Claude Diebolt
Michael Haupert
Editors

Handbook of Cliometrics

Springer Reference

Handbook of Cliometrics

Claude Diebolt • Michael Haupert
Editors

Handbook of Cliometrics

With 59 Figures and 20 Tables

Springer Reference

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 978-3-642-40405-4 ISBN 978-3-642-40406-1 (eBook)
ISBN 978-3-642-40407-8 (print and electronic bundle)
DOI 10.1007/978-3-642-40406-1

Library of Congress Control Number: 2015943062

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media (www.springer.com)

An Introduction to the Handbook of Cliometrics

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 microfoundations, 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 a handbook such as this one, the most difficult question is what to include. The possibilities were endless, but the space was limited. Topics that are not included are not by any means considered to be lacking in importance or historical significance. We simply had to make difficult choices, and in the end, we decided that variety over time, topic, and geography would be our goal. The final selection of chapters represents a sampling of the topics that cliometrics has helped to transform over the past half century. It ranges from those that have long been at the center of cliometric analysis, such as Greg Clark's chapter on the industrial revolution and Larry Neal's chapter on financial markets, to 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 and Thomas Rahlf's contribution on statistical inference. In between, we have included articles by Peter Temin and Stanley Engerman, 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 cliometicians, like Matt Jaremski and Emanuele Felice. The common link in the chapters 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 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 7 sections, grouping the 22 contributions by general topic, starting with 2 chapters on the history of economic history and cliometrics. The first is Michael Haupert's brief overview of the evolution of economic history, highlighting the literature in the history of the discipline that begot cliometrics and

shaped its development. Peter Temin also looks at the past to explain the present state of cliometrics and then goes further, using recent contributions to the literature to make some predictions about the future of cliometrics as a discipline, highlighting the ways in which economic history and economic development benefit from their interaction. Both of these articles emphasize the growth of the influence of cliometrics on the field of economic history.

The second section focuses on human capital, beginning with broad topical coverage by Claudia Goldin, who focuses on institutions that encourage investment in human capital. In particular, she looks at two major components of human capital: education and health. Robert Margo provides an extensive review of clio's contributions to historical labor markets, using the United States as his backdrop. Lee Craig's essay on standards of living highlights some of the lessons we have learned from merging cliometrics with the fields of demography, biology, and nutrition. This is followed by two specific surveys of the role of clio in age heaping and church book registry. Franziska Tollnek and Joerg Baten show how age heaping has been used to shed light on topics as diverse as education, gender gaps, and cross-country differentials in long-run growth. Jacob Weisdorf looks at how church book registries have been used to look at similar questions.

Section 3 takes the big picture into consideration with five papers on economic growth. We begin with an essay on growth theories and the contribution of cliometrics to them by Claude Diebolt and Faustine Perrin, followed by Greg Clark's look at the industrial revolution. Clark looks at the change in productivity growth rates and the impact of institutions and human capital on growth. James Foreman-Peck surveys the cliometric models used to explain the demographic transition that led to modern economic growth, and Emanuele Felice discusses historical estimates of GDP. Finally, Markus Lampe and Paul Sharp look at the contributions of clio to international trade.

Section 4 focuses on financial markets. Larry Neal enumerates three reasons why cliometricians study financial markets. These include an increased appreciation for the role they play in growth and development, the availability of vast data sets, and improved analytical techniques. John James contributes an essay on payment systems, noting the similarities between their development and the evolution of economic institutions. He focuses on the role of cliometrics in building applicable data sets and analyzing them in novel ways, such as network modeling. Matt Jaremski provides an exhaustive survey of empirical approaches to the study of financial panics, while Caroline Fohlin examines the role of financial systems in economic development.

In Section 5, Jochen Streb and Stanley Engerman and Nathan Rosenberg offer two takes on the role of clio in our study of the history of innovation. Engerman and Rosenberg stress that since theoretical models cannot deal with the full complexity of the process of invention and innovation, some historical study is necessary to develop a full understanding of these processes. Streb discusses the cliometric impact on the use of patent statistics to model invention and innovation.

The section on statistics and business cycles has essays by Thomas Rahlf on statistical inference and Terence Mills on the use of cliometrics to study trends,

cycles, and structural breaks. Rahlf gives a historical overview of the emergence of concepts of particular interest to cliometricians. Mills looks at the evolution of methods of calculating trends and growth rates.

Finally, we turn to the role of government. Price Fishback looks at the New Deal and how its effects on the Depression have been modeled. Jari Eloranta takes a broader view of the role of government – in scope, time period, and geography – when he looks at war and the many ways it has been analyzed by economists, historians, and sociologists, among others.

We enjoyed the process of putting the handbook together. What began as an innocent query (Why isn't there a handbook of cliometrics?) grew into a final project that we are excited to share with you. The process was long and exhausting but worthwhile. 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.

Shortly before the final touches were put on the handbook, we were deeply saddened by the sudden and unexpected loss of one of our contributors. John James was one of the first to finish a chapter, enthusiastically agreeing to our request and delivering his essay on payment systems a few months later. Sadly, he passed away on November 28, 2014. For all those who knew John, his enthusiasm, scholarship, and dedication to the project were predictable. His loss will be felt by all of us. In his memory, we dedicate this handbook to cliometricians everywhere.

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

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

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

About the Editors

Claude Diebolt is a CNRS research professor of economics and a fellow of the University of Strasbourg Institute for Advanced Study. His research focuses on the cliometrics of growth and economic cycles. He is the founder, the publication director, and the managing editor of the journal *Cliometrica*. He is also the founding president of the Association Française de Cliométrie, the chairman of the board of trustees of the Cliometric Society, and the Economic History Association representative at the International Economic History Association. Claude Diebolt's research has been published in more than 30 books and in around 125 academic journals such as the *American Economic Review*, the *Journal of Monetary Economics*, the *Journal of Macroeconomics*, *Explorations in Economic History*, the *European Journal of the History of Economic Thought*, etc. His work has been translated into Chinese, Czech, English, German, and Spanish.

Michael Haupert is Professor of Economics at the University of Wisconsin-La Crosse and Executive Director of 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 previously served as the editor of the newsletter of the Cliometric Society and currently edits the newsletter of the Economic History Association, for which he has begun writing a series of biographies of past presidents of the Association. He has authored two books on the history of the entertainment industry in America and has published more than 100 articles in journals such as *Cliometrica*, *The Journal of Economic History*, *The Journal of Money, Credit, and Banking*, and *The Journal of Economic Education*.

Contributors

Joerg Baten University of Tuebingen and CESifo, Tuebingen, Germany

Gregory Clark University of California, Davis, CA, USA

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

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

Jari Eloranta Appalachian State University and University of Jyvaskyla, Boone, NC, USA

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

Emanuele Felice Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Barcelona, Spain

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

Caroline Fohlin Johns Hopkins University, Baltimore, MD, USA and Emory University, Atlanta, GA, USA

James Foreman-Peck Cardiff University, Cardiff, UK

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

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

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

Matthew Jaremski Colgate University and NBER, NY, USA

Markus Lampe Universidad Carlos III Madrid, Madrid, Spain

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

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

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

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

Thomas Rahlf German Research Foundation, Bonn, Germany

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

Paul Sharp University of Southern Denmark, Odense M, Denmark

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

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

Franziska Tollnek University of Tuebingen, Tuebingen, Germany

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

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*,
DOI 10.1007/978-3-642-40406-1_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

[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.

more about 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.

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

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

[15]Weintraub (2002), p. 21

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.

After WWI, economists became less theoretical and more statistical in their approach. The creation of the National Bureau of Economic Research, which is

[16]See Tribe (2000).

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

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

[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.

nearly 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

[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.

was a "science," constructing, testing, and applying technically sophisticated models. Econometrics was on the rise and economic historians were divided 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,

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

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

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.

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

[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).

observations (as the Newmarch data set) on 1945 punch cards, but the computer then did all of the computational work.

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

[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.

the transaction costs and the greater the drag on economic growth and development. 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

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

Peter Temin

Contents

Abstract

I argue in this paper 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

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*,
DOI 10.1007/978-3-642-40406-1_28

as Cliometrics, and a non-random survey of recent research emphasizing the emerging methodology of the New Economic History.

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

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*,
DOI 10.1007/978-3-642-40406-1_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 1987.[1] It was later adopted by various writers but did not become a 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

[1]Fisher cites J.S. Nicholson, "The Living Capital of the United Kingdom," for the term "living capital" as opposed to "dead capital."

[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).

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 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

[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.

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 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

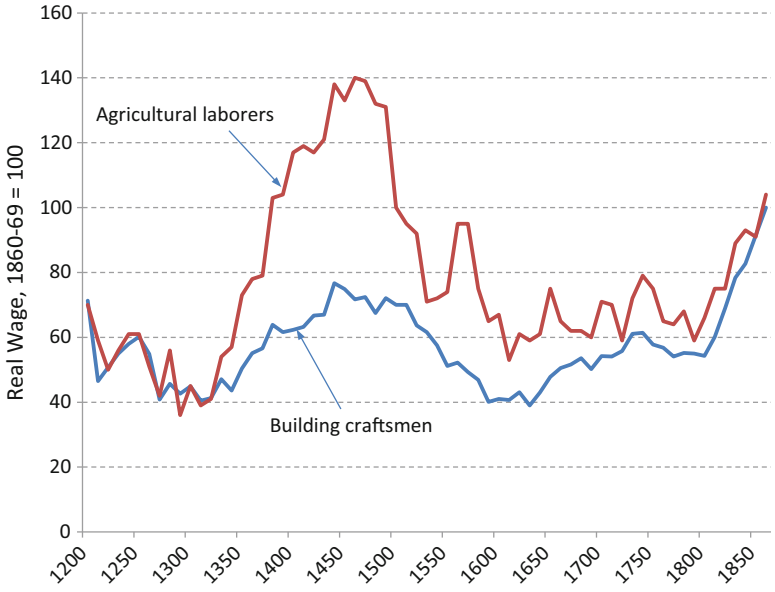

Fig. 1 English laborer's real wages, 1209–1865 (Sources and Notes: Clark (2007b), Table A2 for agricultural laborer real day wages and Clark (2005), Tables A2 for building craftsman real wages. Both series set 1860–69 = 100. The building craftsman series is for decades, the midpoint of which is used 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)

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 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.

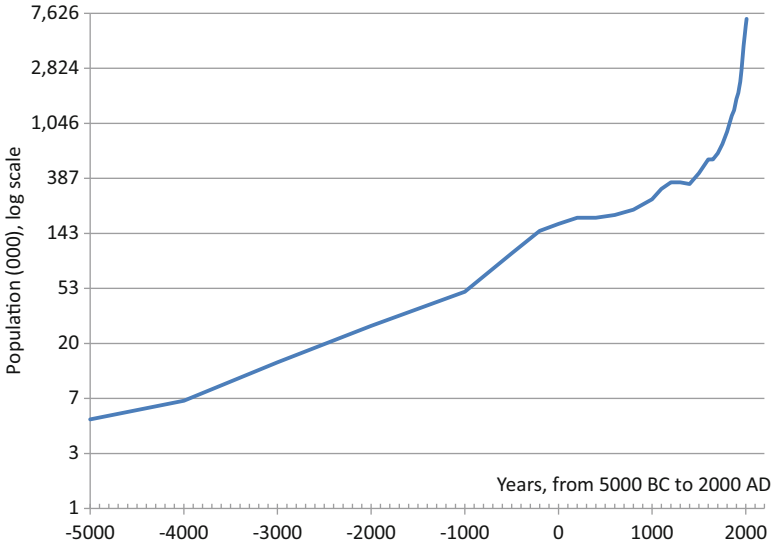

Fig. 2 World population (000) from the Neolithic era to 2010 (Source: Kremer (1993), Table 1 and recent world population estimates after 1980)

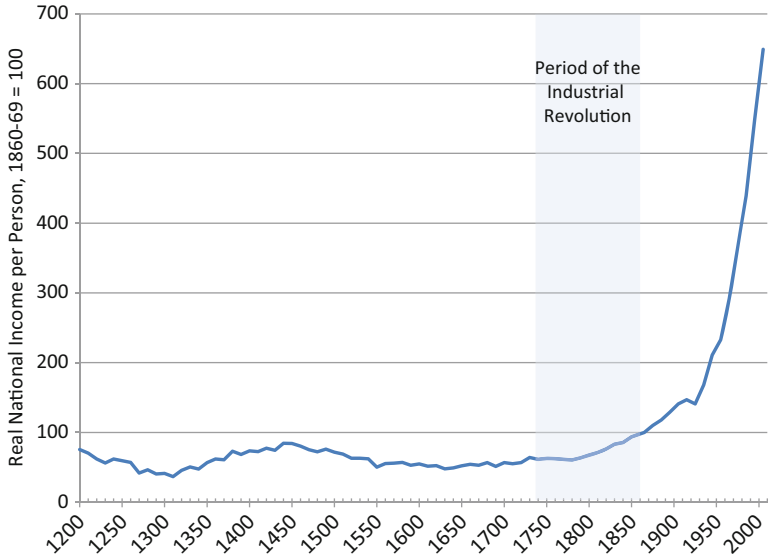

Fig. 3 Real national income per person, England: 1200–2000 (1860–69 = 100) (Source and Notes: Clark (2009), Table 28, column labeled Real National Income/N (PNDP) for 1200–1850, and Table 34 for 1860–2000, where the midpoint of the decade given is graphed. The approximate dating of the industrial revolution is from Clark (2007a), Fig. 10.2)

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 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. 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,

[9]Mark Twain provides a similar answer in *A Connecticut Yankee in King Arthur's Court*.

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 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.

[10]Fogel (1989) provides a definitive treatment of the subject.

[11]See Goldin (1976) on slavery in US cities from 1820 to 1860.

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 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

[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.

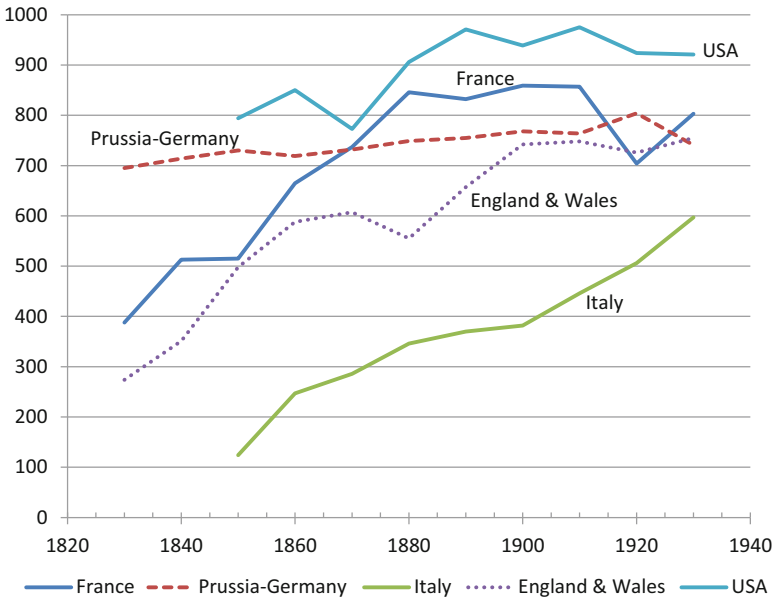

Fig. 4 Public and private primary school students per thousand 5- to 14-year-olds (Sources: All data other than United States, 1850–1870: Lindert (2004a), Table 5.1; Lindert (2004b), Table A1 contains the raw data for the numerator. United States 1850–1870: Carter et al. (2006) Bc 438–446 for enrollment 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)

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. 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.

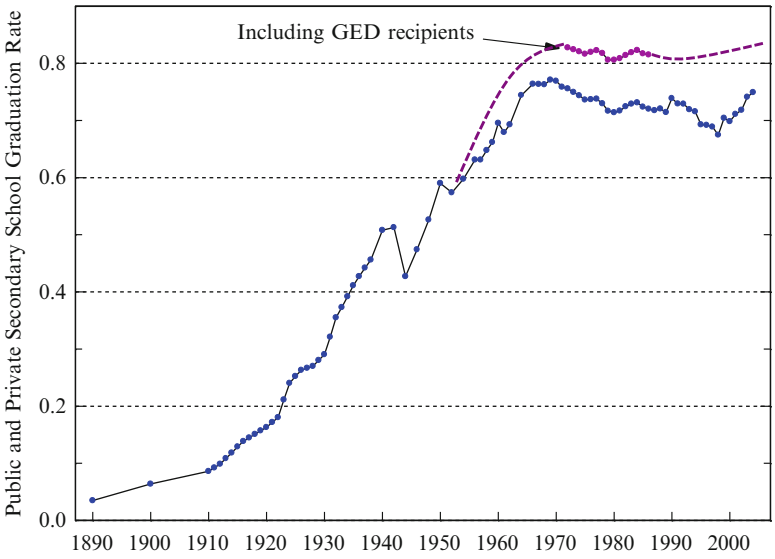

Fig. 5 Public and private secondary school enrollment and graduation rates: United States, 1890–2005 (Source: Goldin and Katz (2008), Fig. 9.2)

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 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

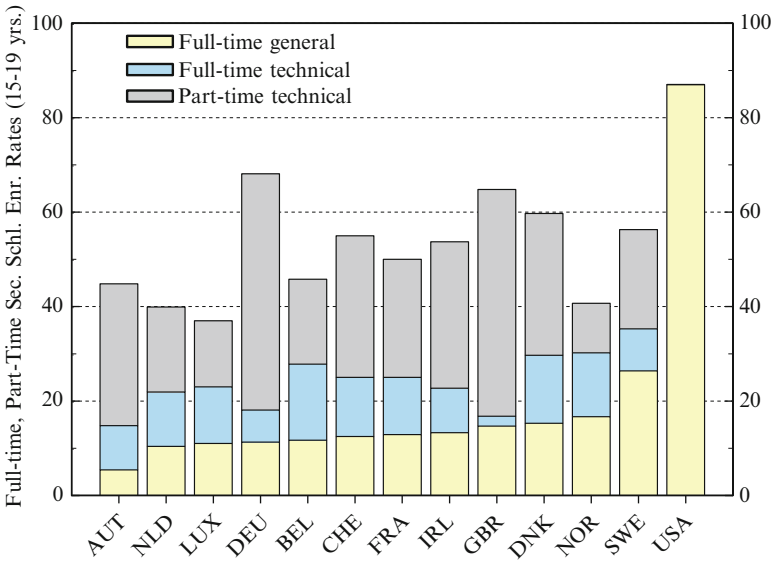

Fig. 6 Secondary school enrollment rates: OECD, 1955/56 (Source: Goldin and Katz (2008), Fig. 1.7)

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

[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), Chapters 3 and 4.

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 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.

[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.

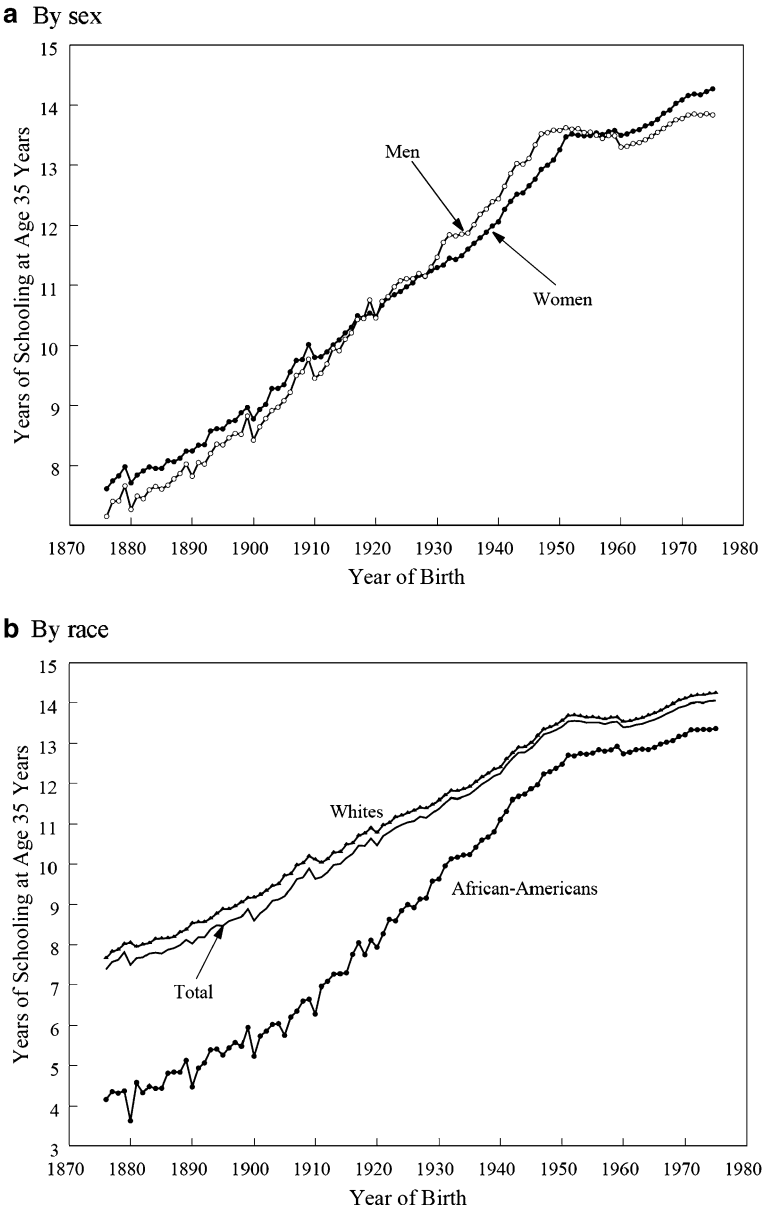

Fig. 7 Years of schooling at age 35 years by year of birth, US cohorts born 1875–1975 (Source: Goldin and Katz (2008), Figs. 1.5 and 1.6)

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 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).

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)

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

[15]Many places have added two other transitions: preschool to kindergarten and middle school or junior high school to high school.

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 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.

[16]Goldin and Katz (2008), Chaps. 4 and 5

[17]This result is given in Goldin and Katz (2008), Table 2.5.

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

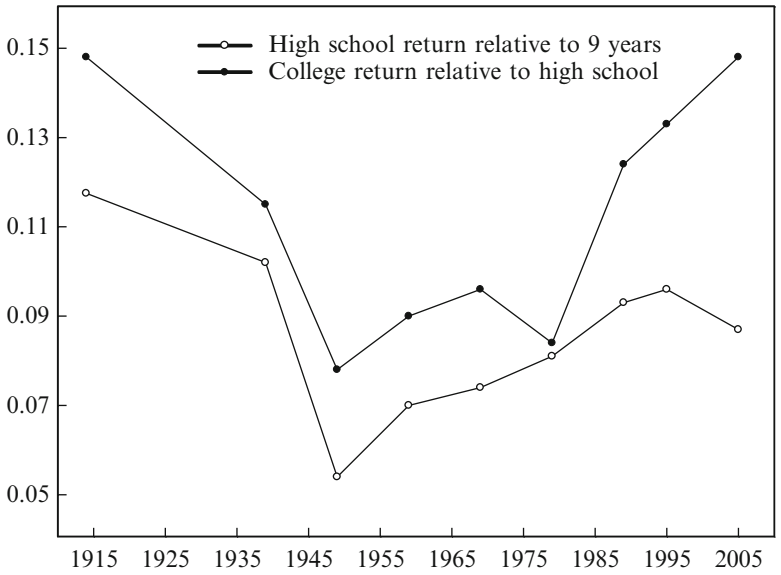

Fig. 8 Returns to a year of school for young men: 1914–2005 (Source: Goldin and Katz (2008), Fig. 2.9 and Table 2.7 for notes)

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]

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.

[18]For a discussion of higher education in the United States, see Goldin and Katz (1999, 2008).

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 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]

[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.

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.

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

[20]See Weil (2007) for a clever way to separate the effects of health on income from the reverse causality.

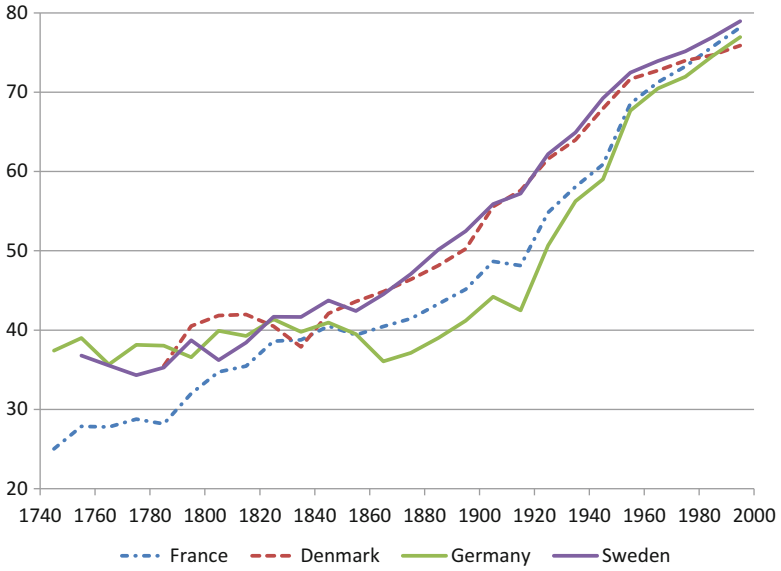

Fig. 9 Expectation of life at birth (period rates) for four European nations, 1745–1995 (Source: Floud et al. (2011), Table 5.1)

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 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.

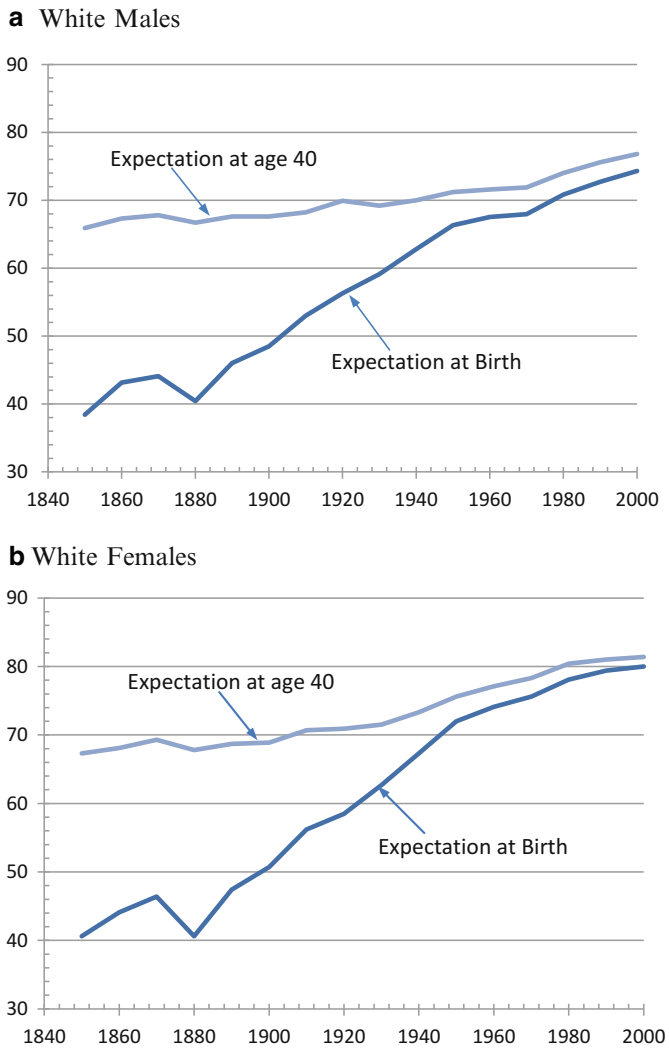

Fig. 10 Expectation of life at birth and at age 40 (period rates) for US whites by sex: 1850–2000 (Source: Carter et al. (2006), Tables Ab644-655)

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 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.

[21]The division among the three phases is the author's, not necessarily that of the various contributors to the literature.

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]

[22]On the changing relationship between health and economic development, see Preston (1975).

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 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).

$$Q = f(A, [E \cdot P\lambda], K, X) \tag{2}$$

[23]See Cutler et al. (2006), Fig. 3 for US data and Floud et al. (2011), Fig. 4.5 for England and Wales.

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*,
DOI 10.1007/978-3-642-40406-1_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:

$$Y = F(L, K, T) \tag{1}$$

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 occupation" to the census, the individual was part of the labor force. Examples of gainful occupations include farmer, carpenter, domestic servant, and clerk.[3]

[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.

[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).

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 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]

[4]It can also occur if information flows between A and B improve so that workers have more accurate knowledge of labor market conditions.

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 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).

[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.

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 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

[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).

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, 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

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))

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]

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

[7]For additional analyses of historical data on the length of the work data, see Whaples (1990), Costa (2000), and Vandenbroucke (2009).

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 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

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

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

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*,
DOI 10.1007/978-3-642-40406-1_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

[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

early-developing country, real GDP per capita expanded at 1.3 % annually from 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 microlevel 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

[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.

associated with medical care, a point conceded, if grudgingly, by Deaton (2006, p. 110).[6] 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

[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).

[7]Which involves allocating nutrition across various physical activities.

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 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

[8]Floud et al. 2011, p. 117

[9]Floud et al. 2011, pp. 116–118

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 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.

[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).

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 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:

$$\text{Height}_j = g\left(\text{Income}_j, Y_j\right) + \mu_j \quad (2)$$

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 production and the biological standard of living. Technological change in

[16]United Nations 1999, Table 1

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

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*,
DOI 10.1007/978-3-642-40406-1_24

Abstract

In this article, 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 sect. "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

[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.

numeracy rates is very strong on a regional or countrywide basis. Clearly, we are only able to 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]

[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).

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 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

[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).

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 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

[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.

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).

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

[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.

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 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

[15]De Moor and Van Zanden (2010) use the Whipple index for their calculations. We translated the numbers into ABCC values for convenience.

(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 (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

[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).

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, 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

[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.

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 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).

[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.

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 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

[22]East Asia is dominated by Chinese data, since Japan is considered part of the industrialized countries.

[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.

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 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.

[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).

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 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.

[26]The data are arranged in birth decades.

[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).

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 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,

[28]Included countries are Afghanistan, Bangladesh, India, Iran, Sri Lanka, Nepal, Pakistan, Hong Kong, Indonesia, Cambodia, Federation of Malaya, Sarawak, the Philippines, and Thailand.

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).

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 article 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. doi: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

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*,
DOI 10.1007/978-3-642-40406-1_1

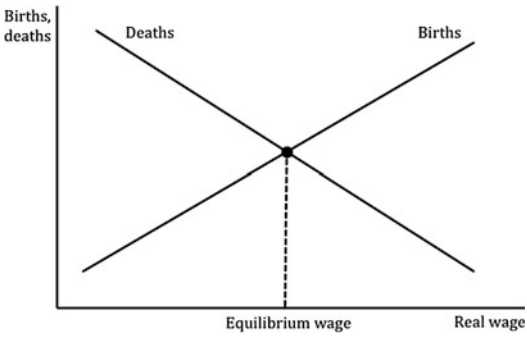

Fig. 1 The Malthusian model

on people's earning possibilities, and a death or a miscarriage often followed from 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

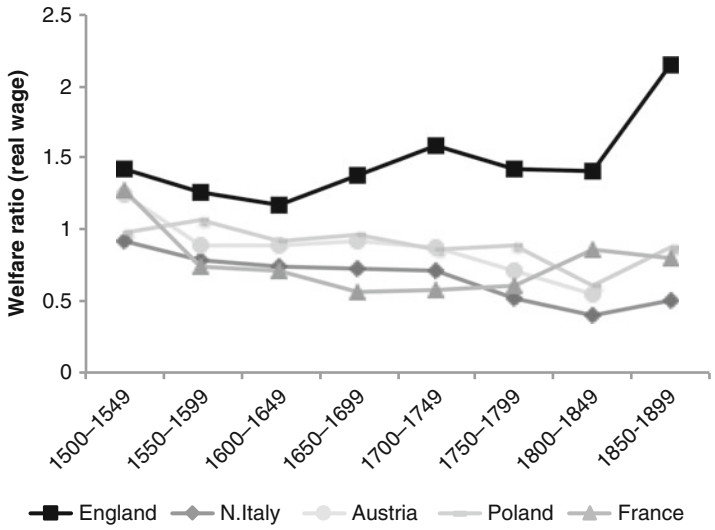

Fig. 2 The Great Divergence within Europe (Source: Allen (2001))

permanent shifts in the positions of the birth and death schedules? The answers 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 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 socioeconomic 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).

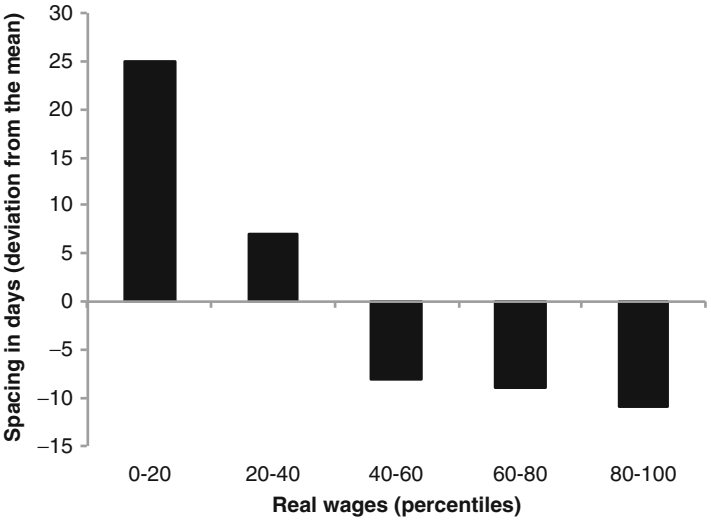

Fig. 3 Birth-spacing intervals by real wage percentiles (Source: Cinnirella et al. (2013))

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 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 CAMPOP statistics, Boberg-Fazlic et al. (2011) divided the

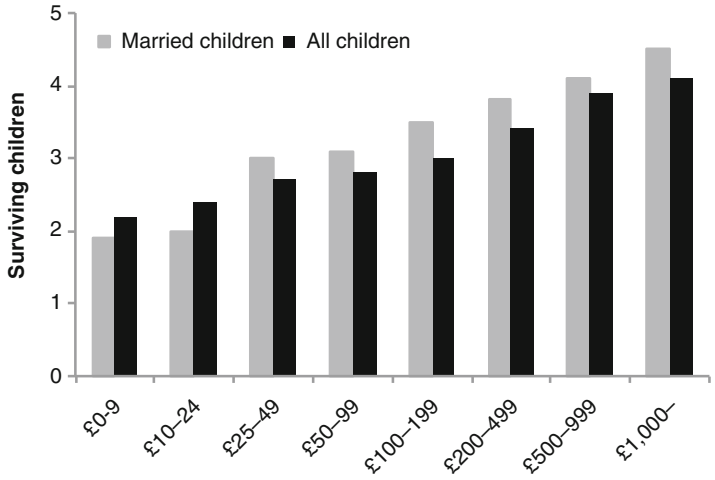

Fig. 4 Reproductive success by wealth (Source: Clark and Hamilton (2006))

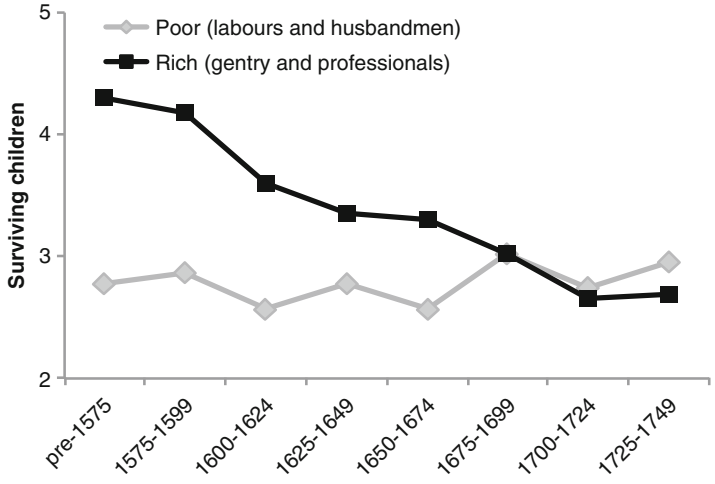

Fig. 5 Reproductive success of rich and poor (Source: Boberg-Fazlic et al. (2011))

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.

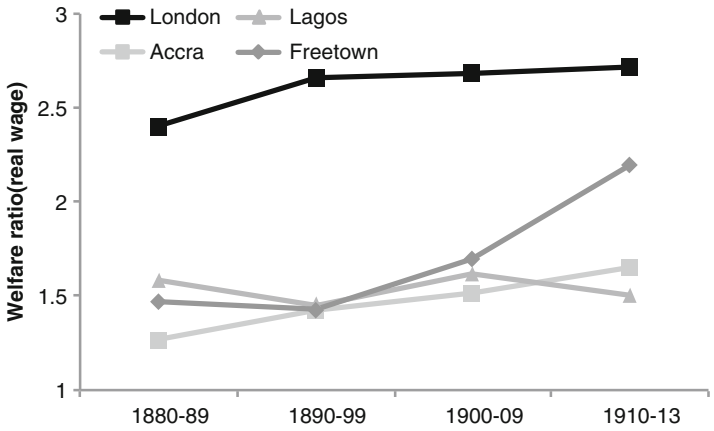

Fig. 6 The Great Divergence between Europe and Africa (Source: Allen (2001), Frankema and van Waijenburg (2012))

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

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

Fig. 7 The shares of Kampala men and women in salaried work (Source: Meier zu Selhausen and Weisdorf (2014))

to 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

Growth Theories

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*,
DOI 10.1007/978-3-642-40406-1_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.

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.

[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.

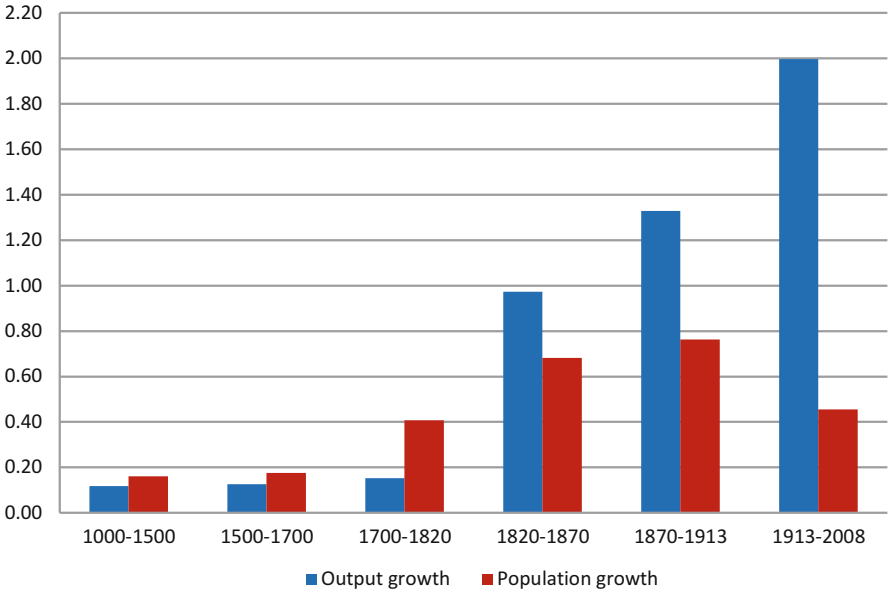

Fig. 1 GDP per capita and population growth rates in Western Europe (30 countries) (Source: Data from Maddison (2008))

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 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 % 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

[2]The demand for education increased from the end of the period.

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.

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?

[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.

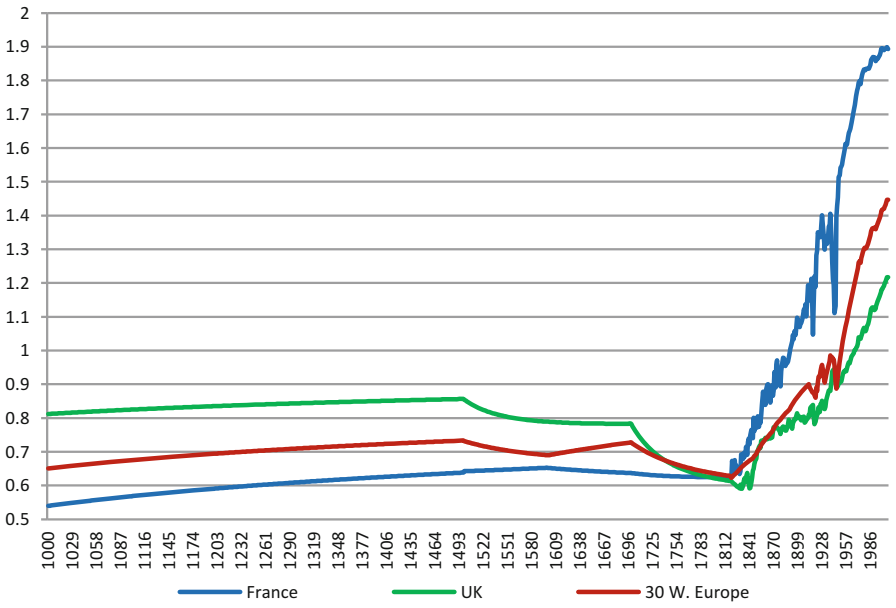

Fig. 2 Ratio of output-to-population growth rates in France, UK and Western Europe, 1000–2008 (Source: Using data from Maddison (2008))

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 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

[5]The term was coined first by Galor (2005).

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 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.

[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.

[7]Within the framework of the neoclassical growth model with endogenous fertility, the authors attempt to determine the optimal population growth rate.

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 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 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

[8]Tamura (1994) finds the same result.

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.

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.

[9] In this paper, 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 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 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).

[10]Notably the pioneering work of Jacob Mincer (1962) that contributed to the development of economic analysis of the household.

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

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 paper 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 paper 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*,
DOI 10.1007/978-3-642-40406-1_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

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

[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).

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.

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.

[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.

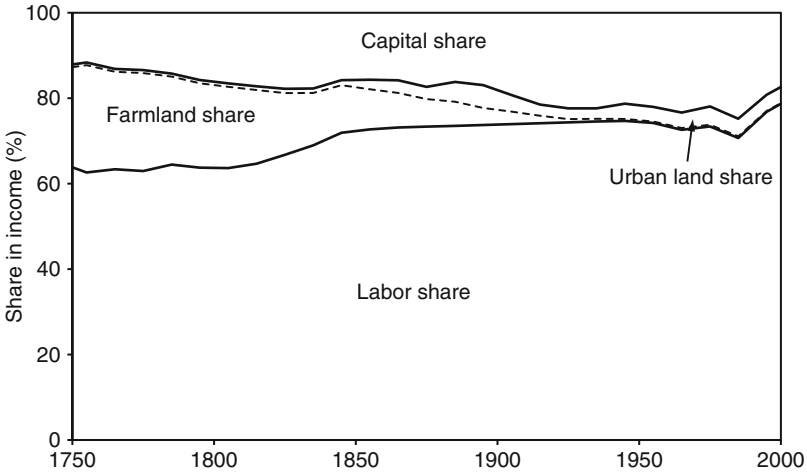

Fig. 1 Factor shares 1750–2000, England (Source: Clark 2007a, Fig. 14.4)

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

Some economists, most notably Paul Romer, have theorized that this correlation derives from physical capital accumulation creating substantial external benefits not 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

$$A = \frac{r^a p_k^a w^b s^c}{p} \tag{3}$$

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.

[5]For a more detailed explanation, see Clark (2007a, pp. 379–382).

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)

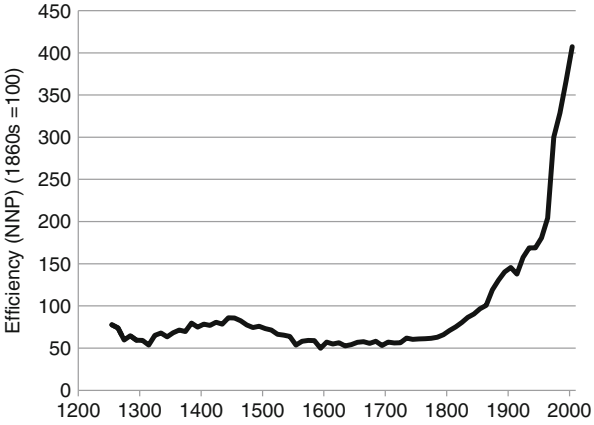

Fig. 2 Estimated efficiency of the english economy, 1250–2000 (Source: Clark 2010)

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

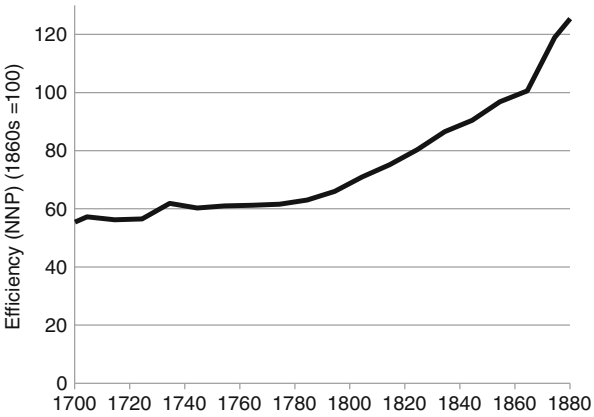

Fig. 3 Efficiency levels, England, 1700–1880 (Source: Clark 2010)

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 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

[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)

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.

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

[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)

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, 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,

[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).

had accessible deposits of coal near population centers. That, rather than differences in innovative potential, explains British success and Chinese failure. While the absence 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

[12]For more details, see Clark and Jacks (2007).

[13]Wicker (1957).

[14]Overton (1985, Table 1).

[15]Overton (1991, pp. 309–310).

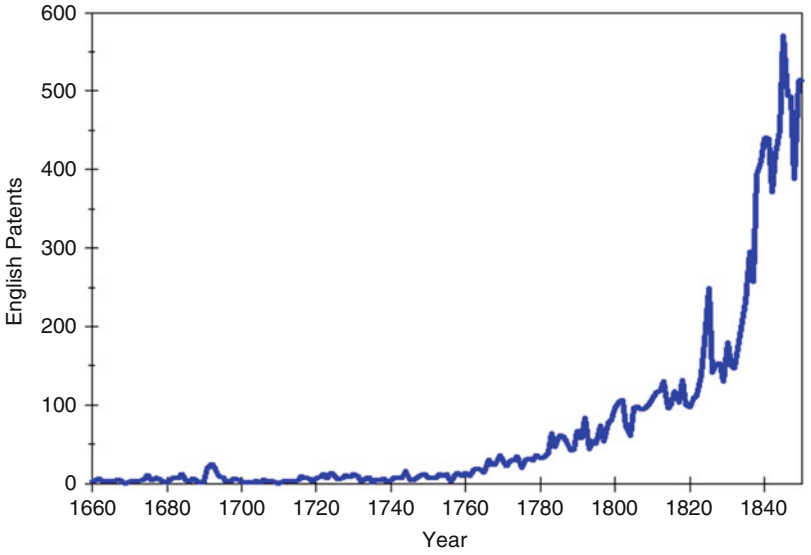

Fig. 4 Patents per year, England, 1660–1851 (Source: Mitchell 1988)

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 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

[16]Clark (1998), Clark and Jamelske (2005).

[17]This is also the analysis of Mokyr (2003, 2012).

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 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.

[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).

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)

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. 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

[21]By 1665 the State of Holland was able to reduce rates on its long term debt to 4 %.

[22]De Vries (1978).

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))

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 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 Shleifer 1993; Greif 2006; North 1981, 1994; North and Thomas 1973; North and Weingast 1989; North et al. 2012; Rosenthal 1992).[25]

[23] See Eq. 3 above.

[24] Allen (2009).

[25] Clark (1996, 2007a, b) criticizes this approach.

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.

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

[26]The recent rise of China is, however, an exception to the general association of growth and democracy.

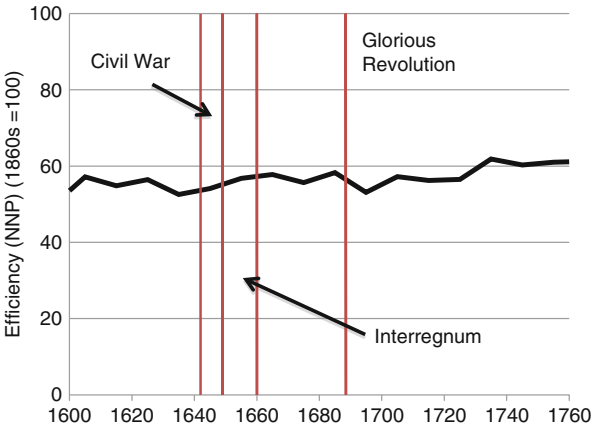

Fig. 6 Economic efficiency and political changes, England, 1600–1770 (Source: Clark 2010)

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.

The Netherlands of the seventeenth century was just one of many earlier European societies that had property holder franchises. From 1223 to 1797, Venice

[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).

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 that the concept of intellectual property was alien to cultural norms in imperial

[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).

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

[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))

to pass 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.

[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)

The reason for this is that Parliament could, on grounds of the public good, extend patents 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.

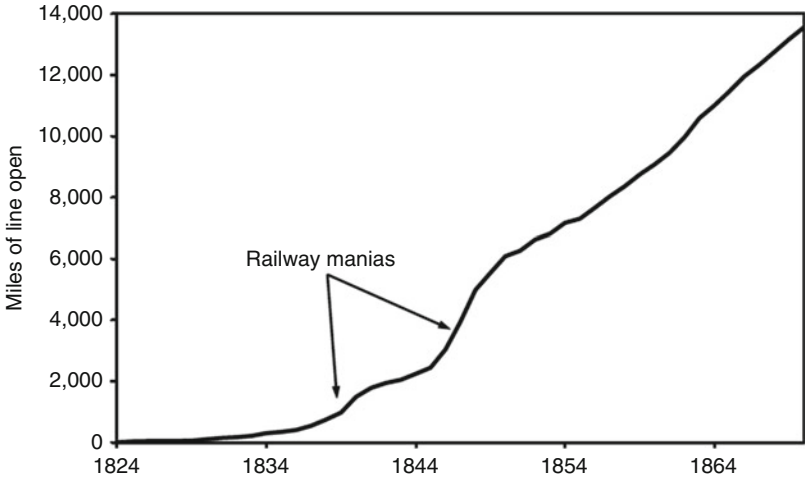

Fig. 8 English railroad construction, 1825–1869 (Source: Mitchell and Deane 1971, p. 225)

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

[35]The industrialization of the United States created much greater private and family fortunes.

this era) in a trial in 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.

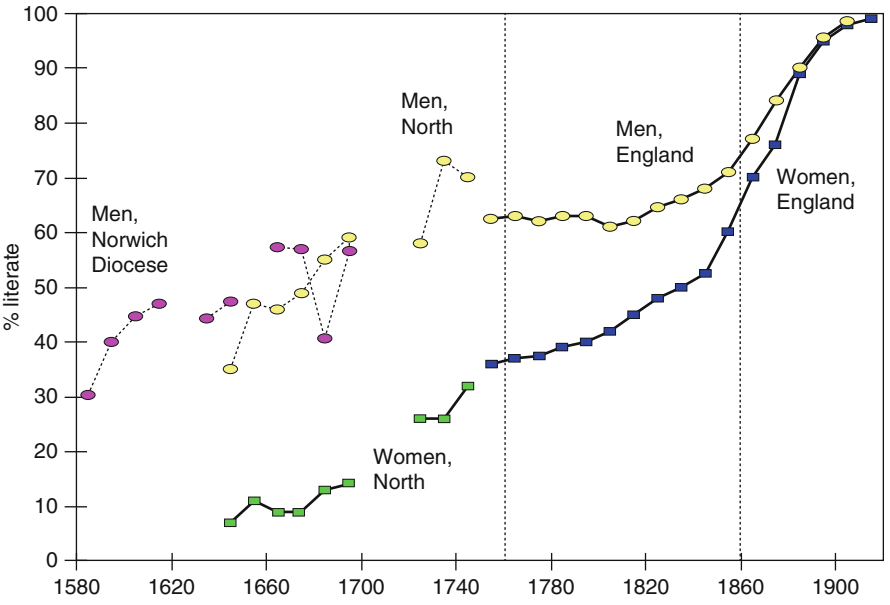

Fig. 9 Literacy in England, 1580–1920 (Source: Clark 2007a, Fig. 9.3, p. 179)

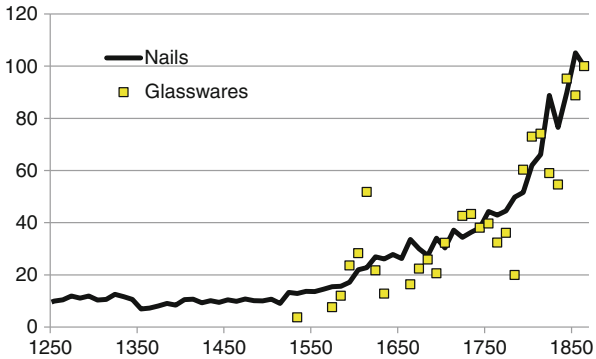

Fig. 10 Efficiency of production of nails and glassware, by decade, 1250–1869 (Source: Clark 2010)

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.

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 preindustrial 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

[39]Shineberg (1971, p. 65).
[40]Townsend (1983, p. 6).
[41]Wrigley et al. (1997, p. 313).

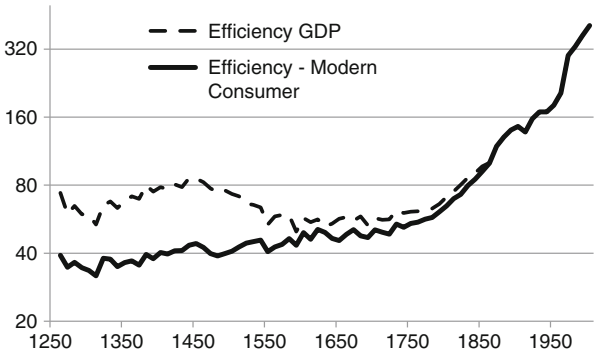

Fig. 11 Economic efficiency from the perspective of a modern consumer, England, 1250–2009. *Notes*: The weights in consumption for the modern consumer are assumed to be half from the consumption basket of the pre-industrial worker. But the other half is composed of books (.1), manufactured goods (.1), clothing (.1), sugar (.03), spices (.03), drink (.05), light (.05), soap (.02), and paper (.02) (Source: Clark 2010)

vertical axis so that the slope of the line measures the rate of efficiency growth. Thus, 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

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)

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.

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

[42]Hopkins (1966, p. 249).

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, 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

[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.

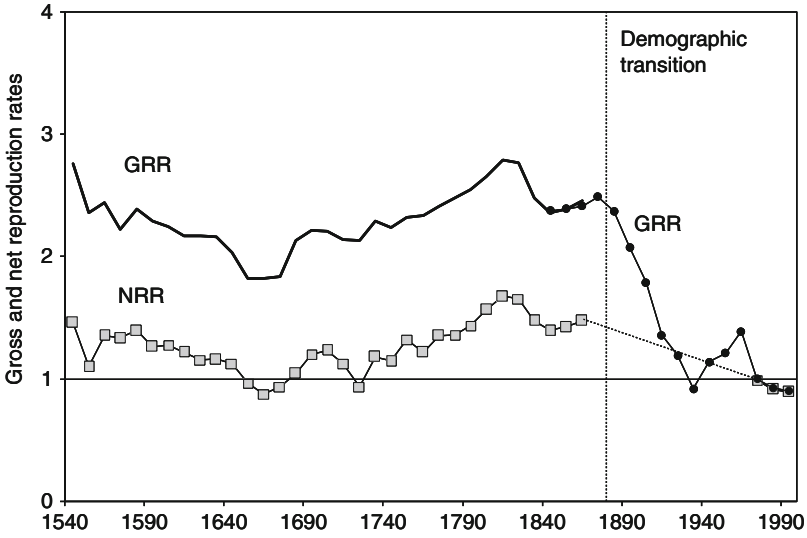

Fig. 12 The fertility history of England, 1540–2000 (Source: Clark 2007a, Fig. 14.6, p. 290)

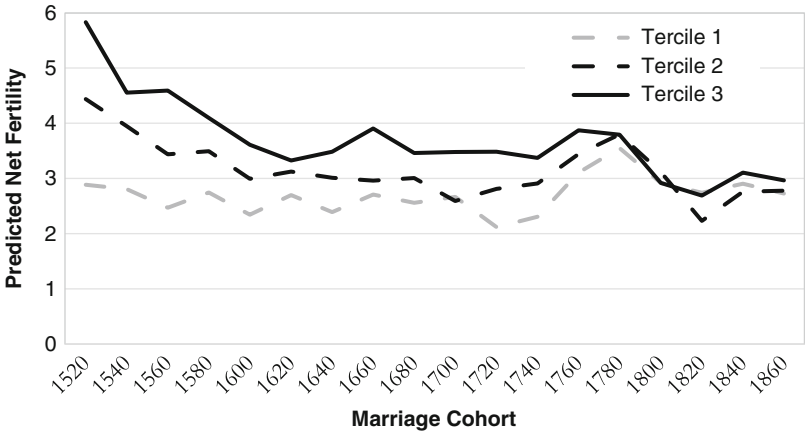

Fig. 13 Net fertility by wealth terciles, marriage cohorts, 1520–1879 (Source: Clark and Cummins 2015)

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 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).

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)

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 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

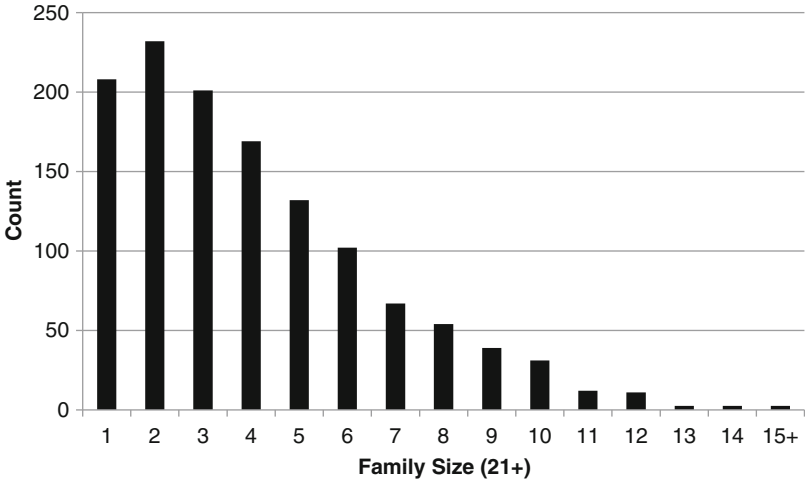

Fig. 15 The distribution of family size in english upper classes, marriages 1770–1879 (Source: Clark and Cummins 2014)

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

$$\mathrm{DOXB}_{t+1} = a + b_1\mathrm{DOXB}_t + b_2\ln(\mathrm{Wealth})_t + b_3N + e_t$$

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

$$\ln W_{t+1} = b_0 + b_1 \ln W_t + b_2 \ln N + b_3 \text{DFALIVE} + e_t$$

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 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

Economic-Demographic Interactions in Long-Run Growth

James Foreman-Peck

Contents

Abstract

Cliometrics confirms that Malthus's model of the preindustrial economy, in which increases in productivity raise population but higher population drives down wages, is a good description for much of demographic-economic history. 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

J. Foreman-Peck (✉)
Cardiff University, Cardiff, UK
e-mail: foreman-peckj@cardiff.ac.uk

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
DOI 10.1007/978-3-642-40406-1_17

until the end of the nineteenth century. Interactions between the economy and migration 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

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, commonly due to human capital accumulation.

A country's population is potentially influenced not merely by natural increase – the excess of fertility over mortality – but also by migration. For much of history, simply establishing the size of population 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 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 particular times have considered themselves overpopulated and encouraged emigration. Others have believed they were underpopulated and promoted immigration or family-friendly policies. Most populations change primarily, not from migration but from "natural increase," the excess of births over deaths. Natural increase of different 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.

The 1601 Poor Law in England was unusual in the early acknowledgement of some state responsibility for sections of the population in difficulty. Elsewhere their relief was 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 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 identification; and section "Time Series Analyses" outlines the distinctive approaches of some recent studies that use time series methods.

Data

How far cliometrics can proceed plausibly 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 2007). The culmination of work by Thorold Rogers (1866), Beveridge (1939), and Phelps-Brown and Hopkins (1955, 1956, 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 (2014) 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 arrived, 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 intriguing. 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 2014). Most 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 early European census of population and assets is that of 1086 in England, in the Doomsday Book. By the end of the eighteenth century, governments were beginning to undertake censuses regularly. The first official US population census was in 1790, and the first British official census was 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 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 choose not to register, perhaps for religious reasons. Migration

[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.

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, it is clear that 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 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).

[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, 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 to the civilian population as the Black Death, because of disease spread and 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) find that the disappearance of the positive check coincided with the introduction of systematic poor relief. They are able to 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 cultivated 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)

We can simulate these oscillations, in response to a mortality shock, for instance, to show a key feature of the Malthusian model. This exercise in "deterministic calibration" illustrates 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,

$$w_t = a - b.P_t + u_t \qquad a, b > 0 \tag{1}$$

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.[6]) 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 to 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 P_t increases with the previous period wage because higher wages encourage earlier marriage and because children are "normal goods" (the diminution of the preventive check): as household income rises, more children become possible.[7] As well as the effect on births, at low living standards, higher wages reduce positive checks to premature death.

Population reflects the balance between births and deaths, along with migration. Ignoring migration, that is, focusing on "natural increase," population measured at

[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. (2011) 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]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.

[7]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.

the end of period t is equal to population at the end of period $t - 1$, plus the excess of birth (B) over deaths (D) over the period t:

$$P_t = P_{t-1} - D_t + B_t$$

For the moment, we will not distinguish theoretically between positive and preventive checks, simply recognizing that both, and their net effect, may depend upon the level of wages.[8] Where v_t is a random disturbance term with mean of zero and c and d are parameters,

$$P_t = c + d.w_{t-1} + v_t \qquad c, d > 0 \tag{2}$$

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.[9]

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 preventative 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]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.

[9]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 (Voth and Voigtlander 2012), 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 and Voigtlander 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 $a = 1$, $b = 0.8$, $c = 1$, 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

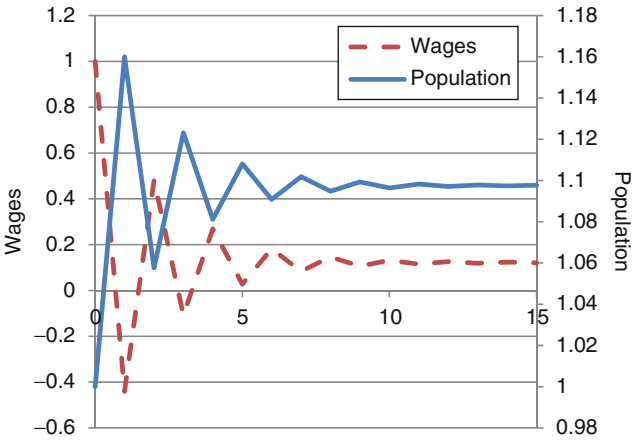

Fig. 1 Simulated Malthusian cycles

of the Malthusian process can be shown with an Excel spreadsheet[10] and by rearranging the system as follows:

$$P_t + 0.64.P_{t-1} = 1.8$$

$$w_t + 0.64\, w_{t-1} = 0.2$$

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

[10]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 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 "= − $A $1*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

A breakout from the above Malthusian equilibrium can be achieved in two general ways. The first is by incomes becoming sufficiently high that the time opportunity cost of children ensures that the negative substitution effect dominates the positive income effect. Then the *d* wage coefficient falls in the population Eq. 2, perhaps even becoming negative. This could be one interpretation of the high-mortality fourteenth and fifteenth centuries in Western Europe that pushed wages to unprecedented heights. With high wages, there is a high value of time and so a desire to limit numbers of children – through the Western European marriage pattern, which appears to have arisen around this time. A variant explanation for fertility change is that the mortality shock and higher male wages 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 (Voigtlander and Voth 2013). However, women's wages in England do not seem to have conformed to this pattern (Humphries and Weisdorf 2014).

Phase two of the demographic transition in Europe as conventionally understood refers not to this earlier period but to the nineteenth century, when population growth generally increased strongly and real wages did not fall any further. This last suggests the second breakout possibility at work, 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 faster technical progress would be absorbed by greater populations, as birth rates rose. But real wages do not fall in this scenario. By contrast, a permanent mortality shock, reducing death rates, would instead lower wages as population increased, until positive or preventive checks intervened. The next phase of the demographic transition involves such checks, especially a fall in the *d* coefficient.

The early French fertility transition described by Weir (1984) after 1830 is consistent with rising living standards reducing the demand for children; a slightly rising age at marriage was accompanied by falling marital fertility. Unlike France, only after 1900 did sustained fertility decline began in much of Bavaria, when the signs of rising opportunity costs of children become apparent. By 1910, urbanization strongly reduced rural marital fertility (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. A time series analysis of the British fertility decline found that illegitimacy lagged accounted for almost one-third of the decline, where illegitimacy is assumed to be an indirect measure of contraceptive costs (Crafts 1984). Crafts (1984) maintained that the cheaper or more widely understood was contraception, the lower would be both illegitimacy and the legitimate fertility.

The second breakout possibility (changing the sign of b, the population coefficient in the wage equation) is in effect postulated to have occurred by population size boosting technological advance in unified growth theory (Galor and Weil 2000). The interpretation could be extended to include market widening, such as the European discovery of the Americas – Smithian growth (cf Acemoglu et al. 2005).

Alternatively, if a, the intercept in the wage equation, is increasing due to technological progress – perhaps from human capital accumulation – the negative effects of population on wages operating through b might be offset. An early discussion of the difference between US and British technology stemming from land-labor ratios might be invoked to explain different rates of technical progress after a demographic shock. Labor scarcity in the USA – and/or perhaps land abundance – supposedly encouraged the adoption of more capital-intensive techniques and stronger technical progress ((Habakkuk 1967; David 1975, Chap. 1). These ideas might be transposed to other economies experiencing shocks that radically altered factor ratios. Then greater technical progress might be triggered by shocks increasing labor scarcity (such as the Thirty Years War in Germany or the high-mortality fourteenth century in Western Europe).

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). This balance was likely to change, and at different times in different countries. French exceptionalism appears with widespread evidence of marital fertility control in France in the 1790s (reducing c), nearly 100 years before comparable evidence in England (Weir 1994). 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.

Natural selection in favor of higher child quality (Galor and Moav 2002) cumulatively has this consequence of raising technical progress, or increasing a, as may falling mortality. Higher life expectancy can raise the returns to investment in human capital because there is a longer 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 preferences for child quality

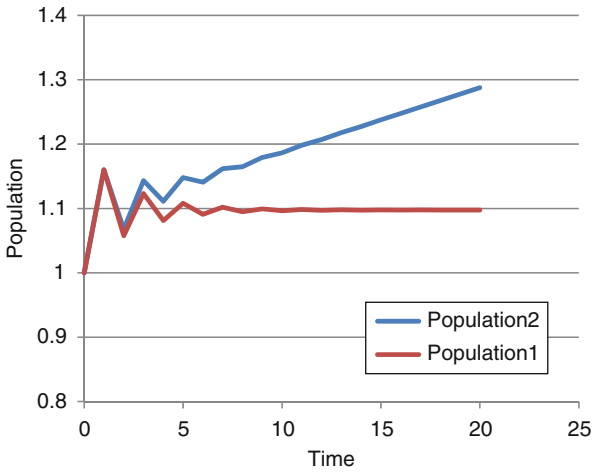

Fig. 2 Population in the break out from the Malthusian cycle

(Foreman-Peck 2011). In this last case, higher female marriage age is hypothesized to increase the rate of human capital accumulation through "child quality" family choice as well. Assuming literacy is a 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 and literacy appears to contribute to output in a cross-Europe production function for the years 1870 to1910.

With less emphasis on family consensus, Diebolt and Perrin (2013a, b) identify gender empowerment increase as the key for the shift towards child quality. Empowerment increased the amount of time invested by women in their education, and fertility therefore declined. These choices shifted outcomes from quantity of children towards quality of children. Female literacy was associated with falling fertility in France in the period 1881–1911, which Diebolt and Perrin describe as evidence for a child quality-quantity trade-off.

Inspection of Eq. 3 suggests that if a increases continuously, then population will eventually "take off." The same applies to the wage difference equation. Figure 2 shows the effect of 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. But the peak in period three is a little higher than in the model with the static equilibrium. After period six, population is growing continuously, though obviously the strength of the time trend determines when the breakout occurs.

The end of this phase of the demographic transition comes as incomes rise, and preferences change towards less time-intensive activities than child rearing (that also become more abundant), bringing down birth rates. Technological change favoring brain over brawn altered the gender division of labor so as to favor fewer

children as well.[11] In terms of the simple model, technological progress, reflected in continuous increases in a, must raise real wages continuously. Since real wages are growing, the opportunity cost of children will be rising. This implies that d, the response of population to wages, may be falling. 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 two changes: the rise in a and the fall in d.

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; O'Rourke and Williamson 1999; Hatton and Williamson 2006; 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 means (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 exactly 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

[11] Alesina et al. 2011 maintain that plough-based agriculture led to lower female participation in wider society and the economy, and conversely. Presumably, plowing required greater physical strength than shepherding.

"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 particular 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

[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).

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 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 populations, 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 we know or can estimate the values of the parameters of the favored model. If we find an association in time series data between wages and population, in principle, it could reflect relations generated by one or both of the equations in the model we have been discussing. We cannot infer any of the parameters a, b, c, and d from the estimated relation without further information; the original equations are not identified.

If we can distinguish shocks to, or variables affecting, one equation that do not affect the other, then we have 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 6 million around 1300 on the basis of 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.

[13]Williamson (2005) also discusses the corollary that the position of blacks deteriorated after 1970 because of competition from immigrants.

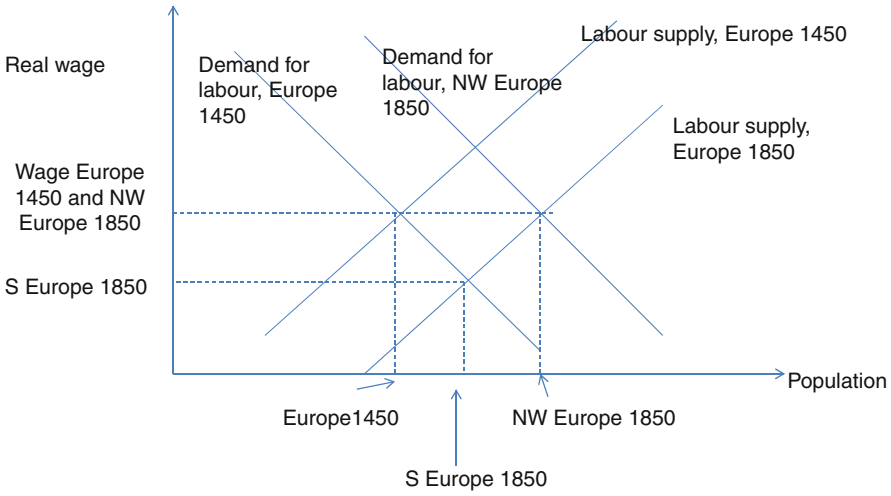

Fig. 3 Allen's (2001) interpretation of European divergence 1450–1850

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 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 in particular. 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 labour (eq. 1), and also utilise the many studies that provide parameter estimates for calibration. They assume that migration was exogenous to national

populations or labour forces; it shifted equation 2, thereby identifying equation 1. Wage changes are then found from the change in population or labour force due to migration over the period.

Time Series Analyses

We have discussed calibration of structural equation models. The econometric alternative for estimating identified economic-demographic models is generally concerned with time series analysis. The study of preventive and positive shocks in particular 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 a number of 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 of 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 find 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 preventative checks working through marriages and agree with Nicolini that positive checks were insignificant. They suggest that growing population actually 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) are able to 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 (Bouccine 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 p. 7). Foreman-Peck and Zhou (2014) 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, whereas the demographic series begin only in 1541.

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 there had been mortality shocks. 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.

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

[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.

east, 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).

Human capital accumulation is frequently emphasized as the source of the breakout to modern economic growth, often helped by lower mortality that raised the rate of return or simply encouraged household choice for child quality. But the inability to measure human capital, other than by weak proxies such as literacy, has been a handicap. Nonetheless across Europe and across English counties, literacy tended to be higher in countries with later female age at marriage, and there is some evidence that national outputs responded to greater literacy.

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.

In the Malthusian economy, there is strong evidence that population, or natural increase, did respond to higher wages. This encourages the expectation that subsidy policies like those attributed to the English 1601 Poor Law would encourage fertility of those subsidized, as Malthus feared. There is evidence that this happened and that the "social safety net" reduced premature deaths by breaking the link between harvest failure and mortality.

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

Alesina A, Giuliano P, Nunn N (2011) On the origins of gender roles: women and the plough. National Bureau of Economic Research, Cambridge, MA

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 (2007) How prosperous were the Romans? Evidence from Diocletian's price edict of 301, Economics discussion paper no 363. University of Oxford, Oxford
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
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, van Leuween B (2011) Arable acreage in England, 1270–1871. LSE unpublished
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
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
David PA (1975) Technical choice innovation and economic growth. Essays on American and British experience in the nineteenth century. Cambridge University Press, London
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
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
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 (2014) The rise of the English economy 1300–1900: a lasting response to shocks. Cardiff economics working paper E2014/3
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
Habbakuk HJ (1967) American and British technology in the nineteenth century. The search for labour saving inventions. Cambridge University Press, Cambridge
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 (2014) The wages of women in England 1260–1850. University of Oxford, unpublished
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
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 (2014) Human capital and industrialization: theory and evidence from the enlightenment. Strasbourg Beta workshop paper
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
Voigtländer N, Voth H-J (2013) How the West 'Invented' fertility restriction. Am Econ Rev 103(2013):2227–2264
Voth H-J, Voigtlander N (2012) The three horsemen of riches: plague, war, and urbanization in early modern Europe California Center for Population Research, University of California Los Angeles, PWP-CCPR-2010-032
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
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

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 the estimates are

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 (✉)
Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Barcelona, Spain
e-mail: emanuele.felice@gmail.com

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
DOI 10.1007/978-3-642-40406-1_5

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 nonquantitative 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 an 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 published in 1934, with reference to the US economy from 1929

[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.

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. 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, GDP was

[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.

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 nonmonetary dimensions. Even the Human Development Index, which is gaining consensus among economic historians (Crafts 1997, 2002; Prados de la Escosura 2013a, b), 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

[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).

production function: $Y = f(L, K)$. A widely accepted specification of this function is the Cobb–Douglas form

$$Y = A \times L^{\alpha} \times K^{\beta} \quad (1)$$

and in particular the one with $\alpha + \beta = 1$ (i.e., with constant returns to scale)

$$Y = A \times L^{\alpha} \times K^{1-\alpha}. \quad (2)$$

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 were 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, there 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 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),

[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).

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. 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 2,000 years for most countries, also accomplishes this rule: although some of his assumptions for the nineteenth century are questionable – or may simply look too crude[10] – an

[9]For modern times, outstanding examples are Feinstein (1972) for the United Kingdom and Prados de la Escosura (2003) for Spain.

[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).

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 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*

[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.

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

$$GDP^{ti}/Q^{ti} = GDP^{(t+1)i}/Q^{(t+1)i}, \tag{3}$$

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

$$GDP^{ti} = \left(Q^{ti}/Q^{yi}\right) \times GDP^{yi}, \tag{4}$$

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 spanning long periods of time is all but a chimera. Once this human (social) 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

[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).

(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.

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, 1910/11 (1861–1938/39); the United States, 1929 (1890–1929) and 1987 (1929–1950); the

[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).

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

$$y_{i_{in}\ i\min\ \text{prices}}^{\frac{i\max - i}{i\max - i\min}} \times y_{i_{in}\ i\max\ \text{prices}}^{\frac{i - i\min}{i\max - i\min}} \tag{6}$$

where i is the year of the series y, $i$min is the early benchmark, and $i$max 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

[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.

[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.

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] (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 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,

[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 eight 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).

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 a 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 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

[20]In that article, an excellent discussion of the literature about these issues and the different shortcut methods is also provided (pp. 2–8).

[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).

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 estimates by both Maddison (upper quadrant) and Prados de la Escosura (lower quadrant) are available for an unchanged minimum number of 20 countries.[23]

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

[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).

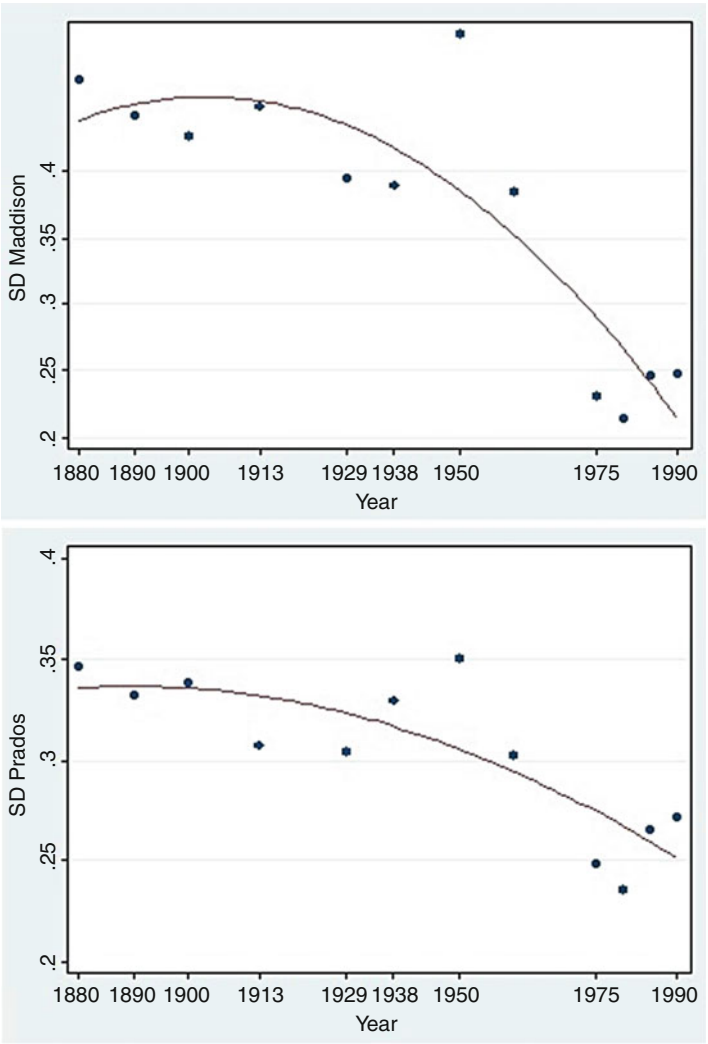

Fig. 1 σ-Convergence in GDP per capita from 1880 to 1990, according to different GDP estimates (Sources and notes: see text)

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.

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

$$\gamma_k = s * Af(k)/k - (\delta + n) \tag{7}$$

[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.

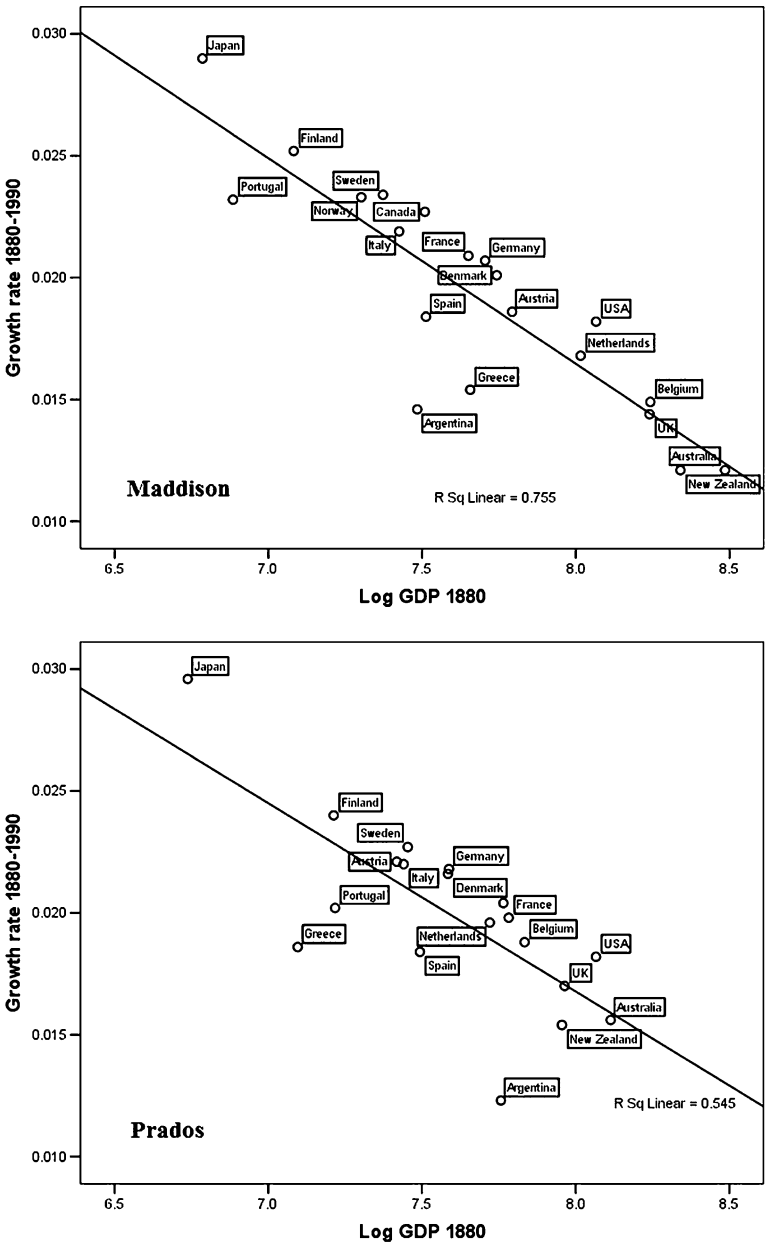

Fig. 2 β-Convergence in GDP per capita from 1880 to 1990, according to different GDP estimates (Sources and notes: see text)

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 remains 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 one 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).

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]

[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.

[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.

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, 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

[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).

(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 paper 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. doi: 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

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). Rivista di Politica Economica 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. Rivista di Storia Economica 19(3):359–381
Feinstein CH (1972) National income, expenditure and output for the United Kingdom, 1855–1965. Cambridge University Press, Cambridge
Feinstein CH, Thomas M (2002) Making history count. A primer in quantitative methods for historians. Cambridge University Press, Cambridge
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, 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. Annali di statistica 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 del Lavoro Q Rev 44(177):225–241
Maddison A (1994) Confessions of a chiffrephile. Banca Nazionale del Lavoro 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, vol I. Development Centre Studies, 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. Rivista di Storia Economica 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
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 (2013a) Human development in Africa: a long-run perspective. Explor Econ Hist 50(2):179–204
Prados de la Escosura L (2013b) World human development, 1870–2007. Working papers in economic history 13-01, Universidad Carlos III
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 Change 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 point 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 (✉)
Universidad Carlos III Madrid, Madrid, Spain
e-mail: markus.lampe@uc3m.es

P. Sharp
University of Southern Denmark, Odense M, Denmark
e-mail: pauls@sam.sdu.dk

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
DOI 10.1007/978-3-642-40406-1_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. 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 autarchy 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 to early 2014, 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."

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, Johnson, and Robinson (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 towards 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, Johnson, and Robinson (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, O'Rourke, and Taylor (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 payments 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 travellers 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, Kauppila (2009) at Finland during the Great Depression, and Taylor, Basu, and McLean (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 (2013) 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.

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 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 and 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.

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). Recent research has moved towards 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).

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 towards the new customs area. Kauppila (2008) has studied the impact of tariffs on industrial activity and prices in interwar Finland. Tirado, Pons, Paluzie, and Martínez-Galarraga (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, Rosés, and Sanz-Villaroya (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) 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. Finally, Henriksen, Lampe, and Sharp (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.[1] 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.

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[2], 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.[3]

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

[1]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.

[2]See the useful discussion on this in Persson (2004).

[3]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.

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 towards 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, 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 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) and Federico and Tena-Junguito (2013).

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 (2013) 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 travelled 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-Marín (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.[4] 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.

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

[4]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).

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. 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[5]: 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 is risky, except with the addition of other information, particularly on the existence of trade. This is because

[5]Following Cournot (1838)

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, 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.[6]

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.[7] 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,[8] 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 upwards 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.

[6]See, for example, Obstfeld and Taylor (1997), Jacks (2005, 2006a, b).

[7]See, for example, Chartres (1995), Ljungberg (1996), Peña and Sánchez-Albornoz (1984), and Bessler (1990).

[8]See, for example, Shiue and Keller (2007), Persson (1999), and Bateman (2011).

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 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 towards 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.[9]

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.[10] Similarly, but looking more generally at Europe and the Americas, Dobado-González et al. (2012), using a new methodology[11] 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), as well as those within Europe. They would not, of course, classify this as evidence of globalization.

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

[9]See the recent paper by Hynes et al (2012).

[10]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).

[11]Their methodology makes use of the residual dispersion of univariate models of relative prices between markets.

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.[12]

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 use 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 or panel designs. The gravity model departs from a simple but theoretically microfounded 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.

[12]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.

However, theoretical motivations and econometric applications have shown that the simple, "naïve" gravity equation, following Head and Mayer (2013, Eq. 4, p. 12)

$$X_{ni} = GY_i^a Y_n^b \phi_{ni} \tag{1}$$

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. 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 (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, 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.

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, p. 124–125), Flandreau and Maurel (2005, p. 139), and Estevadeordal et al (2003, p. 374) find indications of a significantly negative relationship before World War I. For the same period, Jacks (2006b, p. 220) 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 López-Córdova and Meissner (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, Jacks' (2014) study of the effects of the imperial preference system resulting from the 1932 Ottawa Agreements on Canadian trade patterns at the commodity level merits attention. He 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." In contrast, Gowa and Hicks (2013) find 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 towards the preference group.

Another potentially transaction cost-reducing military–politico–economic institution, somewhat related to the interwar trade blocs 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, Mayer, and Ries (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, and 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. 2014).

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. 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. 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.

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, Nicita, and Olarreaga (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 (DWL) 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, 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, pp. 887–888) 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 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 Volckert (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 towards 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, p. 181). 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 his 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. 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
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
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, vol VIII, The Cambridge economic history of Europe. 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. Göteborg 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, Murphy TE, Studer R, Coşkun Tunçer A (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. Scandinavian Economic History Review 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 Vries J (2010) The limits of globalization in the early modern world. Econ Hist Rev 63(3):710–733
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) Autarquía y atraso tecnológico en la industria textil española, 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
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 (2013) World trade 1800–1938. Paper presented at the international conference on Trade policies in Europe in the long nineteenth century, University of Bordeaux, March 2013
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
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, Rubio M d M (2012) On the accuracy of Latin American trade statistics: a non-parametric test for 1925. In: Yañez 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
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, vol I, The Cambridge economic history of modern Britain. 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 (2013) Gravity equations: workhorse, toolkit, and cookbook. CEPII working paper 2013–2027
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
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 M (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. Explorations in Economic History 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 (2011) 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
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
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(2014):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 (2014) Same but different: dialects and trade. German Econ Rev doi: 10.1111/geer.12047
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 (2013) Market potential and the rise of US productivity leadership. NBER working papers 18819
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
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
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
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
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, 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
O'Rourke KH, Taylor AM, Williamson JG (1997) Factor price convergence in the late 19th century. Int Econ Rev 37(3):499–530
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.
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
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
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. 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 d P, 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
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, Tâmega 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
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
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 (S) (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

Part IV

Finance

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*,
DOI 10.1007/978-3-642-40406-1_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

[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.

British funds in 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

[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.

actual 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.

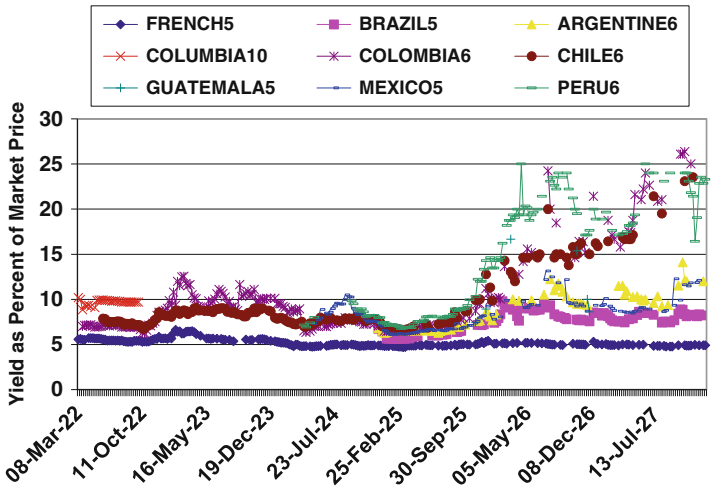

Fig. 1 French and Latin American Bond Yields in London, 1822–1827

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 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). 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

[4]See, for example, Davis and Gallman (2001) and Kindleberger and Aliber (2011).

to inject high-powered money into the economy during a crisis and the increased 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

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

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*,
DOI 10.1007/978-3-642-40406-1_13

buyers and 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

[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).

northern (from Flanders) and southern (from northern Italy) European merchants. Once a merchant's coin holdings had been presented, the value of which was 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

[2]A more comprehensive history of the bill of exchange than that which follows here may be found in Denzel (2010, pp. xxii–xlvi).

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 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

[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."

transfers of Bank of England accounts). Multilateral net clearing certainly 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.

[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).

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 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]

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

[5]By 2010 Federal Reserve banks processed slightly less than half the value of interbank paid checks.

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

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, NY, USA
e-mail: mjaremski@colgate.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
DOI 10.1007/978-3-642-40406-1_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, population, agriculture, and geographic location. As might be expected, acquisitions were more likely to take

[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.

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. 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.

[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.

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 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

[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.

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 $(N \times 1)$ vectors 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 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.

[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.

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.

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

[8]Given the varying frequencies, each observation is a bank-month.

[9]The approach was first developed in Kaplan and Meier (1958).

$$S(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

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 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

$$x_{1,t} = a_{1,0} + \sum_{i=1}^{k} a_{1,i}x_{1,t-i} + \sum_{i=1}^{k} b_{1,i}x_{2,t-i} + \sum_{i=1}^{k} c_{1,i}x_{3,t-i} + u_{1,t}$$

$$x_{2,t} = b_{2,0} + \sum_{i=1}^{k} a_{2,i}x_{1,t-i} + \sum_{i=1}^{k} b_{2,i}x_{2,t-i} + \sum_{i=1}^{k} c_{2,i}x_{3,t-i} + u_{2,t}$$

$$x_{3,t} = c_{3,0} + \sum_{i=1}^{k} a_{3,i}x_{1,t-i} + \sum_{i=1}^{k} b_{3,i}x_{2,t-i} + \sum_{i=1}^{k} c_{3,i}x_{3,t-i} + u_{3,t}$$

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.

[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.

[11]The authors use the asymptotic chi-square test developed by Sims (1980) for the determination of lag order.

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 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

[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).

[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.

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 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.

[15]For instance, Kupiec and Ramirez (2013) apply a panel VAR to study financial panics across the various states.

[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.

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 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

[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.

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 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

Financial Systems

Caroline Fohlin

Contents

Abstract

This paper 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 paper 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 similar functions; thus arguing

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*,
DOI 10.1007/978-3-642-40406-1_7

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 general meetings of their current

[5]This section is based on Fohlin (2012).

[6]See Fohlin (2012, Chapter 3) and on Germany specifically see Fohlin (2005, 2007a, b).

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 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."

---

[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.

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.

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

[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 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, 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.

[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

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 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

[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.
[18]See Amidei and Giordano (2010).

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 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).

[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).

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 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

[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.

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 capital organizations. Venture capitalists fund predominantly untested projects for which the market has yet to enter the picture, and therefore asymmetric information

[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.

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 about the idiosyncratic influences of the government than the innate substitutability of financial markets and universal banks.[29]

[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.

[29]See Fohlin (2000).

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 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]

[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.

[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.

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. 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

[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).

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 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

[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 %.

[35]Sylla (2006) offers a critical appraisal of the Rajan and Zingales "great reversals" thesis.

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, 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

[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.

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)

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]

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

[37]For much more detail, see Fohlin (2012), Chapter 7, on which this section is based.

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 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.

[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.

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.

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

[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).

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 predepression 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 the

[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).

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

[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).

short 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

[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.

time of 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 Groβbanken 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

Part V

Innovation

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 or the future. Consideration of these

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*,
DOI 10.1007/978-3-642-40406-1_26

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 paper. 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

[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.

by the 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 paper 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 paper 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*,
DOI 10.1007/978-3-642-40406-1_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

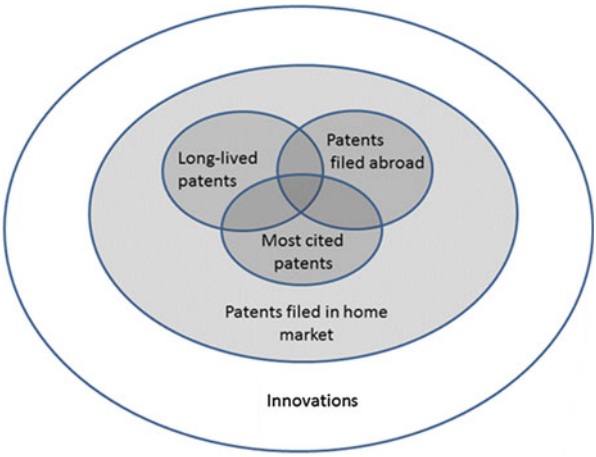

Fig. 1 Identifying valuable patents

methods. We already know that the set of patents filed in a particular country is only a more or less large subset 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.

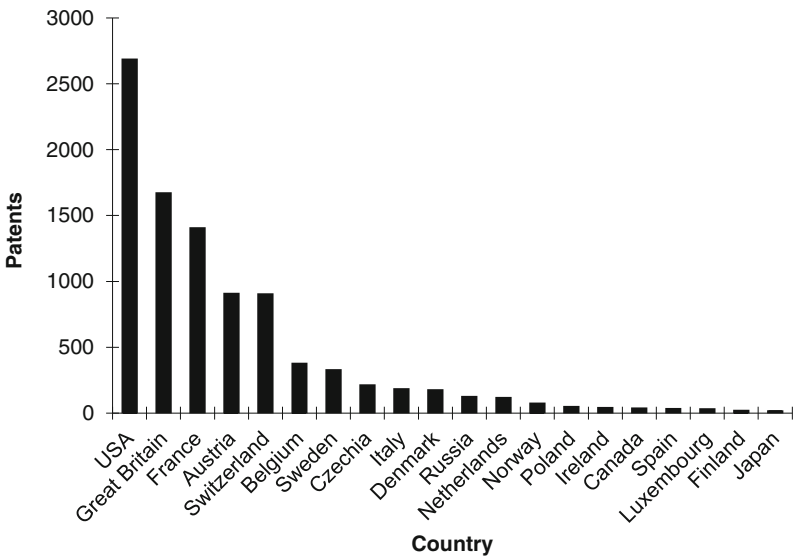

Fig. 2 Long-lived German patents of the 20 most innovative foreign countries before the First World War (Source: Degner and Streb 2013, p. 24)

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 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 from

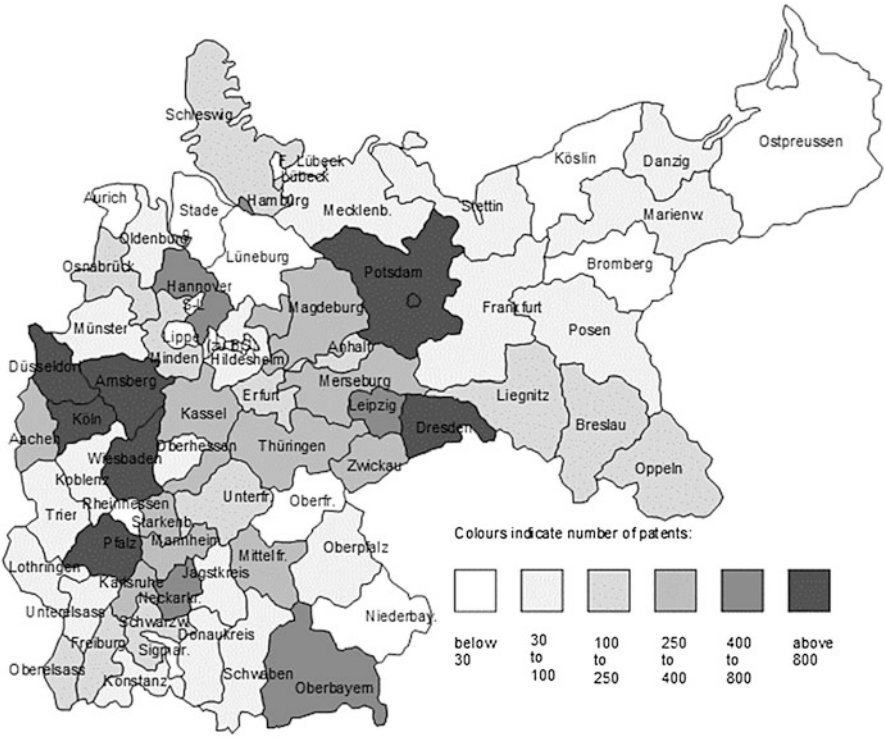

Fig. 3 The geographical distribution of high-value patents in Germany, 1878–1914 (Source: Streb et al. 2006, p. 364)

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

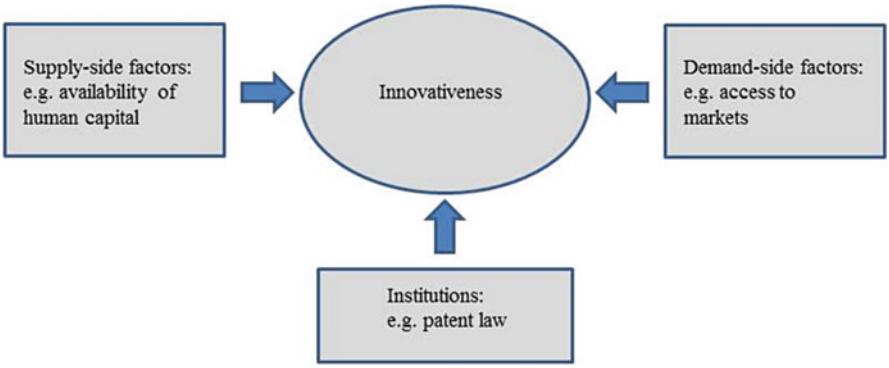

Fig. 4 Determinants of innovation

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

Part VI

Statistics and Cycles

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

T. Rahlf (✉)
German Research Foundation, Bonn, Germany
e-mail: thomas.rahlf@dfg.de

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
DOI 10.1007/978-3-642-40406-1_16

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. 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:

[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).

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
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

[5]Stigler (1986, p. 122).

[6]See, for example, Yule (1895, 1896a, b) and Pearson (1898).

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 "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

[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).

frequency curves and finally found one in the form of his chi-squared (χ^2) test, which he made public in 1900.

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

[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).

socioeconomic data carried out by Yule or studies, such as those of Gosset, of the correlations between 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

[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.

definition subscribed to 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

[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.

> no scientific worker has a fixed level of significance at which from year to year, and in all 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 considering an alternative hypothesis. Neyman and Pearson surely did not intend to

[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.

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):

1. A hypothesis A is, in extreme cases, certainly true or certainly false. Intermediate degrees of belief in A are permitted.

[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."

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:

> 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]

[31] Ramsey (1931a, b).

[32] Keynes (1921, S. 4), quoted by Kyburg/Smokler (1964, p. 9).

[33] Cf. de Finetti (1937).

[34] De Finetti (1981, p. 657).

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 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.

[35] Jeffreys (1939, p. 8).

[36] Ibid., p. 401.

[37] Although the hypothesis can still be false in terms of rule 4.

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 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]

[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).

[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).

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 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.

[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).

[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."

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 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

[51] See above, p. 86f.

[52] Cf. Stegmüller (1973, p. 32ff, particularly p. 37).

[53] Howson (1995, p. 27).

[54] See above, p. 99.

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 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

[55] Lindley (1991, p. 493).

[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.

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 socioeconomic 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 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 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.

[58]One of the few exceptions, who assigned independent significance to the trend, was S. Kuznets. See Kuznets (1930a, b) in particular.

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]

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

[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)).

[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.

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 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.

[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).

[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.

"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 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,

[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?".

[67] Haavelmo (1994, p. 75).

[68] Haavelmo (1944, pp. 13, 22f, 24).

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?

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

[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).

[70]See, for example, Sims (1980).

[71]They were also subject to the same statistical limitations, such as stationarity and linearity.

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]

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]

[72] Zellner (1971, p. 11).

[73] See references in Rahlf (1998).

[74] Leamer (1994, p. ix).

[75] Keuzenkamp (1995, p. 243) therefore uses, for Hendry's approach, the more apposite term "diagnostic *checks*" rather than "diagnostic *tests*."

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 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,

[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.

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 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

[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).

[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).

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 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

[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.

[90]Conrad/Meyer (1957, p. 544). Specific examples can be found in Conrad/Meyer (1964).

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.

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

[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.

[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.

> 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]

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

[93]Kalman (1982a, p. 19f). Original author's italics.

[94]Ibid., p. 20.

[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.

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.

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.

[98]Ibid., p. 171.

[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.

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', 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

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*,
DOI 10.1007/978-3-642-40406-1_21

allow all components to be stochastic and 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,

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.

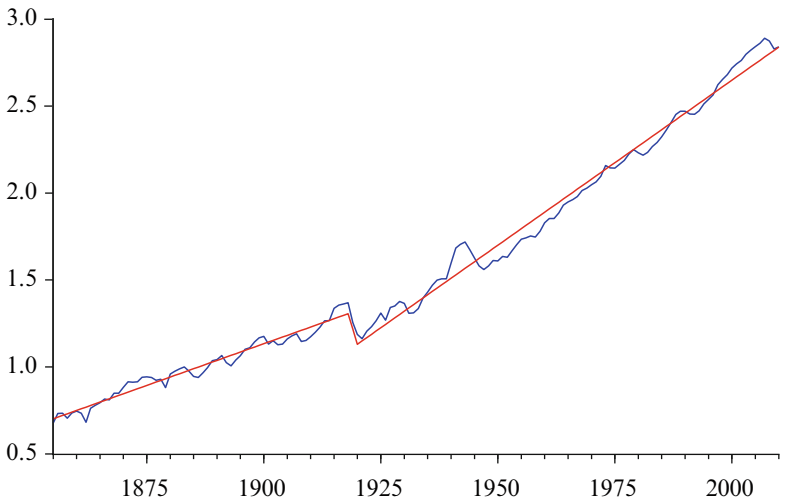

Fig. 1 Logarithms of UK per capita GDP for 1855–2010 with a fitted segmented linear trend having breaks at 1918 and 1920

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

$$x_t = \underset{(0.0095)}{0.7039} + \underset{(0.0006)}{0.0095}\, t - \underset{(0.1311)}{0.2112}\, d_{1t} - \underset{(0.0538)}{0.4353}\, d_{2t} - \underset{(0.1158)}{0.7703}\, d_{3t} + \underset{(0.0017)}{0.0210}\, td_{1t} + \underset{(0.0005)}{0.0216}\, td_{2t} + \underset{(0.0008)}{0.0239}\, td_{3t} + \hat{u}_t \qquad R^2 = 0.9967$$

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

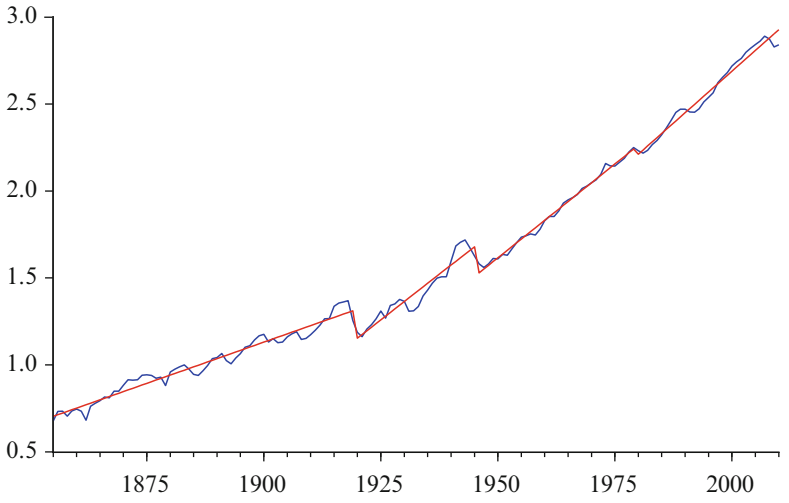

Fig. 2 Logarithms of UK per capita GDP for 1855–2010 with a fitted breaking linear trend having breaks at 1918, 1946, and 1979

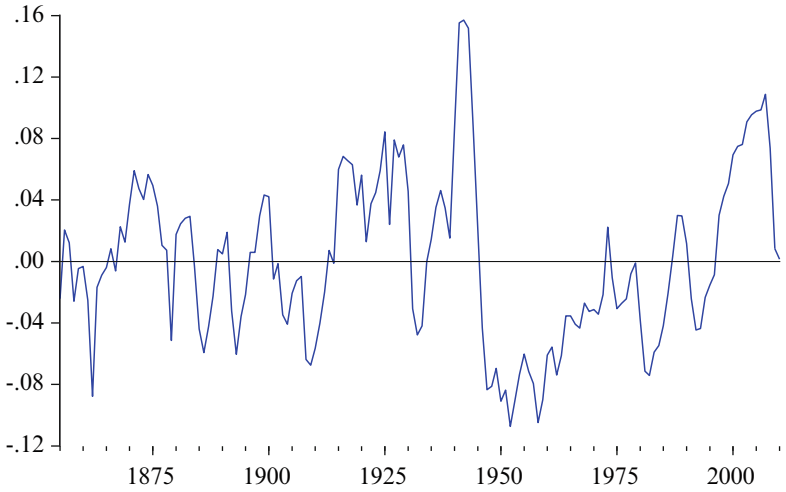

Fig. 3 Cyclical component of UK per capita GDP for 1855–2010 obtained by residual from the segmented trend model

$\psi_t = 1.060\psi_{t-1} - 0.220\psi_{t-2} + e_t$, the roots of this autoregression are 0.78 and 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

$$|a(\omega)|^2 = \left(\sum\nolimits_j a_j \cos \omega j\right)^2 + \left(\sum\nolimits_j a_j \sin \omega j\right)^2$$

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

$$a_{B,0} = a_{c,2,0} - a_{c,1,0} = \frac{4}{3} - \frac{1}{4} - \left(\zeta_{c,2,n} - \zeta_{c,1,n}\right) \tag{16}$$

$$a_{B,j} = a_{c,2,j} - a_{c,1,j} = \frac{1}{\pi j}\left(\sin\frac{4\pi j}{3} - \sin\frac{\pi j}{4}\right) - \left(\zeta_{c,2,n} - \zeta_{c,1,n}\right) \quad j = 1, \ldots, n$$

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).

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

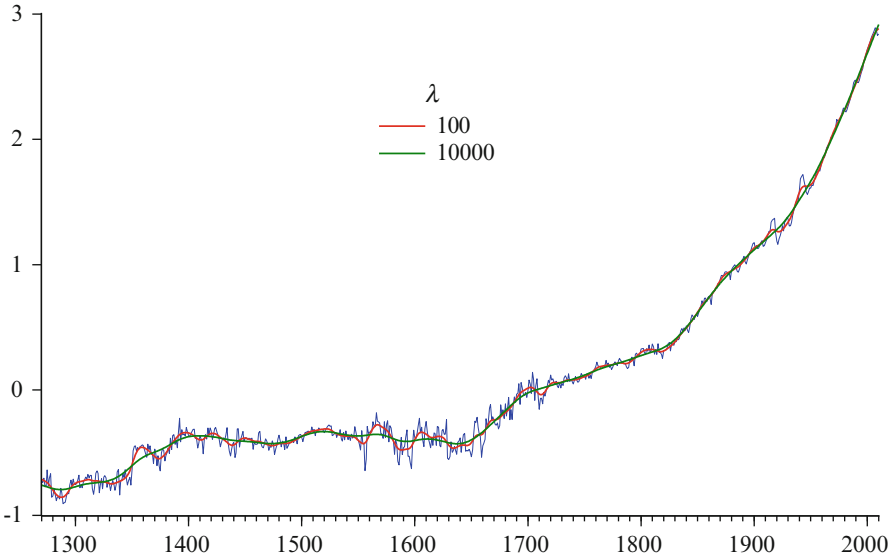

Fig. 4 Logarithms of British per capita GDP for 1270–2010 with H-P trends for $\lambda = 100$ and 10, 000

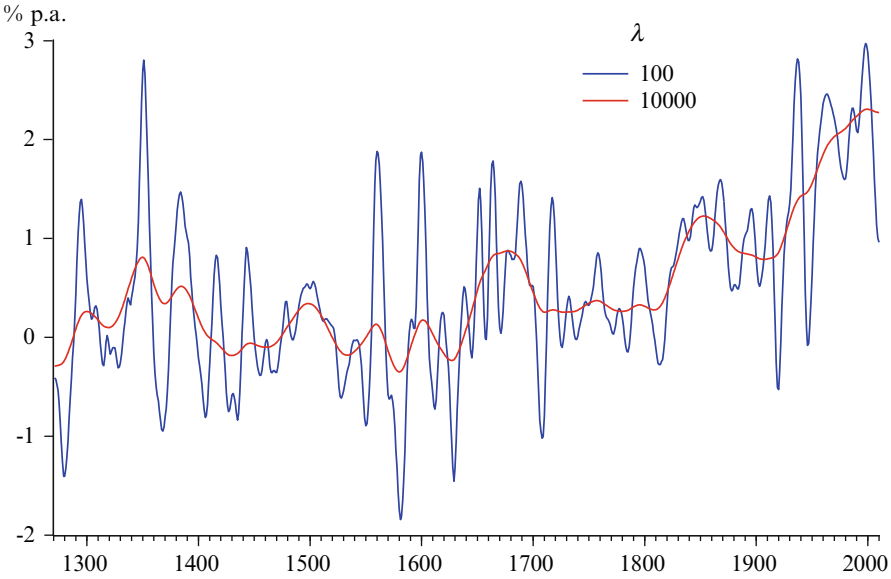

Fig. 5 Trend growth rates of British per capita GDP for $\lambda = 100$ and 10, 000

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 (MMSE) estimates of the components are $\hat{\mu}_t = a_\mu(B)x_t$ and $\hat{\psi}_t = x_t - \hat{\mu}_t = \left(1 - a_\mu(B)\right)x_t = a_\psi(B)x_t$, where

$$a_\mu(B) = \frac{(1+B)^r}{(1+B)^r + (1-B)^m}$$

and

$$\left|a_\mu(B)\right| = \frac{|1+B|^{2r}}{|1+B|^{2r} + \lambda|1-B|^{2m}} \tag{17}$$

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 $\left((1+2\lambda) + \sqrt{1+4\lambda}\right)/2\lambda$ (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 $\theta_q(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$ 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.

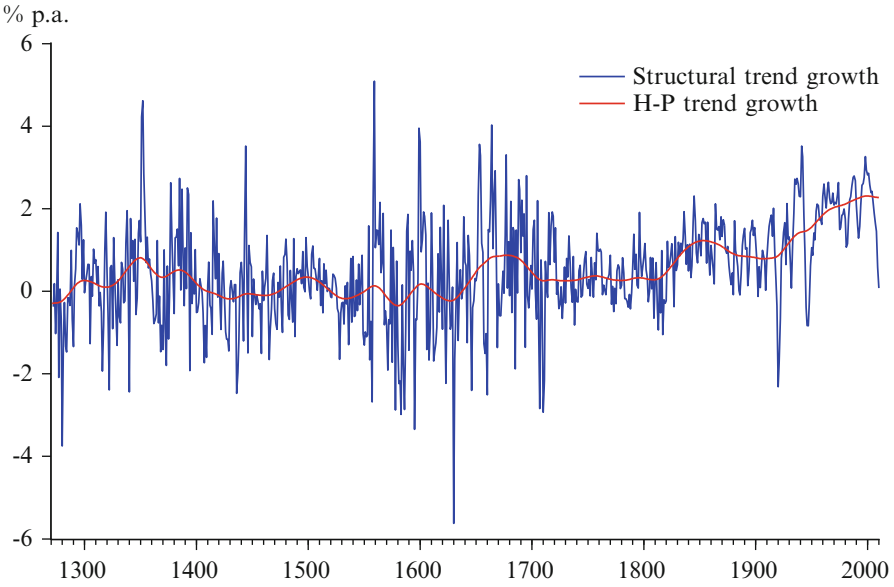

Fig. 6 Trend growth for British per capita GDP, 1270–2010, computed from a structural model, compared to H-P trend growth

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:

$$\mu_t = \mu_{t-1} + c + \zeta_t \qquad \zeta_t \sim WN\left(0, \sigma_\zeta^2\right) \tag{22}$$

$$\psi_t = \phi_1 \psi_{t-1} + \phi_2 \psi_{t-2} + \kappa_t \qquad \kappa_t \sim WN\left(0, \sigma_\kappa^2\right)$$

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

$$\left(1 - \phi_1 B - \phi_2 B^2\right)(\Delta x_t - c) = \left(1 - \theta_1 B - \theta_2 B^2\right) a_t \tag{23}$$

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

$$\mu_t = \mu_{t-1} + 0.486 + 0.554 a_t$$

and the ARMA(1,1) cycle is

$$(1 + 0.303B)\psi_t = (1 + 0.622B)(0.446 a_t)$$

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

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t \sim MWN\left(\mathbf{0}, \Sigma_{\boldsymbol{\eta}}\right) \tag{24}$$

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\zeta}_t \qquad \boldsymbol{\zeta}_t \sim MWN\left(\mathbf{0}, \Sigma_{\boldsymbol{\zeta}}\right)$$

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

$$y_t = \boldsymbol{\xi}'_{\perp,1}\mathbf{z}_t + \widetilde{\boldsymbol{\pi}}_0 + \widetilde{\boldsymbol{\delta}}\boldsymbol{w}_t + \widetilde{\boldsymbol{\varepsilon}}_t \tag{27}$$

where $\widetilde{\boldsymbol{\pi}}_0 = \boldsymbol{\xi}'_\perp\boldsymbol{\pi}_0$, 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

Part VII

Government

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*,
DOI 10.1007/978-3-642-40406-1_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

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

(*continued*)

Table 1 (continued)

	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
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/

were likely to get 40–60 % less in nominal terms than in the late 1920s in some cities.[1]

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

[1]See the new estimates developed by Fishback and Kollmann (2014).

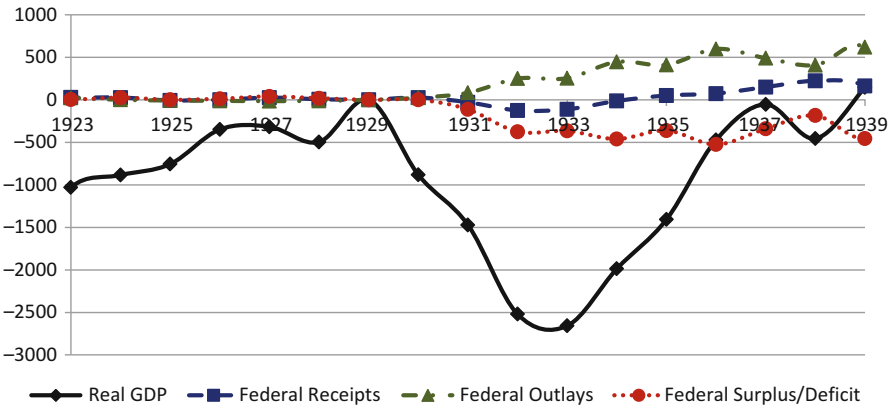

Fig. 1 Per capita gross domestic product and federal government revenue, outlay, and surplus/deficit value minus the 1929 value, (in 2013 Dollars), 1923–1939 (Source: Fiscal years ran from July 1 through June 30, such that the 1930 value covers the period July 1, 1929, through June 30, 1930. Federal 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)

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.

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 overinvestment 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

[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).

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 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]

[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.

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 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

[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.

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]

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

[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.

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.

[9]Ellen McGrattan (2012) develops a model that shows the negative consequences of taxes on corporate income and dividends in the early 1930s.

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.

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.

[10]See the Inaugural Address of Franklin Delano Roosevelt (1933), downloaded on 10 June 2010, from http://www2.bartleby.com/124/pres49.html.

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 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

[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.

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 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

[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).

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]

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.

[13]Eichengreen (1992), Temin (1989), Temin and Wigmore (1990), and Kindleberger (1986) talk extensively about the move off of the gold standard.

[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).

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]

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]

[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.

[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.

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 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

[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).

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]

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

[19]See Costa (1999), Stoian and Fishback (2010), Parsons (1991) Balan-Cohen (2009), and Friedberg (1999).

[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).

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.

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

[22]For analysis of the HOLC, see Courtemanche and Snowden (2011), Fishback et al. (2011, 2013), Harriss (1951), and Rose (2011).

[23]See Temin and Wigmore (1990) and Eggertsson (2008, 2012) for this argument.

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. 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

[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.

[25]For an economic overview of the changes in union policy, see Freeman (1998).

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 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

[26]For additional discussions of more New Deal programs and long-run changes relative to the past and present, see Fishback and Wallis (2013).

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
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 article 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

J. Eloranta (✉)
Appalachian State University and University of Jyvaskyla, Boone, NC, USA
e-mail: elorantaj@appstate.edu

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
DOI 10.1007/978-3-642-40406-1_4

will change in the near 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,

[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/

the 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

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

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 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 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?

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

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 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,

[3]See Fishback and Kachanovskaya (2010) as an example of this debate.

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).

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

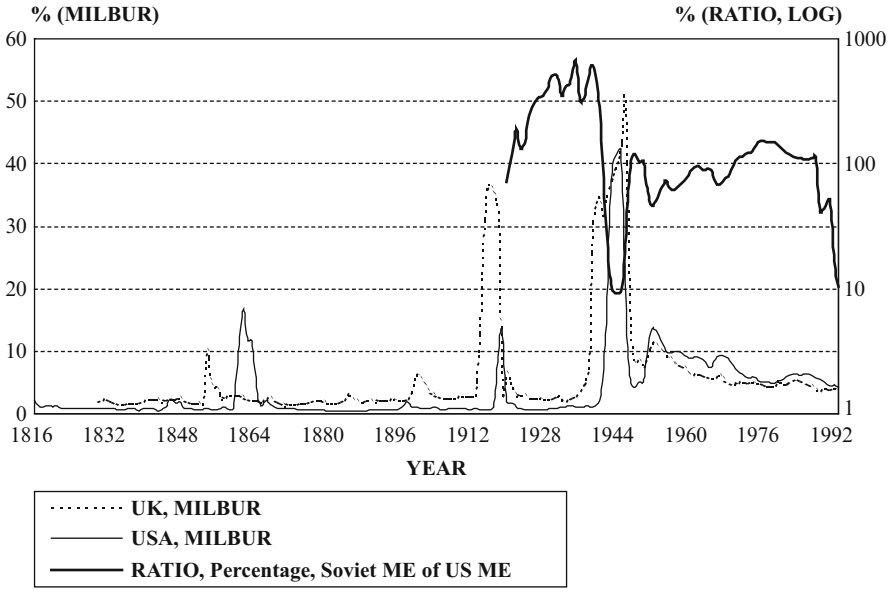

Fig. 1 Military burdens (= MILBUR) of the United States and the United Kingdom and the Soviet military spending as a percentage of the US military spending (*ME*), 1816–1993 (Sources: see Eloranta (2005) for details on the sources and methods of calculation)

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 article 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

Index

C. Diebolt, M. Haupert (eds.), *Handbook of Cliometrics*,
DOI 10.1007/978-3-642-40406-1