures, motor vehicle and textile industry plants, 1929, 1931, 1933, 1935 [United States] (ICPSR 31761). Inter-university Consortium for Political and Social Research [distributor]
Bureau of the Census (1932) Fifteenth census of the United States: manufactures: 1929. United States Government Printing Office
Bureau of the Census (1936) Biennial census of manufactures: 1933. United States Government Printing Office
Bureau of the Census (1938) Biennial census of manufactures: 1935. United States Government Printing Office
Chicu M, Vickers C, Ziebarth NL (2013) Cementing the case for collusion under the NRA. Explor Econ Hist 50:487–507
Coxe T (1814) A statement of the arts and manufactures of the United States of America, for the year 1810
David P (1991) Computer and dynamo: the modern productivity paradox in a not-too distant mirror. In: Technology and productivity: the challenge for economic policy. OECD Publishing, Paris pp 315–347
Davis SJ, Haltiwanger JC (1990) Gross job creation and destruction: microeconomic evidence and macroeconomic implications. NBER Macroecon Annu 1990 5:123–168
Davis SJ, Haltiwanger J (1991) Wage dispersion between and within U.S. manufacturing plants, 1963–86. Brookings papers on economic activity microeconomics, 1991, pp 115–200
Fishbein MH (1973) The census of manufactures: 1810–1890. In: Fishbein MH (ed) The National archives and statistical research. Ohio University Press, Athens, pp 1–31
Foster L, Haltiwanger J, Syverson C (2008) Reallocation, firm turnover, and efficiency: selection on productivity or profitability? Am Econ Rev 98:394–425
Fuchs-Schuendeln N, Hassan TA (2016) Natural experiments in macroeconomics. In: Taylor JB, Uhlig H (eds) Handbook of macroeconomics, vol 2A. Elsevier, Amsterdam, pp 923–1012
Gauthier JG (2007) History of the 2007 economic census. Technical report, U.S. Census Bureau
Haines MR (2010) Historical, demographic, economic, and social data: The United States, 1790–2002. ICPSR02896-v3. Inter-university Consortium for Political and Social Research

Hansen ME, Ziebarth NL (2017) Credit relationships and business bankruptcy during the Great Depression. AEJ: Macroecon 9:228–255
Hsieh C-T, Klenow PJ (2009) Misallocation and manufacturing TFP in China and India. Q J Econ 124:1403–1448
Lee C (2014) Was the Great Depression cleansing? Evidence from the American automobile industry, 1929–1935. Unpublished, University of Michigan
Loualiche E, Vickers C, Ziebarth NL (2017) Firm networks in the Great Depression. Unpublished, Auburn University
Margo RA (2015) Economies of scale in nineteenth century American manufacturing: a solution to the entrepreneurial labor input problem. In: Collins W, Margo RA (eds) Enterprising America: business, banks, and credit markets in historical perspective. University of Chicago Press, Chicago, pp 215–244
Mathy GP, Ziebarth NL (2017) How much does political uncertainty matter? The case of Louisiana under Huey Long. J Econ Hist 77:90–126
Mincer J (1975) Education, experience, and the distribution of earnings and employment: an overview. In: Thomas Juster F (ed) Education, income, and human behavior. NBER, New York, pp 71–94
Morin M (2016) The labor market consequences of electricity adoption: concrete evidence from the Great Depression. Unpublished
Nakamura E, Steinsson J (2013) Price rigidity: microeconomic evidence and macroeconomic implications. Annu Rev Econ 5:133–163
National Archives and Record Administration (2017) Microfilm publications and original records digitized by our digitization partners. https://www.archives.gov/digitization/digitized-by-partners
Nucci AR (1998) The Center for Economic Studies Program to assemble economic census establishment information. Bus Econ Hist 27:248–256
Ohanian L (2001) Why did productivity fall so much during the Great Depression? Am Econ Rev Pap Proc 91:34–38
Raff DMG (1998) Representative firm-analysis and the character of competition: glimpses from the Great Depression. Am Econ Rev 88:57–61
Raff DMG, Bresnahan TF, Lee C, Levenstein M (2015a) United States census of manufactures, 1929–1935, cotton goods industry (ICPSR 35605). Inter-university Consortium for Political and Social Research [distributor]
Raff DMG, Bresnahan TF, Lee C, Levenstein M (2015b) United States census of manufactures, 1929–1935, motor vehicle industry (ICPSR 35604). Inter-university Consortium for Political and Social Research [distributor]
Richardson G, Troost W (2009) Monetary intervention mitigated banking panics during the Great Depression: quasi-experimental evidence from a Federal Reserve District Border, 1929–1933. J Polit Econ 117:1031–1073
Scott P, Ziebarth NL (2015) The determinants of plant survival in the U.S. radio equipment industry during the Great Depression. J Econ Hist 75:1097–1127
Sokoloff KL (1982) Industrialization and the growth of the manufacturing sector in the northeast, 1820–1850. PhD dissertation, Harvard University
Sokoloff KL (1984) Was the transition from the Artisinal shop to the non-mechanized factory associated with gains in efficiency?: evidence from the U.S. manufacturing censuses of 1820 and 1850. Explor Econ Hist 21:351–382
Sokoloff KL (1986) Productivity growth in manufacturing during early industrialization: evidence from the American northeast, 1820–1860. In: Engerman SL, Gallman RE (eds) Long-term factors in American economic growth. University of Chicago Press, Chicago, pp 679–736
United States Census Bureau (2015) Federal statistical research data centers. Online
Vickers C, Ziebarth NL (2014) Did the NRA foster collusion? Evidence from the macaroni industry. J Econ Hist 74:831–862

Vickers C, Ziebarth NL (2018) United States Census of Manufactures, 1929–1935 (ICPSR 37114). Inter-university Consortium for Political Science [distributor]
White TK (2014) Recovering the item-level edit and imputation flags in the 1977–1997 censuses of manufactures. CES working paper 14–37
Ziebarth NL (2013a) Are China and India backward? Evidence from the 19th century U.S. census of manufactures. Rev Econ Dyn 16:86–99
Ziebarth NL (2013b) Identifying the effects of bank failures from a natural experiment in Mississippi during the Great Depression. AEJ Macroecon 5:81–101
Ziebarth NL (2015) The Great Depression through the eyes of the census of manufactures. Hist Methods 48:185–194
Ziebarth NL (2017) Misallocation and productivity during the Great Depression. Unpublished, Auburn University

Decolonizing with Data

The Cliometric Turn in African Economic History

Johan Fourie and Nonso Obikili

Contents

Abstract

Our understanding of Africa's economic past – the causes and consequences of precolonial polities, the slave trade, state formation, the Scramble for Africa, European settlement, and independence – has improved markedly over the last two decades. Much of this is the result of the cliometric turn in African economic history, what some have called a "renaissance." While acknowledging that cliometrics is not new to African history, this chapter examines the major recent contributions, noting their methodological advances and dividing them into four broad themes: persistence of deep traits, slavery, colonialism, and independence. We conclude with a brief bibliometric exercise, noting the lack of Africans working at the frontier of African cliometrics.

J. Fourie (✉)
LEAP, Department of Economics, Stellenbosch University, Stellenbosch, South Africa

Stellenbosch University, Stellenbosch, South Africa
e-mail: johanf@sun.ac.za

N. Obikili
LEAP, Department of Economics, Stellenbosch University, Stellenbosch, South Africa
e-mail: me@nonsoobikili.com

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0_69

Keywords
Africa · History · Poverty · Reversal of fortunes · Sub-Saharan · Trade · Slavery · Colonialism · Missionaries · Independence

Introduction

Africa was not always the poorest continent. Although by 2016, Sub-Saharan per capita incomes (at 1632 constant 2010 US$) was the lowest of the seven major world regions (The World Bank 2017),[1] this was not true, at least for certain parts of sub-Saharan Africa, only five decades earlier. This is demonstrated in an award-winning *Journal of Economic History* – article by Dutch scholars Ewout Frankema and Marlous van Waijenburg - a paper that has done much to showcase the contribution that quantitative African economic history can make to improve our understanding of Africa's development path. Frankema and Van Waijenburg show that mid-twentieth century urban unskilled real wages were well above subsistence levels, rising significantly over time. They also show that in parts of West Africa and Mauritius, wages were considerably higher than those of Asian laborers at the same time (Frankema and Van Waijenburg 2012).

We also have good reason to suspect that Africa was not the poorest continent five centuries earlier. In what is now a seminal contribution, Acemoglu et al. (2002) famously described a "reversal of fortunes"; the dense populations in parts of precolonial sub-Saharan Africa suggested high living standards around 1500, on par or even above living standards elsewhere. After the onset of the Industrial Revolution in Europe and then North America, and the consequences of colonization, these fortunes were reversed (Acemoglu and Robinson 2010).

What, then, were the reasons for the demise of Africa's comparative fortunes? Where and when did living standards rise, and why did it falter? What inhibited Africans from reaping the benefits of the technological and institutional innovations of the last two centuries? And, given the continent's existing low living standards, to what extent can African economic history inform contemporary policy-making?

These are only some of the questions that have sparked a "renaissance" in African economic history over the last decade (Austin and Broadberry 2014). A new generation of economists and economic historians are rewriting African economic history, aided by larger datasets and innovative empirical techniques (Fourie 2016). This is a sharp turnaround from the "recession" in African economic history scholarship that began in the 1980s, the result of formalization in economics, the cultural shift in history, and the "Afro-pessimism" related to the poor economic performance of many African countries (Collier and Gunning 1999; Hopkins 2009;

[1]Only slightly below per capita income in South Asia, at $1690.

Austin and Broadberry 2014). It is therefore no surprise that with the rise in Africa's fortunes post-2000, and with the expectation of a demographic dividend in the coming century[2], interest in understanding Africa's economic past is at an all-time high. To give one example: in the period 1997–2008, only 10 papers on Africa were published in the leading five economic history journals. Since then, 35 papers have appeared.

The "renaissance" has been characterized by two approaches. The "history matters"-school usually seek to establish a causal relationship between a variable in the (deep) past – like settler mortality in the case of Acemoglu et al. (2002) and an outcome variable in the present (or recent past). These approaches rely on rigorous econometric techniques, seek to establish singular causal relationships, and typically use already-published source material, such as the Roome map, the Murdock Atlas, or FAO crop suitability indicators, instead of collecting new primary materials from archives. It mainly exploits within-African spatial variation in development outcomes at one point in time rather than fluctuations and trends across time. In the econometric jargon, these are "small T, large N" studies.

By contrast, the "historical reconstruction"-school typically seeks to fill gaps about our knowledge of Africa's economic past, such as long-term trends in population, taxation, wages, inequality, biological standards of living, education, social mobility, etc. Frankema and Van Waijenburg's (2012) contribution is an excellent example of this approach. In these studies, historical accounting and basic quantitative methods are often preferred to econometric techniques that allow for causal interpretation. Although these studies are also comparative (British versus French, settler vs. peasant economies), their comparisons typically have a fairly small N. In other words: small N, large T.

This chapter will showcase the breadth and depth of both schools, noting, in particular, the use of cliometric analysis in understanding the causes and consequences of African historical development. We first discuss the latest evidence on Africa's fluctuating fortunes. We then turn to the possible explanations for the divergence of African economies in the late twentieth century. These include factors and events deep in African history, including the spread of agriculture, disease, and cultural attributes. Slavery in Africa has received much attention as a cause for Africa's slow growth – and we review the literature on that in section four. In section five, we review the colonial impact of missionaries, settlers and the rise of post-colonial states, and its contribution to African economic performance. Finally, we discuss African economic history scholarship. Who gets to write about Africa's economic past? We conclude that more needs to be done to draw African scholars into the field.

[2]Today, one out of six people on Earth live in Africa. That is likely to increase to one in four in 2050, and to one in three by 2100. See Pison (2017).

Fortunes, Reversed and Revised

Africa was not always the poorest continent but measuring its fluctuating fortunes has not been easy (Jerven 2018). The lack of written records, especially for the precolonial period, complicates any long-run analysis and has forced economic historians to find creative approaches to measure the trajectory of living standards across time.

Unreliable population estimates, especially for the precolonial period, is one example of the difficulty of plotting continent-wide historical living standards. Frankema and Jerven (2014, p. 908), in an attempt to backwardly project sub-Saharan population estimates, conclude: "For the pre-colonial period the empirical evidence is so thin that it suffices to point to Thornton's work on baptismal records from missionaries in the kingdom of Kongo. The colonial censuses are in turn widely discredited, and therefore not used as authoritative benchmarks, and while the population in post-colonial Africa is better recorded, census taking has remained uneven, irregular, and incomplete." Despite these shortcomings, Frankema and Jerven (2014) calculate that sub-Saharan Africa housed 240 million people in 1950. They then rely on population growth rates of other regions, notably South-East Asia, to arrive at a population estimate of 100 million in 1850. This is in sharp contrast to earlier revisions by Manning (2010), who had used Indian population growth rates to calculate a much higher precolonial population number for sub-Saharan Africa. Figure 1 compares the two revised estimates.

Getting population numbers right is necessary to understand the process of economic development in Africa. Population density is frequently used as a proxy for precolonial prosperity; more densely populated areas, it is assumed, live off a greater surplus (Acemoglu et al. 2002). But as Frankema and Jerven note, population

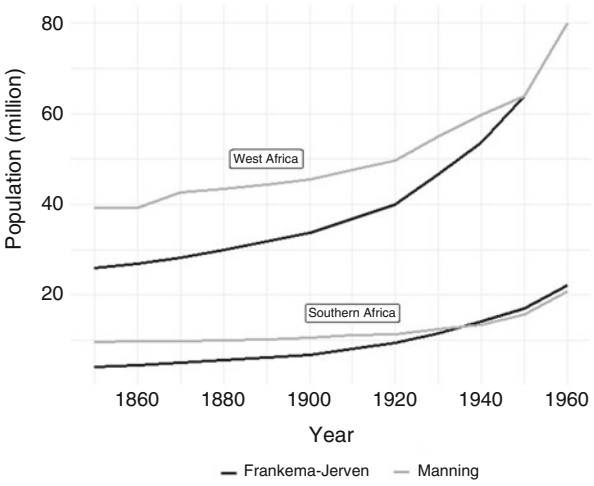


Fig. 1 Population size for Southern and West Africa. (Source: Frankema and Jerven (2014), redrawn)

numbers, even into the colonial era, may be severely biased and thus result in incorrect assertions about economic outcomes. Fourie and Green (2015) offer one example. They combine household-level settler production in the eighteenth-century Dutch Cape Colony with anecdotal accounts about Khoesan farm laborers to show that the Khoesan, while not formally recorded by the colonial authorities, was an important component of farm labor on settler farms. By obtaining more accurate population numbers, Fourie and Green (2015) show that previous research had overestimated slave productivity, social inequality, and the level of gross domestic product.

Gross domestic product, of course, provides a far more reliable picture of the rise and fall of living standards than mere population numbers. But such estimates are notoriously unreliable, even today. Morten Jerven (2010 p. 147), analyzing historical gross domestic product (GDP) estimates of African territories, agrees: "Data on the post-colonial period are less reliable than is commonly thought." In a widely acclaimed book, Jerven (2013) warns against the uncritical use of postcolonial GDP statistics in cross-country regressions, noting the errors and biases in the sources from which these estimates are generated. For these reasons, and despite some attempts to construct historical GDP series of African countries or regions (Fourie and Van Zanden 2013; Bolt and Van Zanden 2014; Inklaar et al. 2018), few African countries can boast a long-run GDP series.

In the absence of reliable GDP statistics, wages offer an alternative. Frankema and Van Waijenburg (2012, p. 896) note that real wages have "the advantage that they better reflect the living standards of ordinary African workers... they focus on the purchasing power of African laborers leaving aside the significantly higher income levels of European settlers and/or Asian migrant workers." While they were not the first to calculate real wages in Africa (Bowden et al. 2008; De Zwart 2011; Du Plessis and Du Plessis 2012), they were the first to do so across several countries and for multiple years of the colonial period. Using a standardized basket of goods to calculate real wages (Allen et al. 2011, p. 922), their results were the first to reveal the unexpectedly high relative living standards of West Africans compared to unskilled laborers in Asian countries during most of the colonial period, a result, they suggest, that calls "for a reinterpretation of the path-dependence nature of African economic development."

One concern is that their results only capture urban laborers. De Haas (2017) reconstructs typical farm size, production, and income to calculate real income from farming activities in Uganda. His novel approach suggests that rural farmers, similar to their urban unskilled counterparts, were living well above subsistence levels at a level remarkably constant over time. De Haas (2017) also shows that during the 1950s and 1960s, urban wages and rural incomes strongly diverged. Internal and external political pressures in the wake of independence drove urban wages upwards, but cash crop prices and resultant rural incomes did not experience a similar sustained improvement. Increasingly, urban laborers became an economically privileged group. If the experience of Uganda can be generalized, the large rural-urban income gap that characterized many postcolonial African economies

may find its roots in this period. Bossuroy and Cogneau (2013) use backward projection of present day household survey data to show that rural-urban income gaps were particularly large in three former French colonies compared to three former British colonies. Cohort analysis, as employed by Bossuroy and Cogneau (2013), is an innovative way to uncover trends during the late-colonial and post-colonial period.

In response to the poor aggregate data quality of African economic history, an exciting development is the turn to a "history from below," or the use of individual-level records (Fourie 2016). More precise measures of income, wealth, and production are available, for example, in probate inventories and tax censuses. These records are expensive to transcribe but are invaluable in studying the transfer of wealth across generations, notably in populations where almost all individuals are farmers. Fourie (2013) uses more than 2500 probate inventories of the eighteenth-century Dutch Cape Colony to show that earlier depictions of the Cape as a "social and economic backwater" are not supported by the empirical evidence. These records are, unfortunately, limited to settlers only.

Economic historians are becoming more creative. Individual-level records that survive in government or church archives, like military attestation records or baptism and marriage records, encoded information that can now be used for reasons orthogonal to its original purpose, circumventing the potential biases of the colonial authorities. Church records, for example, allow investigations into social mobility (Meier zu Selhausen et al. 2017; Cilliers and Fourie (2017) and gender inequality (Meier zu Selhausen 2014; Meier zu Selhausen and Weisdorf 2016). There are several new, large research projects underway to digitize and transcribe more of these and similar records.

Attestation forms often include the height of recruits. Human height, or stature, is widely used as a proxy for living standards, as it captures not only genetic traits but also environmental conditions like access to nutrients and the disease environment (Steckel 1995; Baten and Blum 2012). While stature was first used to document the living standards of Africans shipped to the Americas as slaves (Steckel 1979; Moradi and Baten 2005; Moradi 2010) were the first to use African heights, obtained from twentieth-century household surveys, to plot living standards and inequality on the continent. The challenge was to find evidence for the early colonial and even precolonial period. Moradi (2009) turned to individual-level records to do this. Using the attestations of Kenyan army recruits, Moradi (2009, p. 719) finds an upward trend during both the colonial and postcolonial period. His results suggest that "however bad colonial policies and devastating short-term crises were, the net outcome of colonial times was a significant progress in nutrition and health." More recently, Mpeta et al. (2018) have used a combination of military attestations, cadaver records, and household surveys, to plot the changes in the heights of black South Africans over the twentieth century. As earlier sources that measure height are uncovered and transcribed, the living standards of Africa's diverse people will be traced deeper into history.

Deep Roots of Divergent Development

The lack of empirical evidence at the individual level for the precolonial period has not prevented a new generation of social scientists from investigating how deep historical factors still shape African economic development today. In fact, the ability to now spatially map environmental, political, and cultural factors in deep history and overlay it with contemporary outcomes allow economists to expose correlations – and even causality – that were simply not possible in the age before sufficient computing power, accessible software and, most importantly, rigorous econometric techniques. The deep roots of divergent development have indeed attracted most interest from cliometricians of Africa.

These deep roots, some argue, go back as far as the migration of *Homo sapiens* out of Africa 70,000 years ago. In a seminal, if controversial study, Ashraf and Galor (2013) present empirical results that show an inverted-u correlation between the distance from Ethiopia – considered the route by which *Homo sapiens* left Africa – and incomes today. They argue that too much and too little genetic diversity is bad, which is why Native American and African populations have lagged behind Europe and Asia.

In response to Ashraf and Galor, anthropologists and other social scientists noted several inconsistencies in their data quality and assumptions (Guedes et al. 2013). They note, for example, the poor data quality used for calculating population densities in 1500 – Ashraf and Galor use figures, they show, that are largely outdated. Ashraf and Galor also make assumptions inconsistent with research in other fields; Guedes et al. (2013) point to the lack of research that links genetic diversity to general diversity.

Some have attempted to replicate Ashraf and Galor's methods in different settings. Using African countries only, Asongu and Kodila-Tedika (2017) find contrasting results, suggesting that "poverty is not in the African DNA". Others have used the migratory distance, the instrument used by Ashraf and Galor, to investigate outcomes like cultural traits instead of genetic inheritance (Gorodnichenko and Roland 2017; Desmet et al. 2017), or have used more precise genetic traits, like DRD4 exon III allele frequencies, to infer its impact on economic development (Gören 2017). As more precise genetic information becomes available, especially in Africa, with its diverse genetic composition, deep history as reflected in genetic inheritance is likely to be a fertile area for new research.

The reason for Africa's genetic diversity is, inter alia, the result of a rich variety of environmental conditions and its effect on natural selection. But it was not only humans that evolved to fit their environment. Insects did too. The TseTse fly, found in tropical areas throughout Africa but not on any other continent, transmits a parasite that is harmful to humans and lethal to livestock. Alsan (2014) is the first to investigate empirically the long-run effects of the TseTse fly on economic outcomes. She first shows, repeated in Fig. 2, that the TseTse fly was predominantly found in the regions most suitable for agriculture. Alsan then demonstrates that the presence of the TseTse fly reduced the agricultural surplus of farmers, as they were

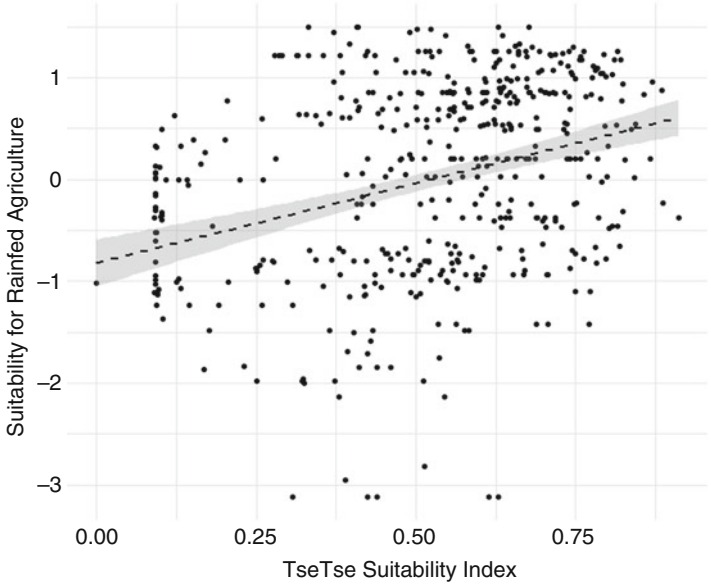


Fig. 2 Correlation between rainfed agriculture and TseTse fly suitability. (Source: Alsan (2014), redrawn)

less likely to use domesticated animals and the plow, had lower population densities, and were less politically centralized.

Climate and environmental conditions also shaped the type and speed of agricultural adoption. Michalopoulos et al. (2016) find that individuals from ethnicities that derived a larger share of subsistence from agriculture in the precolonial era are today more educated and wealthy. The reasons for this, they suggest, are differences in attitudes and beliefs and differential treatment by others. One alternative mechanism through which these early agricultural practices may persist is the complexity of precolonial political regimes that resulted from them.

Gennnaioli and Rainer (2007) show that the strength of precolonial political institutions was an important factor in the capacity of colonial and postcolonial governments' provision of public goods. In a groundbreaking study, Michalopoulos and Papaioannou (2013) show how the spatial distribution of precolonial ethnicities affects contemporary economic performance. Areas that had higher levels of political centralization in precolonial times today exhibit more economic activity, as proxied for by satellite images of light density at night. This association, they find, is independent from geographic features or other observable ethnic-specific cultural and economic variables.

The persistence of political centralization – or formal institutions – and their interaction with attitudes and cultural norms and beliefs – or informal institutions – within Africa is the subject of a large recent research project on the Kuba Kingdom of Central Africa. Lowes et al. (2017) conduct experiments on

descendants of individuals that lived within and just outside the borders of the seventeenth-century Kuba Kingdom, a centralized state with an unwritten constitution, a judicial system with courts and juries, a police force, taxation, and public goods provision. The experiments – the Resource Allocation Game and the Standard Ultimatum Game – are performed on 499 individuals. One subgroup of these individuals are the descendants of the Woot, a group of culturally similar people that had lived both inside and outside the Kuba Kingdom. Lowes et al. find that the descendants of those groups that settled outside the Kuba Kingdom are today *more* likely to have strong norms of rule-of-law and a lower propensity to cheat than the descendants of those that lived within the Kuba Kingdom, with its strong formal institutions. They argue that this is consistent with a "model where endogenous investments to inculcate values in children decline when there is an increase in the effectiveness of formal institutions that enforce socially desirable behavior" (Lowes et al. 2017, p. 1065).

Reflecting their novel methodological contributions, both the Michalopolous and Papaioannou (2013) and Lowes et al. (2017) studies are published in *Econometrica*. They reflect the frontier of cliometrics in Africa; first, in using innovative contemporary outcome variables – satellite images and experiments – and, second, their careful use of *causal* interpretations. Michalopoulos and Papaioannou (2013, p. 114) explicitly acknowledge that their evidence should not be interpreted causally – "(s)ince we do not have random assignment on ethnic institutions, this correlation does not necessarily imply causation." While Lowes et al. (2017, p. 1089) can exploit the random assignment of those treated in the Kuba Kingdom and those outside its borders, they are careful to note that their experiment can only test the "causal impact of a particular bundle of state institutions." The exercise of causally linking anthropological evidence of the precolonial period to both formal and informal institutions today are fraught with difficulties. Carefully selected instruments may offer a solution. Next we will see how this strategy was used to investigate the consequences of Africa's most infamous historical episode: the Atlantic slave trade.

The Slave Trades: Causes, Consequences, and Controversies

Historians of Africa have long studied the devastation caused by Africa's slave trades. With an approximate 12 million Africans shipped in the Atlantic slave trade, and another combined 6 million in the 3 other trades – the trans-Saharan, Red Sea, and Indian Ocean – between 1400 and 1900, the population of Africa, according to some estimates, was only half of what it would have been without the slave trades (Manning 1990).

The study of Africa's slave trades was one of the first topics in African economic history that made use of large data sets and statistical analyses (Eltis 1977; Eltis 1987; Inikori 1976). A generation later, and with the advancement of computing power and easily accessible statistical software, the African slave trades remains one

of the most studied topics in African history. We separate these studies into two main focus areas: first, the study of the trade itself, its size, causes, and mechanics, and, second, its consequences.

Demand and supply explain why the Atlantic slave trade, by far the largest of the overseas trades, arose in the sixteenth century, and why the majority of slaves was of African origin. On the demand side, African labor was highly productive in much of the New World, "discovered" and colonized by Europeans from the late fifteenth century. Eltis et al. (2005, p. 696) calculate that, in the Caribbean over the period 1674–1790, "total factor productivity in slave agriculture increased markedly, and the demand for slave labour increased by a factor of at least four."

On the supply side, Africans' resistance to tropical diseases and their proximity to the Americas made them more attractive than European, Indian, and Chinese laborers (Bertocchi and Dimico 2014; Angeles 2013). This was possible, Angeles (2013) argues, because of the low costs of slave capture in Africa, which was mostly undertaken by Africans themselves. Because slaves were mainly obtained from different ethnic groups, Africa's ethnic fragmentation, itself the consequence of few large states and limited penetration of any of the world's major religions, is one reason for the low cost of slave labor on the continent, and the profitability of the slave trade.

Climatological conditions also affected slave supply. Fenske and Kala (2015) find that more slaves were exported in colder years along the African coast. This is because lower temperatures reduced mortality and raised agricultural yields, lowering the costs of slave transport. Their results suggest that a temperature increase of one degree celsius reduced annual exports by roughly 3000 slaves per port. Rainfall, or the lack thereof, also mattered. Levi Boxell (2017) shows that nineteenth century drought increased the number of slaves exported from a given region. He also uses geocoded data on nineteenth century African conflicts to show that drought increased the likelihood of conflict, but only in the slave exporting regions of Africa.

European technology, notably guns, played a key role in the slave trade too. Whatley (2017) use a Vector Error Correction Model of annual slave and trade statistics to show that gunpowder imports and slave exports were co-integrated in a long-run relationship. Gunpowder imports "produced" additional slave exports, and additional slave exports attracted additional gunpowder imports. He makes use of several placebo tests, as well as an instrumental variable of excess capacity in the British gunpowder industry to support the Gun-Slave hypothesis.

The slave trade itself was highly inefficient. Dalton and Leung (2015) find that voyage output, measured as the number of slaves that disembarked in the Americas, varied substantially across voyages within a European country. The dispersion in output was the highest across Portuguese voyages, lower across French voyages, and lowest across British voyages. Dalton and Leung (2015) then calculate the total factor productivity gains had the dispersion of distortions disappeared. Their results show that the dispersion of distortions had the smallest damage to total factor productivity in Great Britain, followed by Portugal, and then France.

While historians are interested in the causes of African slavery, economists tend to focus on its consequences.[3] What had been difficult to ascertain, however, was the extent to which the slave trades were responsible for the poor economic performance of many African countries by the late twentieth century. Nathan Nunn's job market paper, published in the *Quarterly Journal of Economics* in 2008, was a first attempt at a *causal* interpretation (Nunn 2008). Nunn first shows that countries that had higher numbers of slaves removed are also poorer today. He then makes two arguments in favor of a causal relationship: first, from historical and basic descriptive evidence, it would seem that it was the more prosperous regions, and not the poorest regions, that selected into the slave trades. The novelty of Nunn's contribution, however, rests on his second approach, the use of an instrumental variable. His instrument is the distance from each African country to the slave markets in the Americas: the greater the distance, the lower the number of slaves shipped from that African country. This requires the author to make the assumption that the distances are unrelated to economic outcomes today, except through the effect of the slave trade. Nunn's IV-estimates support the OLS-estimates and his argument: the slave trade still had a negative effect on African economies in the late twentieth-century.

Nunn's causal interpretation ignited interest in identifying the *mechanisms* that explain the persistent effect of slavery he had found. Nunn himself first attempted to address this in the final section of his 2008 paper, noting ethnic fractionalization and state development as two plausible explanations. But it would be his later work, on ruggedness and trust, that would offer more sophisticated explanations for slavery's persistence.

With Diego Puga, Nunn postulated that one mechanism through which the slave trades could affect present-day outcomes is geography (Nunn and Puga 2012). Rugged terrain, they argue, afforded protection to those being raided during the slave trades. Many Africans thus escaped to rugged areas, terrain that is "tough to farm, costly to traverse, and often inhospitable to live in" (Nunn and Puga 2012, p. 20). These areas, however, while offering protection, were less likely to generate large surpluses or offer better trade opportunities. Africans' economic prospects were thus stymied by their rugged location, a consequence of the centuries' long slave trade.

Soon after Nunn and Puga (2012), Nunn published another paper, this time with Leonard Wantchekon, that postulated a second mechanism to explain the persistent effects of the slave trade. Nunn and Wantchekon's "Slave Trade and the Origins of Mistrust in Africa," published in the *American Economic Review* (2011), argues that the greater involvement of the slave trade resulted in lower levels of trust, as seen in Fig. 3, and that these cultural norms persisted over time. They use the distance of ethnic groups from the coast at the time of the slave trade as an instrument for the number of slaves taken. A number of falsification tests that examine the reduced-form relationship between distance from the coast and trust inside and outside of Africa suggest that the exclusion restriction is satisfied. Nunn and Wantchekon

[3]See Bertocchi (2016) for an extensive review of the slave trade legacies.

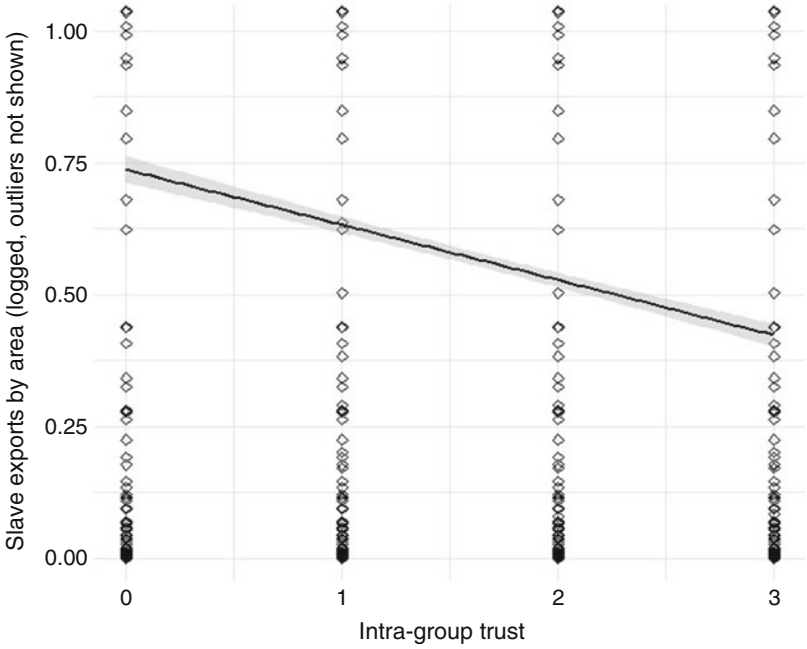


Fig. 3 Correlation between slave exports and intragroup trust. (Source: Calculated from Nunn and Wantchekon (2011))

(2011, p. 3223) explain: "Places farther from the coast had fewer slaves taken, and therefore exhibit higher levels of trust today. If distance from the coast affects trust only through the slave trade, then there should be no relationship between distance from the coast and trust outside of Africa, where there was no slave trade. This is exactly what we find."

That the slave trade causally affects trust several centuries later is an important empirical finding, but it still does not explain the mechanism – or channel – that is responsible for this persistence. Nunn and Wantchekon (2011) put forward two possible explanations. First, the slave trade altered the cultural norms of the ethnic groups exposed to it, making them less trusting of others. Second, the slave trade may have caused the deterioration of legal and political institutions. Individuals in heavily affected regions may be less trusting today because their leaders and institutions are less trustworthy. The authors conduct three exercises to measure the size and significance of each of these two mechanisms. All three exercises show that both channels are important, but that the internal channel – the effect through cultural norms – is at least twice as large as the external channel.[4]

[4]In a replication study with an updated dataset, Deconinck and Verpoorten (2013) confirm the authors' results.

Did the slave trade perhaps bring beneficial outcomes for African traders and farmers? Rönnbäck (2015) argues that the demand for provisions from the external slave trade was too small to have any substantial effect on African commercial agriculture. Focusing on the Gold Coast, he shows that some African laborers located in the coastal European enclaves experienced an initial boost to living standards, but then their conditions declined. It was only a small group of highly privileged employees that benefited consistently, exacerbating social stratification. Dalrymple-Smith and Frankema (2017) agree. Exploring the provisioning strategies of 187 British, French, Dutch, and Danish slave voyages conducted between 1681 and 1807, they show that during the eighteenth century, an increasing share of the foodstuffs required to feed African slaves were taken on board in Europe instead of West Africa. Although there was considerable variation in provisioning strategies among slave trading nations and across main regions of slave embarkation, the average slave trade-induced demand impulse was weak.

The focus, however, has been on the long-term persistent consequences of the trade, and the mechanisms through which the shock of slavery persists into the present. Education is one example. Using slave data and colonial censuses of Nigeria and Ghana, Obikili (2015) finds a negative and significant correlation between slave export intensity before the colonial era and literacy rates during the colonial era. Using contemporary data, he shows that this relationship persists into the present.

Violence and conflict, and how it reinforces itself in a low-level equilibrium, is another channel. Fenske and Kala (2017) find a discontinuous increase in conflict after 1807 in areas affected by the slave trade. In areas where the slave trade declined, they argue, political leaders resorted to violence to maintain their influence. The slave trade shifted south and east, associated with an increase in violence in those regions. The slave trade also resulted in political fragmentation. Although this channel was first proposed by Nunn (2008), it has found additional support in Obikili (2016). He shows that villages and towns of ethnic groups with higher slave exports were more politically fragmented during the precolonial era, and that this fragmentation is reflected in political outcomes today.

And then there are the myriad social beliefs, norms, and other informal institutions that were affected. Obikili (2016) shows, for example, correlations between slavery, political fragmentation, and the propensity to bribe or be corrupt. The slave trade also helps to explain different rates of polygamy in western and eastern Africa. More male slaves were exported from western Africa, while more female slaves were exported in the Indian Ocean trade. Dalton and Leung (2014) link historical slave trade data with current rates of polygamy and find that the transatlantic slave trades cause polygamy at the ethnic group level while the Indian Ocean slave trades do not.

And its long-term consequences can even affect outcomes like access to finance. Pierce and Snyder (2017) show that the slave trade is strongly correlated to reduced access to the formal and trade credit markets. The effect is particularly strong for capital investment in smaller firms that are not business groups. Because the slave trade cannot explain any other business obstacle, the authors argue that its effects persist through informal institutions like mistrust and weakened institutions. Levine

et al. (2017, p. i) confirm this channel, noting that the "slave trade is strongly, negatively related to the information sharing and trust mechanisms but not to the legal mechanism."

In the end, Nunn's cliometric contribution not only sparked interest in the consequences of African slavery, but also caused some debate in the field of African economic history, notably from historians concerned with the quality of the source material (Reid 2011; Austin and Broadberry 2014). Gareth Austin (2008), for example, while welcoming the new interest in Africa's past, cautioned against the "compression of history," the practice of conflating different data-generating processes across decades or even centuries. Anthony Hopkins, too, welcomed the new approaches for "their boldness, their freshness and their potential for re-engaging historians in the study of Africa's economic past" but criticized them on methodological and empirical grounds (Hopkins 2009, p. 155). He notes, in particular, the poor data quality – "the population figures they assemble ... are insufficiently robust to carry the explanatory weight placed on them" (Hopkins 2009, p. 166) and then notes that the "regression analysis is only as robust as the numerical evidence it draws on" (Hopkins 2009, p. 168).[5]

Debates about data quality and source bias are not limited to slave data, of course. Colonial-era written records would introduce new forms of measurement error and bias that would have to be accounted for.

Colonialism and Independence

The age of slavery was followed by the age of colonialism. By the nineteenth-century, European missionaries and explorers were entering deeper into the African interior in search of souls and treasure. They were soon followed by settlers and imperialists, claiming large parts of the continent for European powers.

There are at least two economic questions that deserve our attention: First, what explains the emergence of colonialism? Colonialism – or colonization – was, of course, not one thing**.** It was a heterogenous *process* of political, economic, and psychological subjugation and dispossession with the aim of advancing the political and financial power of the colonizer.

But why did it emerge when it did, and what explains the variety of colonial regimes? A second question, perhaps more difficult to answer, is: What were its consequences? What did missionaries do, and how would that shape Africans' attitudes, beliefs, and freedoms? What was the response to the arrival of European settlers, with new crops, technologies, and diseases, and how did it affect local production systems, labor markets, and demographic trends? These are difficult

[5]Hopkins' empirical concerns would trigger a response from James Fenske, then a young Economics PhD graduate from Yale University. Fenske cited the interest in African economic history that Nunn's work had generated, noting that these studies are "not distinguished by their broad theories, but by their careful focus on causal inference"(Fenske 2010, p. 177). Both Hopkins and Morten Jerven responded, with another response by Fenske (Hopkins 2011; Jerven 2011; Fenske 2011).

questions to answer, of course, because, as Heldring and Robinson (2012, p. 4) observe, "we have to think about what the trajectories of African societies would have been in the absence of colonialism." Such counter-factual thinking requires us to be precise about our assumptions and honest about the potential biases. This is where the rigor of cliometrics can be of great value.

To elucidate these two questions, we first consider the effect of missionaries, take a detour to preindustrial South Africa, and then consider the Scramble for Africa and colonial era. Christian missionaries brought profound changes to African societies.[6] New religious beliefs are, of course, the most obvious change. Nunn (2010) shows that Africans who inhabit regions where European missionaries settled are today more likely to be Christian. In other words, missionaries seem to have had the effect that was, ostensibly, their main intention – to convert souls. But missionaries also brought education. Gallego and Woodberry (2010) use regional data for 180 provinces in African countries to show that Protestant mission stations in the past are more correlated with schooling variables today than similar measures for Catholic mission stations. The reason, they argue, is the increased competition between Protestant and Catholic stations within Catholic colonies. Although the overarching aim of mission stations may have been for religious conversation, Frankema (2012) uses the Blue Books to show that prior to 1940, mission stations explain nearly all of the variation in school enrollment rates for African countries. Says Frankema (2012 p 336): "Christian education was effective in leading indigenous people into the Christian faith and essential to raise the number of converts over time, because educated converts helped spread the Christian message in the local vernacular."

Missionaries not only created a demand for reading but also the ability to supply books. Cagé and Rueda (2016) build a geocoded dataset of Protestant missions and their printing investments in 1903. They show that regions that were close to mission stations with an early printing press have higher newspaper readership, trust, education, and political participation today. The concern, of course, is that missionaries are not randomly assigned across the continent. Most authors acknowledge this shortcoming but mostly assume that it is unlikely to affect their results. And even if placement was random, the mechanism of persistence remains unclear. Was it indeed early educational attainment that paved the way for current generations to have higher years of schooling and incomes, or were there are unobservables that may explain the correlation? Fourie and Swanepoel (2015) use mission stations in South Africa to argue that migration may explain much of the persistence. Mission stations attracted the best and brightest, often from afar. Once migration is controlled for, they show, the effects of early education disappears. Another concern is the use of missionary maps that are partial to European missionaries. In new work, Jedwab et al. (2018) show that a more careful analysis of mission activity in colonial Ghana

[6]Missionaries had already arrived in South Africa during the eighteenth-century, setting up stations like Genandendal aimed at converting indigenous Khoesan people, but the expansion of Christian missionaries in South Africa and throughout the continent would mostly be a late nineteenth-century phenomenon.

reverses many of the research findings when a map of only European missionaries are used. What is clear is that much more careful work is required to ascertain not only the impact of missionaries but also the mechanism through which the impact persists.

Before the Scramble for Africa at the end of the nineteenth century, Europeans had already settled the southern tip of the continent. As the Dutch East India Company expanded their reach in Asia in the seventeenth century, ship traffic around the Cape of Good Hope increased rapidly. They needed refreshments of fresh water, food, and fuel, and so the Lords XVII decided to establish a station in Table Bay that would supply passing ships. In April 1652, a motley crew of officials and workmen arrived to build a fort, farm vegetables, and trade meat with the indigenous Khoesan. The plan was poor, and the Company was soon forced to settle more land and release workmen to become free farmers. Colonization at the southern tip of Africa had begun.

Fourie (2013) uses probate inventories to ascertain the wealth of the settlers at the Cape. He finds evidence of "remarkable wealth"; the average Cape farmer owned, for example, 54 head of cattle and 350 sheep. The high level of wealth was both a consequence of demand for Cape goods (Boshoff and Fourie 2010) and low production costs, notably the acquisition of land at low cost and the use of imported slaves as farm labor. Human capital and strong property rights, in both land and slaves, also mattered (Fourie and Von Fintel 2014; Fourie and Swanepoel 2018). But access to inexpensive labor was critical to the success of the Cape economy. Malaysia, Indonesia, India, Madagascar, and Mozambique were the main places of origin for Cape slaves. Combining court records with slave records, Baten and Fourie (2015) calculate numeracy rates for the different regions of slave origin, providing an estimate of comparative living standards in the eighteenth-century Indian Ocean economies.

For most of sub-Saharan Africa, though, the colonial experience is tied to the Scramble for Africa at the end of the nineteenth century. A first, obvious question is about the timing of the Scramble. Frankema, et al. (2017) use a new trade dataset to show that nineteenth-century sub-Saharan Africa experienced a terms-of-trade boom in the five decades before the Scramble for Africa (1835–1885). Given the larger weight of West Africa in French imperial trade, the authors argue, it made economic sense for French conquest of the interior of West Africa.

Even though a more systematic process of exploration and annexation was well under way by the 1860s, it would be the Berlin conference, organized by Otto von Bismarck from November 1884 to February 1885, that would embody European colonization in Africa. In a seminal paper, Michalopoulos and Papaioannou (2016, p. 1807) investigate one consequence of the Berlin conference, the arbitrary partitioning of land: "While the Berlin conference discussed only the boundaries of Central Africa (the Congo Free State), it came to symbolize ethnic partitioning because it laid down the principles that would be used among Europeans to divide the continent. The key consideration was to preserve the status quo preventing conflict among Europeans for Africa, as the memories of the European wars of the eighteenth and nineteenth century were alive. As a result, in the overwhelming

majority of cases, European powers drew borders without taking into account local conditions." They employ this exogenous shock as a "quasi-natural" experiment to assess the impact of ethnic partitioning on civil conflict. Using a dataset that reports georeferenced incidents of political violence between 1997 and 2013, they show that the likelihood of conflict is approximately 40% higher in areas where partitioned ethnicities reside as compared to homelands of ethnicities that have not been separated by national borders. In short: the arbitrariness of the colonial partitioning helps explain some economic and political outcomes today.

It is now widely agreed that colonialism had many undesirable economic and political consequences, most notably through its effect on institutions. Acemoglu et al. (2002) famously described a "reversal of fortunes" in global incomes between 1500 and 2000 – and that the *extractive institutions* of colonialism were one reason for this reversal. But what exactly these extractive institutions were, remains the subject of debate. Econometric techniques that allow causal inference can help. As discussed, Michalopoulos and Papaionnou (2016) show an effect working through the arbitrariness of colonial borders. Acemoglu et al. (2014) use a survey of village elders combined with regression analysis to show that the distribution of ruling families in Sierra Leone, first recognized by British colonial authorities, explain development outcomes today. Lowes and Montero (2016) use a geographic regression discontinuity design along former concession boundaries in the Congo Free State to show that one specific type of extractive institution – private companies that used violent tactics to collect rubber – have persistent negative effects on education, wealth and health outcomes today. Lechler and McNamee (2017) use spatial discontinuity of colonial rule within Namibia to show the effect of direct versus indirect colonial rule on democratic participation. Archibong (2018) shows that current ethnic inequalities are the result of historical heterogeneous federal government policies towards different groups in Nigeria.

In some colonies at certain times, European powers, mostly out of self-interest, invested in physical and social infrastructure. Railroads were one such investment. Herranz-Loncán and Fourie (2017) calculate that railways in the Cape Colony can account for between 22% and 25% of the increase in the Colony's labor productivity from 1873 to 1905. Jedwab and Moradi (2016) exploit the construction and eventual demise of colonial railroads in Ghana to study how colonial infrastructure affected later economic outcomes. They show that railroads had a large effect on the distribution of economic activity during the colonial era, and that these effects, despite the railroads falling into disuse, have persisted to date. Replicating the methodology for Kenya, they show how railways determined the location of European settlers, Asian traders, and the main Kenyan cities at independence (Jedwab et al. 2017). Despite the decline of the railways, the spatial distribution of the colonial era persists. Bertazzini (2018) finds similar spatial persistence for the road network built in Ethiopia by Italians between 1935 and 1940.

Education also improved, although, as we have seen, that was mostly as a consequence of missionary activity. Huillery (2009) shows that current educational outcomes in French West Africa have been more specifically determined by colonial investments in education than health and infrastructure. This is because of the strong

persistence of investment: "regions that got more at the early colonial times continued to get more" (Huillery 2009, p. 176). Cogneau and Moradi (2014) use the partition of German Togoland after World War I as a natural experiment to test the impact of British and French colonization. Data of recruits to the Ghanaian colonial army 1908–1955 allow them to show that literacy and religious affiliation diverge at the border between the parts of Togoland under British and French control as early as in the 1920s. This, they claim, is because of policy differences towards missionary schools.

Bolt and Bezemer (2009), on the other hand, argue that the impact on education was not driven only by missionary activity but also by broad exposure to Europeans and European-style education. The find that the colonial education influenced subsequent development and that was influenced by the density of the European population. Wantchekon et al. (2014) use the random allocation of regional schools in colonial Benin to assess the impact of colonial education on the descendants of those that first attended school. They find a significant effect on the first and later generations, as well as large village-level externalities: "Descendants of the untreated in villages with schools do better than those in control villages" (Wantchekon et al. 2014, p. 705).

Colonial governments also invested in health. Lowes and Montero (2017) use medical campaigns by French colonial governments to examine the effects on health attitudes and outcomes today. They show that in places were villagers were forcibly examined, inhabitants today have less trust in medicine, and World Bank projects in the health sector are less successful.

How costly, then, was colonization for the European powers? Huillery (2014) uses France as one case study. She shows that French West Africa took only 0.29% of French annual expenditures, of which only 0.05% was for development. West Africans, instead, carried the heaviest burden, disproportionately funding French civil servants' salaries as a share of local expenditure. Gardner (2012) came to the same conclusion for most of British Africa.

Decolonizing with Data

The data and cliometric revolutions have significantly improved our knowledge of Africa's economic past. There are at least two reasons for this. First, where African histories are largely based on colonial documentation and where such scholarship is often undertaken by non-Africans, the fear is that these histories could suffer from the implicit biases of both the source material and the researcher. Quantitative records often used for purposes orthogonal to its intended reason for collection – suffer less from such biases. Second, in areas where both quantitative and qualitative source material is weak or completely missing, economic historians have uncovered and used innovative alternatives. For example, climate data going back into the distant past can help explain the slave trade (Fenske and Kala 2017), or tree rings in Zimbabwe may help show how an Indonesian volcano caused a period of tribal warfare in early nineteenth-century Southern Africa (Hannaford and Nash 2016).

Such quantitative records help to "decolonize" African economic histories often distorted by the imprints of the colonial regime.

But a second aspect of "decolonization" is to encourage greater participation of scholars from the continent. African cliometrics is mostly a non-African field. To show the relative shortage of African scholars, we use data from ISI Web of Science (WoS) and Elsevier's SCOPUS database to conduct a simple bibliographic exercise. We make use of WoS and SCOPUS for the simple reason that they include information about the attributes of both the authors and the papers they have published in accredited scholarly journals.[7]

Because Economics journals also frequently publish Economic History papers, we take advantage of the available information on author names, paper titles, abstracts, and keywords as organized in WoS and SCOPUS to classify Economic History (EH) papers separately from mainstream Economics papers. To accomplish this, we build up a database of EH papers that include the words "economic history" or "history" in their title, keywords, or abstract, from a sample of journals. The rest we classify as ECON papers. We then use a random 70% (5643) of the EH and ECON papers to train a Support Vector Machine (SVM) machine learning algorithm on the words used in the respective titles, abstracts and keywords of the two groups. We use the remaining random 30% (2419) of the papers to test the prediction accuracy of the resulting algorithm. Below is the Confusion Matrix from the test indicating a 98% prediction accuracy for ECON papers with a 2% confusion for EH papers and 96% prediction accuracy for EH papers with 4% confusion for ECON papers (Table 1).

We then apply the algorithm to the 49,444 papers in a database of 17 Economic History journals[8] and the top 25 Economics journals published since 1992, and classify them as either ECON or EH. We obtain 18,835 EH papers with a 96% accuracy. We then write a function that selects from this list all the papers that mention Africa or any of the current or historical names of African countries in their titles, keywords, or abstracts. This leaves us with a list of 238 papers published in both Economics and Economic History journals. It is from this list that we compile

[7]In parallel to these two databases on research output, Google Scholar has risen to prominence lately. Unlike WoS and SCOPUS, Google Scholar gathers information about any published document – even those published as Working Papers. It therefore has a more extensive list of citations. It follows that the number of citations accounted for in this analysis will be significantly less than what may appear in a Google Scholar search.

[8]These are *Economic History Review, Journal of Economic History, European Review of Economic History, Explorations in Economic History, Cliometrica, Economic History of Developing Regions, South African Journal of Economic History, Australian Economic History Review, African Economic History, Scandinavian Economic History Review, Low Countries Journal of Social and Economic History, Revista de Historia Economica, The Indian Economic and Social History Review, Journal of European Economic History, Revista di Storia Economica,* and *Research in Economic History.*

Table 1 Confusion matrix

ECON		EH	TOTAL	ECON	EH
ECON	1269	30	1299	98%	2%
EH	44	1076	1129	4%	96%
	1313	1106	2419		

Table 2 Euclidean index ranking for economic history scholars on "Africa"

	Author	H-index	E-index	G-index	Cit.	Pub	Country
1	Nunn, N.	5	167.2	7	337	7	USA
2	Austin, G.	5	94.7	7	171	7	UK
3	Williamson, J.	4	88.7	6	113	6	USA
4	Huillery, E.	3	68.3	3	102	3	France
5	Frankema, E.	7	60.9	9	156	9	Netherlands
6	Richardson, D.	3	41.3	3	71	3	USA
7	Eltis, D.	3	37.2	3	60	3	USA
8	Baten, J.	4	35.5	4	58	4	Germany
9	Robinson, J	2	32.1	3	34	3	USA
10	Bates, R.	2	30.1	4	32	4	USA
11	Shatzmiller, M.	3	28.3	3	417	3	Canada
12	Allen, R.	2	27.5	2	32	2	UAE
13	Lewis, F.	2	25.6	2	33	2	USA
14	Fourie, J.	3	24.8	4	43	4	South Africa
15	Von Fintel, D.	3	22.8	4	36	4	South Africa
16	Moradi, A.	2	22.8	2	30	2	United Kingdom
17	Pamuk, S.	2	18.0	2	25	2	Turkey
18	Fenske, J.	4	14.9	5	35	10	UK
19	Jerven, M.	3	13.7	3	19	3	Norway
20	Van Leeuwen, B	2	10.6	2	15	2	Netherlands

an H-index and an Euclidean index of the top economic historians working on Africa.

It is important to note that our method is not perfect. Nunn and Wantchekon's (2011) paper on slavery and trust is, for example, classified by our algorithm as ECON rather than EH, despite the clear relevance to Economic History.[9] Despite these concerns, though, we believe that the analysis provides a fair reflection of the state of the field.

Table 2 reports the results for the top 20 authors in African economic history, ranked according to their Euclidean index (Perry and Reny 2016). Several trends are immediately apparent. Besides two authors affiliated with Stellenbosch University in

[9]We tried several versions of the algorithm, but because the word "History" is not included in either their title, abstract, or keywords, our algorithm fails to classify the paper as EH. For future work, it might be useful to include historical topics – like the Atlantic slave trade, or colonialism – as part of the training algorithm.

South Africa, the leading African economic historians are based outside the continent. Of the 18 scholars based outside, not one is African. This trend is mirrored in the citations of this chapter, with only a handful of African authors cited for their cliometric contributions. Fourie (2016) notes two ways to address this clear imbalance: First, more should be done to recruit African scholars to good PhD programs, notably those in the USA and Europe where most of the leading scholars are based. Second, more should be done to appoint qualified African scholars in these scholars' research programs – as postdocs or tenure-track faculty. Large research programs on African economic history funded by European or US donors still frequently lack African participants. Building stronger networks with African universities can help speed this process (Green and Nyambara 2015; Austin 2015).

Conclusion

The study of African economic history has received a much-needed resuscitation from the cliometric turn in economic history. The combination of innovative statistical techniques that allow for causal analysis, access to larger and more reliable data sets, a growing interest in the histories of developing regions, and the rising fortunes of many African economies have spurred interest in Africa's fluctuating fortunes, past and present. This chapter has provided an overview of the most important contributions of the last two decades.

More can be done, though, to attract African scholars into the field. The good news is that this seems to be happening. More than half of participants at the African Economic History Network Meetings in 2017 were African, many of them Masters or PhD students. A free textbook project, coordinated by Ewout Frankema, Ellen Hillbom, Ushehwedu Kufakurinani, and Felix Meier zu Selhausen is one attempt to expose younger scholars to African economic history. As African countries become wealthier and funding for tertiary education improves, the study of African economic history is likely to gain popularity. The future of African cliometrics hinges on the ability of the field to draw these young scholars into their networks and equip them with the scientific tools and academic freedom to explore the economic histories of their own continent.

References

Acemoglu D, Robinson JA (2010) Why is Africa poor? Econ Hist Dev Reg 25(1):21–50
Acemoglu D, Johnson S, Robinson JA (2002) Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q J Econ 117(4):1231–1294
Acemoglu D, Reed T, Robinson JA (2014) Chiefs: economic development and elite control of civil society in Sierra Leone. J Polit Econ 122(2):319–368
Allen RC, Bassino J-P, Ma D, Moll-Murata C, van Zanden JL (2011) Wages, prices, and living standards in China, 1738–1925: in comparison with Europe, Japan, and India. Econ Hist Rev 64 (s1):8–38
Alsan M (2014) The effect of the tsetse fly on African development. Am Econ Rev 105(1):382–410
Angeles L (2013) On the causes of the African slave trade. Kyklos 91:1–26

Archibong B (2018). Historical origins of persistent inequality in Nigeria. Oxford Dev Stud 46 (3):325–347
Ashraf Q, Galor O (2013) The out of Africa hypothesis, human genetic diversity, and comparative economic development. Am Econ Rev 103(1):1–46
Asongu SA, Kodila-Tedika O (2017) Is poverty in the African DNA (gene)? South Afr J Econ 85 (4):533–552
Austin G (2008) The reversal of fortune thesis and the compression of history: perspectives from African and comparative economic history. J Int Dev 20(8):996–1027
Austin G (2015) African economic history in Africa. Econ Hist Dev Reg 30(1):79–94
Austin G, Broadberry S (2014) Introduction: the renaissance of African economic history. Econ Hist Rev 67(4):893–906
Baten J, Blum M (2012) Growing tall but unequal: new findings and new background evidence on anthropometric welfare in 156 countries, 1810–1989. Econ Hist Dev Reg 27(sup1):S66–S85
Baten J, Fourie J (2015) Numeracy of Africans, Asians, and Europeans during the early modern period: new evidence from Cape Colony court registers. Econ Hist Rev 68(2):632–656
Bertazzini MC (2018) The long-term impact of Italian colonial roads in the Horn of Africa, 1935–2000. In: Economic History Working Papers No: 272/2018. London School of Economics, London
Bertocchi G (2016) The legacies of slavery in and out of Africa. IZA J Migr 5(1):24
Bertocchi G, Dimico A (2014) Slavery, education, and inequality. Eur Econ Rev 70:197–209
Bolt J, Bezemer D (2009) Understanding long-run African growth: colonial institutions or colonial education? J Dev Stud 45(1):24–54
Bolt J, van Zanden JL (2014) The Maddison project: collaborative research on historical national accounts. Econ Hist Rev 67(3):627–651
Boshoff WH, Fourie J (2010) The significance of the Cape trade route to economic activity in the Cape Colony: a medium-term business cycle analysis. Eur Rev Econ Hist 14(3):469–503
Bossuroy T, Cogneau D (2013) Social mobility in five African countries. Rev Income Wealth 59: S84–S110
Bowden S, Chiripanhura B, Mosley P (2008) Measuring and explaining poverty in six African countries: a long-period approach. J Int Dev 20(8):1049–1079
Boxell L (2017) Droughts, conflict, and the African slave trade. In: MPRA Paper No. 81924. Stanford University. Available online: https://mpra.ub.uni-muenchen.de/81924/
Cagé J, Rueda V (2016) The long-term effects of the printing press in Sub-Saharan Africa. Am Econ J Appl Econ 8(3):69–99
Cilliers J, Fourie J (2017) Occupational mobility during South Africa's industrial take-off. S Afr J Econ 86:3–22
Cogneau D, Moradi A (2014) Borders that divide: education and religion in Ghana and Togo since colonial times. J Econ Hist 74(3):694–729
Collier P, Gunning JW (1999) Explaining African economic performance. J Econ Lit 37(1):64–111
Dalrymple-Smith A, Frankema E (2017) Slave ship provisioning in the long 18th century. A boost to West African commercial agriculture? Eur Rev Econ Hist 21(2):185–235
Dalton JT, Leung TC (2014) Why is polygyny more prevalent in Western Africa? An African slave trade perspective. Econ Dev Cult Chang 62(4):599–632
Dalton JT, Leung TC (2015) Dispersion and distortions in the trans-Atlantic slave trade. J Int Econ 96(2):412–425
De Haas M (2017) Measuring rural welfare in colonial Africa: did Uganda's smallholders thrive? Econ Hist Rev 70(2):605–631
De Zwart P (2011) South African living standards in global perspective, 1835–1910. Econ Hist Dev Reg 26(1):49–74
Deconinck K, Verpoorten M (2013) Narrow and scientific replication of the slave trade and the origins of mistrust in Africa. J Appl Econ 28(1):166–169
Desmet K, Ortuño-Ortín I, Wacziarg R (2017) Culture, ethnicity, and diversity. Am Econ Rev 107 (9):2479–2513

Du Plessis S, Du Plessis S (2012) Happy in the service of the company: the purchasing power of VOC salaries at the cape in the 18th century. Econ Hist Dev Reg 27(1):125–149
Eltis D (1977) The export of slaves from Africa, 1821–1843. J Econ Hist 37(2):409–433
Eltis D (1987) Economic growth and the ending of the transatlantic slave trade. Oxford University Press, New York
Eltis D, Lewis FD, Richardson D (2005) Slave prices, the African slave trade, and productivity in the Caribbean, 1674–1807. Econ Hist Rev 58(4):673–700
Fenske J (2010) The causal history of Africa: a response to Hopkins. Econ Hist Devel Reg 25 (2):177–212
Fenske J (2011) The causal history of Africa: replies to Jerven and Hopkins: debate. Econ Hist Dev Reg 26(2):125–131
Fenske J, Kala N (2017) 1807: economic shocks, conflict and the slave trade. J Dev Econ 126:66–76
Fenske J, Kala N (2015) Climate and the slave trade. J Dev Econ 112:19–32
Fourie J (2013) The remarkable wealth of the Dutch Cape Colony: measurements from eighteenth-century probate inventories. Econ Hist Rev 66(2):419–448
Fourie J (2016) The data revolution in African economic history. J Interdiscip Hist 47:193–212
Fourie J, Green E (2015) The missing people: accounting for the productivity of indigenous populations in Cape colonial history. J Afr Hist 56(2):195–215
Fourie J, Swanepoel C (2015) When selection trumps persistence: the lasting effect of missionary education in South Africa 1. Tijdschr Soc Econ Geschiedenis 12(1):1
Fourie J, Swanepoel C (2018) Impending ruin or remarkable wealth? The role of private credit markets in the 18th-century Cape Colony. J South Afr Stud 44(1):7–25
Fourie J, van Zanden JL (2013) GDP in the Dutch Cape Colony: the national accounts of a slave-based society. S Afr J Econ 81(4):467–490
Fourie J, von Fintel D (2014) Settler skills and colonial development: the Huguenot wine-makers in eighteenth-century Dutch South Africa. Econ Hist Rev 67(4):932–963
Frankema EHP (2012) The origins of formal education in sub-Saharan Africa: was British rule more benign? Eur Rev Econ Hist 16(4):335–355
Frankema E, Jerven M (2014) Writing history backwards or sideways: towards a consensus on African population, 1850–2010. Econ Hist Rev 67(4):907–931
Frankema E, Van Waijenburg M (2012) Structural impediments to African growth? New evidence from real wages in British Africa, 1880–1965. J Econ Hist 72(4):895–926
Frankema E, Williamson J, Woltjer P (2017) An economic rationale for the west African scramble? The commercial transition and the commodity price boom of 1835–1885. J Econ Hist 78 (2):1–45
Gallego FA, Woodberry R (2010) Christian missionaries and education in former African colonies: how competition mattered. J Afr Econ 19(3):294–329
Gardner L (2012) Taxing colonial Africa: the political economy of British imperialism. Oxford University Press, Oxford
Gennaioli N, Rainer I (2007) The modern impact of precolonial centralization in Africa. J Econ Growth 12(3):185–234
Gören E (2017) The persistent effects of novelty-seeking traits on comparative economic development. J Dev Econ 126:112–126
Gorodnichenko Y, Roland G (2017) Culture, institutions, and the wealth of nations. Rev Econ Stat 99(3):402–416
Green E, Nyambara P (2015) The internationalization of economic history: perspectives from the African frontier. Econ Hist Dev Reg 30(1):68–78
Guedes J d'A, Bestor T, Carrasco D, Flad R, Fosse E, Herzfeld M, Lamberg-Karlovsky C, Lewis C, Liebmann M, Meadow R, Patterson N, Price M, Reiches M, Richardson S, Shattuck-Heidorn H, Ur J, Urton G, Warinner C (2013) "Is poverty in our genes? A critique of Ashraf and Galor," the 'out of Africa'-hypothesis, human genetic diversity, and comparative economic development., American Economic Review (Forthcoming). Curr Anthropol 54(1):71–79

Hannaford MJ, Nash DJ (2016) Climate, history, society over the last millennium in Southeast Africa. Wiley Interdiscip Rev Clim Chang 7(3):370–392
Heldring L, Robinson JA (2012) Colonialism and economic development in Africa. Tech. rep. National Bureau of Economic Research, Cambridge, MA
Herranz-Loncán, A, Fourie J (2017) For the public benefit? Railways in the British Cape Colony. Eur Rev Econ Hist 22(1):73–100
Hopkins AG (2009) The new economic history of Africa. J Afr Hist 50(2):155–177
Hopkins AG (2011) Causes and confusions in African history. Econ Hist Dev Reg 26(2):107–110
Huillery E (2009) History matters: the long-term impact of colonial public investments in French West Africa. Am Econ J Appl Econ 1(2):176–215
Huillery E (2014) The black man's burden: the cost of colonization of French West Africa. J Econ Hist 74(1):1–38
Inikori JE (1976) Measuring the Atlantic slave trade: an assessment of Curtin and Anstey. J Afr Hist 17(2):197–223
Inklaar R, de Jong H, Bolt J, van Zanden JL (2018) Rebasing 'Maddison': new income comparisons and the shape of long-run economic development. Tech. rep. Groningen Growth and Development Centre, University of Groningen
Jedwab R, Moradi A (2016) The permanent effects of transportation revolutions in poor countries: evidence from Africa. Rev Econ Stat 98(2):268–284
Jedwab R, Kerby E, Moradi A (2017) History, path dependence and development: evidence from colonial railways, settlers and cities in Kenya. Econ J 127(603):1467–1494
Jedwab R, Meier zu Selhausen F, Moradi A (2018). The economics of missionary expansion and the compression of history. Centre for Studies of African Economies Working Paper 2018–07
Jerven M (2010) African growth recurring: an economic history perspective on African growth episodes, 1690–2010. Econ Hist Dev Reg 25(2):127–154
Jerven M (2011) A clash of disciplines? Economists and historians approaching the African past. Econ Hist Dev Reg 26(2):111–124
Jerven M (2013) Poor numbers: how we are misled by African development statistics and what to do about it. Cornell University Press, Ithaca
Jerven M (2018) The history of African poverty by numbers: evidence and vantage points. J Afr Hist 59(2):449–461
Lechler M, McNamee L (2017) Decentralized Despotism? Indirect colonial rule undermines contemporary democratic attitudes. Tech. rep. Munich Discussion Paper
Levine R, Lin C, Xie W (2017) The origins of financial development: how the African slave trade continues to influence modern finance. Tech. rep. National Bureau of Economic Research, Cambridge, MA
Lowes S, Montero E (2016) Blood rubber: the effects of labor coercion on institutions and culture in the DRC. Tech. rep. Working paper, Harvard, http://www.saralowes.com/research.html
Lowes S, Montero E (2017) Mistrust in medicine: the legacy of colonial medical campaigns in Central Africa. Harvard University, Mimeo
Lowes S, Nunn N, Robinson JA, Weigel JL (2017) The evolution of culture and institutions: evidence from the Kuba kingdom. Econometrica 85(4):1065–1091
Manning P (1990) Slavery and African life: occidental, oriental, and African slave trades, vol 67. Cambridge University Press, New York
Manning P (2010) African population: projections, 1850–1960. In: Ittmann, K., Cordell, D.D. and Maddox, G.H. eds., The demographics of empire: the colonial order and the creation of knowledge. Ohio University Press: Athens, USA, pp. 245–275
Meier zu Selhausen F (2014) Missionaries and female empowerment in colonial Uganda: new evidence from Protestant marriage registers, 1880–1945. Econ Hist Dev Reg 29(1):74–112
Meier zu Selhausen F, Weisdorf J (2016) A colonial legacy of African gender inequality? Evidence from Christian Kampala, 1895–2011. Econ Hist Rev 69(1):229–257

Meier zu Selhausen F, Van Leeuwen MHD, Weisdorf JL (2017) Social mobility among Christian Africans: evidence from Anglican marriage registers in Uganda, 1895–2011. Econ Hist Rev 74:1291–1321
Michalopoulos S, Papaioannou E (2013) Pre-colonial ethnic institutions and contemporary African development. Econometrica 81(1):113–152
Michalopoulos S, Papaioannou E (2016) The long-run effects of the scramble for Africa. Am Econ Rev 106(7):1802–1848
Michalopoulos S, Putterman L, Weil DN (2016) The Influence of ancestral lifeways on individual economic outcomes in sub-Saharan Africa. Tech. rep. National Bureau of Economic Research, Cambridge, MA
Moradi A (2009) Towards an objective account of nutrition and health in colonial Kenya: a study of stature in African army recruits and civilians, 1880–1980. J Econ Hist 69(3):719–754
Moradi A (2010) Nutritional status and economic development in sub-Saharan Africa, 1950–1980. Econ Hum Biol 8(1):16–29
Moradi A, Baten J (2005) Inequality in Sub-Saharan Africa: new data and new insights from anthropometric estimates. World Dev 33(8):1233–1265
Mpeta B, Fourie J, Inwood K (2018) Black living standards in South Africa before democracy: new evidence from height. S Afr J Sci 114(1/2):8–8
Nunn N (2008) The long-term effects of Africa's slave trades. Q J Econ 123(1):139–176
Nunn N (2010) Religious conversion in colonial Africa. Am Econ Rev 100(2):147–152
Nunn N, Puga D (2012) Ruggedness: the blessing of bad geography in Africa. Rev Econ Stat 94 (1):20–36
Nunn N, Wantchekon L (2011) The slave trade and the origins of mistrust in Africa. Am Econ Rev 101(7):3221–3252
Obikili N (2015) The impact of the slave trade on literacy in West Africa: evidence from the colonial era. J Afr Econ 25(1):1–27
Obikili N (2016) The trans-Atlantic slave trade and local political fragmentation in Africa. Econ Hist Rev 69(4):1157–1177
Perry M, Reny PJ (2016) How to count citations if you must. Am Econ Rev 106(9):2722–2741
Pierce L, Snyder JA (2017) The historical slave trade and firm access to finance in Africa. Rev Financ Stud 31(1):142–174
Pison G (2017). Theres a strong chance that one-third of all people will be African by 2100. The Conversation. Available at: https://theconversation.com/theres-a-strong-chance-that-one-third-of-all-people-will-be-african-by-2100-84576. Accessed 4 Jan 2018
Reid R (2011) Past and presentism: the precolonial and the foreshortening of African history. J Afr Hist 52(2):135–155
Ronnback K (2015) The transatlantic slave trade and social stratification on the Gold Coast. Econ Hist Dev Reg 30(2):157–181
Steckel RH (1979) Slave height profiles from coastwise manifests. Explor Econ Hist 16(4):363–380
Steckel RH (1995) Stature and the standard of living. J Econ Lit 33(4):1903–1940
The World Bank (2017) GDP per capita (constant 2010 US$). Data retrieved from World Development Indicators, http://databank.worldbank.org/data
Wantchekon L, Klaˇsnja M, Novta N (2014) Education and human capital externalities: evidence from colonial Benin. Q J Econ 130(2):703–757
Whatley WC (2017) The gun-slave hypothesis and the 18th century British slave trade. Explor Econ Hist 67:80–104

Index

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
https://doi.org/10.1007/978-3-030-00181-0

H