Joshua Hall *Editor*

Explorations in Public Sector Economics

Essays by Prominent Economists

Springer

Explorations in Public Sector Economics

Joshua Hall
Editor

Explorations in Public Sector Economics

Essays by Prominent Economists

Springer

Editor
Joshua Hall
Department of Economics
West Virginia University
Morgantown, WV
USA

ISBN 978-3-319-47826-5 ISBN 978-3-319-47828-9 (eBook)
DOI 10.1007/978-3-319-47828-9

Library of Congress Control Number: 2016955917

Printed on acid-free paper

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

To Buchanan and Tullock and Downs and Riker and too many others to mention. We stand on your shoulders.

Preface

The origins of this book go back to a dinner at the house of Lowell E. Gallaway. After dinner, Lowell and I retired to his den while our better halves stayed in the dining room to chat. I had only recently started my academic career and I was eager to pick Lowell's brain about the publication process. After all, by that time he had only published in six decades! Much of our conversation focused on his book with Richard K. Vedder, *Out of Work*, and the difficulty they had in getting some of the individual chapters published as refereed journal articles. This caused me to ask Lowell if he ever had a favorite paper that never got published in some form. Of course he did and he was glad to give me a yellow, faded copy of a 1979 working paper.

While driving home that evening I thought about the paper sitting in my back seat. As far as I knew, the only copies that existed were hard copies in private hands. What a shame that other scholars had no way to access these insights. Upon returning home, I confirmed that the only traces of the paper on the Internet were in citations in papers published around the time the paper was originally written. The fact that the paper was cited at the time suggests that the paper had value for scholars then and thus would be likely to have value for future scholars, if only those following the footnotes to papers published around 1980.

I began to take notes of papers that were cited several times as conference presentations or working papers but never published. A public economist by training, these papers tended to be in the field of public economics, especially public choice. When I went to conferences and would talk to more experienced economists, I began to ask them if they had a favorite paper that never quite found a home. Inevitably they did.

After several years, I found myself in possession of eleven interesting papers by prominent economists on important topics in the area of public economics. These papers had numerous citations and in some cases significant media attention and yet without intervention they might be lost to history. At a minimum, the fact that most of them were not available through libraries or the Internet meant that they could not be of use to scholars in public economics. Thus, this book was born, and I am incredibly grateful to Springer and Lorraine Klimowich for their help and patience in bringing this product to fruition.

The papers in this volume deal with issues that are at the core of the economics analysis of politics. For example, the volume begins with a paper on voting by the late Nobel Laureate Gary S. Becker and Casey B. Mulligan. Why people vote has been a prominent topic in the economic analysis of politics since before Downs seminal analysis in *An Economic Theory of Democracy*. Similarly, the concepts of public goods and externalities are taught in every undergraduate public economics class. Bruce L. Benson and Roger E. Meiners deepen our understanding of these topics with their respective chapters. The volume contains a variety of different methods—theory, empirical, experimental, and historical. However, they all speak to issues (e.g., political corruption, media and presidential voting, school competition, and global warming) that are important parts of public policy and thus the field of public economics today. I hope you agree.

Morgantown, WV, USA Joshua Hall
September 2016

Acknowledgements

I would like to thank the Center for Free Enterprise at West Virginia University for summer support that provided me with the time necessary to finish this project. I would like to thank Richard K. Vedder and Lowell E. Gallaway for introducing me to public sector economics as an undergraduate and graduate student at Ohio University. Kai Cher Tay and Eric Mason provided invaluable research assistance.

Contents

Contributors

Bruce L. Benson Florida State University, Tallahassee, FL, USA

Gary S. Becker University of Chicago, Chicago, IL, USA

J.R. Clark The University of Tennessee at Chattanooga, Chattanooga, TN, USA

R. Morris Coats Nicholls State University, Thibodaux, LA, USA

Thomas R. Dalton Department of Economics, Eller College of Management, University of Arizona, Tuscon, AZ, USA

Arthur Denzau Virginia Polytechnic and State University, Blacksburg, VA, USA

Lowell E. Gallaway Ohio University, Athens, OH, USA

Daniel Houser ICES, George Mason University, Fairfax, VA, USA

Lawrence W. Kenny University of Florida, Gainesville, FL, USA

Robert A. Lawson Southern Methodist University, Dallas, TX, USA

Dwight R. Lee Southern Methodist University, Dallas, TX, USA

Gary D. Libecap University of California, Santa Barbara, Santa Barbara, CA, USA

Edward J. Lopez Western Carolina University, Cullowhee, NC, USA

Roger E. Meiners University of Texas at Arlington, Arlington, TX, USA

Casey B. Mulligan University of Chicago, Chicago, IL, USA

William Robert Nelson Jr. Eqis Capital Management, San Rafael, CA, USA

William S. Peirce Case Western Reserve University, Cleveland, OH, USA

Katie Sherron Florida State University, Tallahassee, FL, USA

Russell S. Sobel The Citadel, Charleston, SC, USA

Richard K. Vedder Ohio University, Athens, OH, USA

Chapter 1
Is Voting Rational or Instrumental?

Gary S. Becker and Casey B. Mulligan

Abstract A fully rational choice approach to politics does not closely resemble modern models of voting behavior that purport to be applications of the economists analysis of rationality to the political sector. For these models do not build voting choices on the fragility of preferences about how to vote, which we show to be a basic implication of the voters paradox. Building a simple model on the fragility of preferences about how to vote delivers an number of different and realistic implications for the demand for public policies and political candidates, the supply of public policies and political candidates, and, ultimately, the determinants of public policy. The model explains why so many studies have found voters not voting in their (narrowly defined) self-interest, why minorities are not exploited under majoritarian voting, why interest groups have an important influence on public policy, why public decisions are so weakly correlated with voting rules, and why conformity is more common in political than private life.

1.1 Is Instrumental Voting Rational?

The modern political economics literature is dominated by voting models, where participants are alleged to be rational in the sense that they vote according to the election outcome that would give them greatest utility, and they are forward looking. For example, 22 studies of public decision-making1 were published in the *American Economic Review* during the 5 years prior to writing this paper.[1] Thirteen (59 %) studies modeled the process with each voter individually voting for the policy serving his self-interest. Another eight (for a total of 95 %) modeled voters voting in their

[1] Those published during the years 1994–98, assigned *Journal of Economic Literature* classification "Economic Models of Political Processes," and had a model of the political process.

G.S. Becker · C.B. Mulligan (✉)
University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA
e-mail: c-mulligan@chicago.edu

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_1

self-interest as groups. The remaining article had social planners making public decisions.[2]

Models of instrumental voting have had a number of implications, but have so far weak empirical support. Examples include Meltzer and Richard's (1981) prediction that more skewed (and probably more unequal) income distributions lead to more government redistribution while Peltzman (1980), Benabou (1996), and many others have found little (or wrong!) cross-country and time-series correlations of inequality and the size of government. Or, as another example, instrumental voting models predict policy outcomes to be highly sensitive to the rules of the voting game, and that cycling and other odd behavior may result from an electoral process.[3] One implication of such results is that different policies ought to be adopted by democratic and nondemocratic governments because by definition one is influenced by a voting game and the other is not, yet little empirical difference is found between these two governments in, say, social security spending in cross-country and time-series studies once one or two basic economic or demographic variables are held constant.[4] Brennan and Hamlin (1998) suggest that, in instrumental voting models, voter turnout ought to be lowest among those with preferences near the center because they have the least at stake in the election.

Like Brennan and Hamlin, we believe that voting theories have utilized models of rationality developed for the market sector that are inappropriate for political behavior. As a result, we contend many of the results obtained are fragile and of modest use in understanding political choices. This may explain why formal public choice theory has far outstripped empirical relevance.

It is well known that it unreasonable to expect that a single voter in a majoritarian election with more than a few voters would affect the outcome. This has led to a puzzle about why so many people vote in elections. It is less often recognized (Brennan and Lomasky 1983 and Caplan 2001 are some exceptions) that it would be just as puzzling if, given that someone decided to cast a vote, that he cast the vote in his personal self-interest in the same way he would make purchasing decision in the marketplace. We show how this reasoning has implications for the demand for public policies and political candidates, the supply of public policies and political candidates, and, ultimately, the determinants of public policy.

To show the contrast between political and market behavior, consider a stripped down version of utility maximization in the market sector. Suppose utility of the representative person i depends on the consumption of one unit of goods X or Y, and advertising for or against X:

---

[2] *Public Choice* had a similar distribution for the four years 1995–98; 77 % individual voting in self-interest, 17 % interest group models, and 6 % social planner models.

[3] (Myerson 1995, p. 79ff) shows how political strategies are sensitive to the rules of the voting game in instrumental voting models. He also suggests on p. 77 that different political strategies would be associated with different *policy* outcomes, but he does not offer a declaration of this point.

[4] Eg., (Easterly and Rebelo 1993, p. 436), Lindert (1994), Pampel and Williamson (1989, p. 102), and Jackman (1975).

$$U(X, Y, A_x, A_y)$$
$$\text{where } \frac{d^2U}{dXdA_y} \equiv U_X A_y < 0, U_X A_x > 0,$$
$$\text{and } U_X \gg U_Y \, for A_X = A_Y = 0 \tag{1.1}$$

A persona can get either X or Y by buying them, so there is a market production function

$$X = B(X), \tag{1.2}$$

where B is the purchase function. We could add $B(X)$ to the utility function to get

$$U(X, Y, B(X), A_x, A_y). \tag{1.3}$$

It is important that buying X is necessary and sufficient for consuming X. Hence, while buying and consuming X can be different sources of utility, they are linked together (according to the function B). But economists usually assume the act of purchase does not itself have utility, so the inclusion of B in the utility function is ignored (or implicitly absorbed into the parts of U relevant to X).

If the cost of X and Y are the same and the $A's = 0$, i buys X because i gets so much more utility from X than from Y. An increase in A_y would reduce i's relative valuation of X, but for i to buy Y, the advertising for Y must be sufficiently powerful to overcome i's tendency to get much more utility from X than from Y.[5]

The typical political choice argument sets up a similar function to Eq. (1.1), but usually assumes that $A_x = A_y = 0$. Then if i gets more utility from X, the assumption is that i votes for candidates who offer X either prospectively or in the past (see, e.g., Persson and Tabellini 1999). If one candidate proposes X and the other Y, and with majority rule, which candidate wins depends on whether the majority of voters prefers X or Y.

Of course, it is recognized that the political process differs from the market because of the collective choice in political decisions. It has also been recognized for centuries that it does not pay in any instrumental way for people to vote because one voter's influence over the outcome of a large election is likely to be negligible. Still, it is almost always assumed that given that a person does vote, he votes for the policy that maximizes his utility in Eq. (1.1), given the negligible values of A assumed.

1.2 A Model of Rational Voting

To model the voting process, we replace the market utility function in (1.3) with the political utility function:

[5]The utility function (1.3) can easily include "random" components to reflect that advertising has an uncertain effect on a person's preferences.

$$U(X, Y, V\{X \text{ or } Y\}, A_x, A_y). \tag{1.4}$$

where V refers to whether i votes for a candidate who will implement X or Y. We still assume that i intrinsically prefers X to Y, but now the voting process V replaces the buying process B. In this context, we define "advertising" is a broad sense including, but not limited to, television and other media designed by political candidates to affect voter preferences. Other relevant examples and media supplied groups (other than the candidates themselves) who have an interest in policy outcomes, conversations among friends, teaching of children, etc.

In the marketplace, the purchase of an object usually leads to the consumption of that object, but in voting there is only a negligible connection between how one votes and political outcomes. The *voting process* becomes of primary relevance here, "... voting is not merely an instrumental exercise designed to raise the probability of victory for the preferred outcome; voting also affords the individual direct consumption benefits. Returns accrued to a vote independently of the effect on political outcomes." (Brennan and Lomasky 1983, p. 188).[6]

Suppose we assume that i feels better by voting for X rather than Y when i intrinsically prefers X to Y. It might be that the preference for consuming X over Y is not sensitive to advertising because the intrinsic utility provided by X greatly exceeds that provided by Y. In the marketplace, people influenced by advertising must pay by getting goods they do not intrinsically value as highly. Political advertising by proponents of Y need not change *i's* feelings about the utility he gets from X relative to Y, but only the utility he gets from voting for Y rather than X. This utility is likely to be quite sensitive to advertising because voting for X or Y has no consequences for whether i actually gets X or Y through the political process.

U_v, the marginal utility of voting for X, is assumed to be positive when political advertising is negligible. But our discussion implies that U_v is very sensitive to spending by relevant interest groups, so that in particular $U_V A_y$ is sizeable and negative. Similarly, $U_V A_x$ is sizeable and positive. That is, rational choice theory suggests, although does not prove, that these cross derivatives with respect to voting and political advertising are much larger than the corresponding cross derivates in the marketplace between advertising by X and Y, and the utilities from consuming X or Y.

If U_v is small, and if the cross derivatives with respect to political advertising are large, then political contests would be decided not by the distribution of preferences for X and Y in the voting population- the usual assumption in voting theories. Rather, they would be decided by the distribution of preferences to vote for X and Y, inclusive of the powerful effects of political advertising to change votes toward or away from candidates promising X.

That is, a fully rational choice approach to politics does not closely resemble modern models of voting behavior that purport to be applications of the economists analysis of rationality to the political sector. For these models do not build voting

[6]Brennan and Lomasky (1983) also demonstrate the difference between the market buying process and the voting process with some simple numerical examples.

choices on the fragility of preferences about how to vote, which is a basic implication of the insight that individual votes have a negligible influence over political outcomes.

A rational approach to political voting would emphasize that the spending of time, money, and energy by interest groups and politicians on influencing how people vote, not on influencing their underlying preferences, has a decisive influence on votes and political outcomes, almost regardless of the distribution of underlying preferences for different policies.

Political battles then become a battle between the spending by different interest groups and political alliances, as represented in our analysis by A_x and A_y. Voting can be treated as functions of these spendings, and would to a first approximation be the number of votes for X would be rising in A_x and falling in A_y, and presumably votes would be sensitive to these expenditures.

Schumpeter (1942, p. 262) Downs (1957), and more recently Citrin and Green (1990) have emphasized another difference between market and political environments decisions are more complicated and less tangible in the latter and deduce that self-interested behavior should be much more common in the market environment. Our approach has a similar flavor, but emphasizes the different constraints faced by individuals in the two environments, rather than the different computational problems encountered.

1.3 Other "Demand Side" Implications of Rational Voting

1.3.1 An Individual Does Not Typically Vote in His Self-interest

There is a lot of evidence that individuals often do not vote for the policies serving their self-interest. (Brennan and Lomasky 1983, p. 188) "...given [voter] turnouts, we cannot explain the direction of individuals votes in terms of bringing about a personally profitable outcome."

For example, micro studies of attitudes and votes on policy proposals to racially integrate schools by busing children from one neighborhood to another have found little correlation between a white persons vote and his having children in school or having children who would be affected by the policy proposal (Sears et al. 1979; Citrin and Green 1990). The youths most opposed to the Vietnam War were females and draft deferred men, who presumably had the least self-interest (Sears et al. 1979, p. 370). Little correlation has been found between opposition to the Vietnam war and having friends or relatives in the military at the time (Sears and Funk 1991). Aggregate votes for the incumbent president tend to be correlated with aggregate economic performance (see studies surveyed by Sears and Funk (1991, p. 17)), which seems to be consistent with self-interest. However, there is little micro correlation between incumbent votes and recent income growth (Sears and Funk 1991). There are number of studies around the world showing that opposition to national taxes is only

weakly correlated with taxes paid (see studies surveyed by Sears and Funk (1991, p. 34ff)). Women working, desiring further schooling, or divorced were not particularly opposed to the Human Life Amendment (see studies surveyed by Sears and Funk (1991, p. 39ff)). Those receiving public services in California and Massachusetts were not particularly opposed to Californias Proposition 13 and Massachusetts Proposition 2 1/2, respectively.[7]

Because a husband and his wife are similar to each other in many ways and are economically interdependent, one expects a husband and his wife to have a great deal in common regarding the public policies they perceive to be in their self-interest. However, empirical studies of voting by husbands and wives have found little correlation between the way a husband and his wife votes, which suggests that at least one of them often votes other than his or her self-interest.

This is not to say that one cannot find evidence consistent with a tendency toward instrumental voting in some instances. It has been shown, for example, that lower income and working-class voters tend to favor left-wing parties (Citrin and Green 1990, p. 5). Poor smokers tend to oppose cigarette taxes the most (Citrin and Green 1990, p. 19). Homeowners tended to support and public employees tended to oppose Californias Proposition 13 and Massachusetts Proposition 2 1/2 (Sears and Funk 1991). One study of busing policy did find the whites most affected by the policy to be opposed. It should also be noted that many of the empirical tests rejecting self-interest are conducted by political scientists who have less stake in the defending the hypothesis. Many economists would also question some of the measures of self-interest used in the literature because tax incidence theory sometimes predicts that winners and losers from policy can be very different from those who receive more government benefits and those who pay more taxes, respectively. Many of the studies also ignore the possibility that Tiebout-style sorting may cause those most affected by policy to be different from those less affected. Nevertheless, the hypothesis that a great deal of voting is not in the voters self-interest is a very difficult one to reject.

1.3.2 Personal Costs and Benefits Matter

Empirical studies have found election turnout can be fairly well predicted by proxies for the costs and benefits of voting. For example, poll taxes and bad weather are associated with low voter turnout. But such a finding is not a defense of the allegedly rational approach to voting, because our approach also predicts that personal costs of voting will discourage such behavior. Here the analogue with market behavior is much better because a persons act of voting requires his payment of these costs. In contrast, the link between the policy or politician for whom he votes and the election outcome-related costs he pays is negligible.

[7] See studies surveyed by Sears and Funk (1991, p. 34ff). Propositions 13 and 2 1/2 were proposals to cap or cut some important sources of state revenue.

Close elections are associated with higher voter turnouts. This finding is qualitatively consistent with the allegedly rational approach, because a persons vote is much more likely to be decisive when the election is anticipated to be close. But our approach is consistent with high turnout in close elections, as long as political advertising, conformity, and the other forces affecting how a person casts his vote are more intense in close elections.[8]

One tougher test of the instrumental approach is quantitative. Because the probably of a decisive vote (in a majoritarian election with candidates) is proportional to $e^{-2(N-1)q^2}$, where N is the number of voters and q is a measure of closeness of the election (the expected gap between the elections outcome and 50 %), the expected benefit is more than exponentially related to closeness. Furthermore, the effect of closeness on expected benefit depends on the number of voters. The instrumental approach therefore predicts voter turnout to be more than exponentially related to closeness and for the marginal effect of closeness to decrease with size of the election, unless the "demand" for voting were also more than exponentially related to the expected benefit so as to "undo" the exponential and interactive relationships between expected benefit, closeness, and election size a quantitative relationship between behavior and price which is rarely seen in market behavior. There are other tougher tests of the instrumental voting models predictions for turnout and closeness. For example, third-party turnout should decline when there is a close race between the two favored candidates.

1.4 "Supply Side" Implications of Rational Voting

1.4.1 Candidates Matter as Much as Policy

We also expect votes to be determined by personal characteristics of the candidates in addition to the policies those candidates advocate.

1.4.2 Information or Misinformation?

Advertising can have an informative role in the market sector, but do we expect the same in the political sector? Our approach suggests that the primary role of political advertising will not be to inform voters of the consequences of policies, because any single vote has practically no affect on policy. Rather we expect the "information" in political advertising to pertain to the character of candidates, and other issues only weakly related to the consequences of policies because the purpose of the advertising

[8] We discuss below the (unsurprising) prediction that political advertising will be more intense in a close election.

is to affect a voters preference for casting his vote in one way or another, not to affect his policy evaluation in one way or another.

Voters certainly have less incentive to verify information provided to them in political advertisements than that provided to them in market sector advertisements, which suggest that misinformation should be more common in political advertising than in market sector advertising. Competition is a force in the political sector as it is in the market sector (Wittman 1995 emphasizes this similarity); not all political information will be misinformation as long as there is the potential for competition in the provision of information.

1.4.3 Supply in Close Races

For the same reason that the instrumental benefits of voting are highest for close elections, the benefits of political advertising are highest in close elections.[9] Cox and Munger (1989) show that, for 1982 U.S. House elections, closeness, turnout, and political campaign expenditures are positively correlated. They suggest that the causality is, at least in part, from closeness to expenditures and from expenditures to turnout. Matsusaka and Palda (1993) adds some additional evidence, showing how closeness and turnout are correlated for congressional elections, but not for California ballot propositions for which national party advertising is not particularly stimulated by closeness.

The number of voters may be a much less important determinant of the benefits of political advertising than it is a determinant of an individuals instrumental benefits of voting. The number of advertisers is the more relevant determinant. Advertising can have a large effect in an election with many voters if there were, say, only two candidates and one advertiser for each the advertising need only change the votes of a few of the marginal voters.[10]

1.5 Equilibrium Implications of Rational Voting

Instrumental and rational voting have different implications for how people vote, and the actions taken by groups to affect votes.

[9] A similar point is made by (Aldrich 1993, pp. 267–268), although he does not call it "advertising."

[10] With two candidates and more than two advertisers, each advertiser much take into account not only the reaction of advertisers of the opposing candidate, but advertisers of the same persuasion who might free ride.

1.5.1 Groups Act in Group Interest

The typical interest group is only a small fraction of the electorate so, by the same argument, can we conclude that groups do not try to sway votes toward their preferred outcome? Perhaps, as compared to individuals, groups are even less likely to act in their members interest because of free-riding within the group (Olson (1971) makes this argument), but we believe that rational voting implies that group-sponsored advertising will dominate individual self-interest as a determinant of public decisions. First of all, it is easier for a group to combat free riding of political contributions as opposed to votes. Monetary contributions are easier than votes to monitor (especially when voting is by secret ballot). Monetary contributions can also be unequally distributed among members so that group decisions weight members intensity of preference. Second, political advertising may be one means by which a group coordinates the votes of its members and helps to alleviate the free rider problem. In other words, political advertising serves the dual purpose of swaying the votes of nonmembers and to encourage members to spend resources and their votes in the groups interest. And, as we derive below from our model, groups are in a sense more willing to "pay" for votes than are individuals.

Hence, we predict that it is much less likely for political advertising sponsored by a group to go against group interest than it is for an individual to vote against his self-interest. For example, we expect a larger fraction of old age interest groups (such as the American Association of Retired Persons, or Senior Citizens Council) to favor social security increases than the fraction of elderly who would favor such increases. We also expect the successful groups to not only enjoy a relatively high fraction of members voting for policies preferred by the group, but also a relatively high fraction of nonmembers voting in the groups interest too. This is the spirit of the approach taken by interest group models of the type developed by Peltzman (1976), Becker (1983, 1985), Becker and Mulligan (2003), Mulligan and Sala-i Martin (1999), and others. They assume that some voters form groups and that the groups act in the interests of the group.

We suggest that rationality and the negligible effect of an individuals vote on electoral outcomes imply that groups are more likely to act in the groups interest than an individual is to act in his self-interest![11] This is an important difference from market behavior, where as Olson (1971) convincingly argues where each individual communicates his own self-interest.

[11] Sometimes the "wasted vote" argument is explicitly used by advertisers to sway votes to induce individuals to deviate from the "preferred" vote. (Aldrich 1993, p. 270) cites an example from the 1980 Presidential election, "the two parties, their nominees, and interest groups, therefore, make the argument publicly that a vote for a third-party candidate will be wasted. Resources were systematically devoted to convincing people that 'a vote for Anderson is a vote for Reagan,' as Carter put it....".

1.5.2 Political Advertising and Democracy

We do not expect political advertising to be unimportant in non-democracies. By definition, voting is less important in a non-democracy, but the free riding just takes another form namely, resistance or revolution against the political party in power. Just as an individuals vote has a negligible effect on the outcome of an election, an individuals participation in a revolution has a negligible effect on the success of the revolution. We expect political advertising to be just as important in a non-democracy although, while political advertising tries to influence votes in a democracy, we expect political advertising to influence willingness to participate in a revolution.

A corollary to the importance of advertising in public decisions is that image is important in politics. A good politician is as important for political success as a good policy, and this is likely to be the case in both democracies and nondemocracies.

1.6 Conclusions

Approaches that stress competition between political spending are out of step with the most common approaches to voting and political choices. However, an analysis where outcomes are dominated by political spending rather than by intrinsic preferences does seem to be the right way to implement rational voting when political decisions are determined by collective choices.

Of course, whether or not we think instrumental voting is the "right assumption" is hardly relevant. The much more important criteria is whether instrumental voting can predict which policies governments adopt and which they do not, as compared to other theories of public decision-making. We have suggested a number of areas in which our "rational" approach substantially improves upon the predictive power of the instrumental-voter approach, but there is room for a lot more research in this area.

Acknowledgements We appreciate the comments of Morris Fiorina and Erzo Luttmer, the research assistance of Ran Tao and Shiqiang Zhan, University of Chicago seminar participants, and the financial support of the Smith Richardson Foundation. This is an unfinished note we wrote in 1999 and did not publish because we thought it needed, and we would eventually find, an additional breakthrough. Joshua Hall convinced us to include it in his book of unpublished papers. Gary had a few minor edits for Halls purpose, which were lost at the time Gary passed away. Throughout his life, Gary was eager to and capable of further integrating and advancing theories of economic and political behaviors.

References

Aldrich JH (1993) Rational choice and turnout. Am J Polit Sci 37(1):246–278
Becker GS (1983) A theory of competition among pressure groups for political influence. Q J Econ 98(3):371–400
Becker GS (1985) Public policies, pressure groups, and dead weight costs. J Public Econ 28(3): 329–347
Becker GS, Mulligan CB (2003) Deadweight costs and the size of government∗. J Law Econ 46(2):293–340
Benabou R (1996) Inequality and growth. NBER Macroeconomics Annual 1996, vol 11. MIT Press, Cambridge, pp 11–92
Brennan G, Hamlin A (1998) Expressive voting and electoral equilibrium. Public Choice 95(1–2):149–175
Brennan G, Lomasky L (1983) Institutional aspects of merit goods analysis. FinanzArchiv 4: 183–206
Caplan B (2001) Rational ignorance versus rational irrationality. Kyklos 54(1):3–26
Citrin J, Green DP (1990) The self-interest motive in american public opinion. Res Micropolitics 3(1):1–28
Cox GW, Munger MC (1989) Closeness, expenditures, and turnout in the 1982 US house elections. Am Polit Sci Rev 83(01):217–231
Downs A (1957) An economic theory of democracy. Harper, New York
Easterly W, Rebelo S (1993) Fiscal policy and economic growth. J Monetary Econ 32(3):417–458
Jackman RW (1975) Politics and social equality: a comparative analysis. Wiley, New York
Lindert PH (1994) The rise of social spending, 1880–1930. Explor Econ Hist 31(1):1–37
Matsusaka JG, Palda F (1993) The Downsian voter meets the ecological fallacy. Public Choice 77(4):855–878
Meltzer AH, Richard SF (1981) A rational theory of the size of government. J Polit Econ 89(5): 914–927
Mulligan CB, Sala-i Martin X (1999) Gerontocracy, retirement, and social security. NBER Working Paper (w7117)
Myerson RB (1995) Analysis of democratic institutions: structure, conduct and performance. J Econ Perspect 9(1):77–89
Olson M (1971) The logic of collective action. Harvard University Press, Cambridge
Pampel FC, Williamson JB (1989) Age, class, politics, and the welfare state. Cambridge University Press, Cambridge
Peltzman S (1976) Toward a more general theory of regulation. J Law Econ 19(2):211–240
Peltzman S (1980) The growth of government. J Law Econ 23(2):209–287
Persson T, Tabellini G (1999) The size and scope of government: comparative politics with rational politicians. Eur Econ Rev 43(4):699–735
Schumpeter JA (1942) Socialism, capitalism and democracy. Harper and Brothers, New York
Sears DO, Funk CL (1991) The role of self-interest in social and political attitudes. Adv Exp Soc Psychol 24(1):1–91
Sears DO, Hensler CP, Speer LK (1979) Whites' opposition to busing: self-interest or symbolic politics? Am Polit Sci Rev 73(2):369–384
Wittman DA (1995) The myth of democratic failure: why political institutions are efficient. University of Chicago Press, Chicago

Chapter 2
Public Choice Issues in International Collective Action: Global Warming Regulation

Daniel Houser and Gary D. Libecap

Abstract Although there is a growing literature on scientific estimates and regulatory instruments for international efforts to control greenhouse gas emissions, the underlying political collective action processes have been neglected. We focus on the impact of uncertainty in assessing the benefits and costs of global warming regulation on constituencies and politicians in the bargaining countries. Uncertainty arises due to basic information problems about emissions and their link to global warming, the possible range of temperature changes, and their likely effects across the planet. These information problems also create uncertainty in calculating the net effects of global warming, determining its effective regulation, and assessing compliance by sovereign countries that may be differentially affected. We outline a two-stage analytical framework that describes the positions taken by representatives of negotiating countries and the internal public choice tradeoffs facing politicians when constituents are faced with differential and uncertain effects. We apply the framework to the Montreal Protocol to Control Substances that Damage the Ozone Layer of 1987 for insights in analyzing the Kyoto Protocol of 1997. Additional information will reduce uncertainty over time, and until uncertainty is lowered we conclude that limited regulatory efforts are most likely to generate internal political support within negotiating countries for international collective action.

2.1 Introduction

Concerns about the accumulation of greenhouse gases in the atmosphere and possible effects on global temperatures have led to a series of international initiatives for

D. Houser (✉)
ICES, George Mason University, 4400 University Drive, MSN 1B2,
Fairfax, VA 22030, USA
e-mail: dhouser@gmu.edu

G.D. Libecap
University of California, Santa Barbara, 4420 Donald Bren Hall,
Santa Barbara, CA 93106, USA
e-mail: glibecap@bren.ucsb.edu

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_2

collective action (Houghton et al. 1990; Houghton and Callander 1992; Houghton 1995). These include the United Nations Framework Convention on Climate Change (FCCC) signed at Rio de Janeiro in 1992 where countries pledged to voluntarily reduce carbon emissions to 1990 levels by 2000; a meeting in 1995 in Berlin of the Conference of Parties (COP), created at the Rio conference, to define a structure for further action; and the Kyoto Protocol on Global Warming of December 1997 (Sparber et al. 1998).

Under the Protocol, thirty-eight developed countries are to reduce greenhouse gas (GHG) emissions by approximately 95 percent of 1990 levels by 2008–2012. The United States is to lower its discharges of carbon dioxide (CO2) to 93 percent of 1990 emissions. These actions will not be without costs, although neither the costs, nor the benefits of emission controls are known with much certainty.

Uncertainty arises because of a lack of conclusive information about (a). The human sources and pace of temperature change; (b). The costs and benefits of global warming and their distribution across countries; (c). The costs, benefits and effectiveness of different forms of regulation; and (d). The extent of treaty compliance by sovereign countries.

The costs and benefits of global climate change and its regulation will be spread unevenly both across and within countries. These heterogeneous and uncertain constituent effects create challenges for politicians in fixing positions during international negotiations. The associated public choice bargaining issues have been neglected in the literature and are the focus of this paper.[1] Our analysis reveals why it is in the interest of politicians in developed countries, such as the U.S., to delay action until more information is obtained.[2] In the following section, we outline a two-stage political bargaining framework for international collective action that emphasizes the role of uncertainty in reducing expected benefits for constituents and politicians. Implications are developed for analyzing the Montreal Protocol on Substances that Deplete the Ozone Layer and the Kyoto Protocol to the United Nations Framework Convention on Climate Change.

[1] A large and growing literature on regulatory instruments and some of the scientific and economic issues involved has emerged, including Hoel (1997), Hollick and Cooper (1997), Houghton (2009), Moore (1998), Paterson (1996), Shogren and Toman (2000), Weyant (1993), and Wiener (1999a, b) examines some of the constituency issues of concern here.

[2] In the spring of 2001, the U.S. and Australia, two countries likely to bear the greatest share of treaty costs, chose to delay action on global warming. Although subject to international criticism, our analysis suggests why these actions were reasonable for domestic politicians.

2.2 Theoretical Framework for Analyzing International Collective Action to Address Open-Access Resource Problems

In this section we develop a simple model that frames the international negotiation of environmental treaties as a collective-action problem. Consistent with the empirical treatment of the two treaties studied in the paper, the model emphasizes the important public choice issues arising during both internal and international negotiation. International negotiations do not take place detached from the underlying political realities within each of the bargaining countries. Rather, a country is represented by an agent who is accountable to domestic constituencies. This agent must adopt an international negotiating position that simultaneously leads to a resolution of the international common-property problem *and* generates the greatest political support among the electorate.[3]

The collective-action literature makes clear the importance of mitigating differences between bargaining parties and inducing important economic actors to join.[4] In our theory, side payments can be used to accomplish both of these ends. The theory incorporates transfer payments because they arise in the empirical section of the paper. In particular, side payments are frequently used to mitigate perceived differences across countries in treaty net benefits. Agreement on side payments also is made more difficult by uncertainty in benefit and cost estimation.

We model the international collective-action problem in two parts: first, we address the international negotiation task, and second, we examine how underlying public choice concerns give way to a single negotiating position in international negotiations. The framework is consistent with and is motivated by Olson's (1971) rational choice basis for collective action.

Consider first international negotiations. Suppose that a group of I countries is negotiating to resolve an open-access resource problem. The total utility that country i derives from treaty agreement is the sum of the expected net benefits of the treaty, θ_i, whose value is determined in the way described below, and any transfer payments t_i that it receives from or provides to other countries as a condition of agreement. Hence, each country *i's* preferences can be expressed in terms of whether there is

[3] See the readings in Keohane et al. (1996) and Putnam (1988). The impact of uncertainty in international negotiations is examined also by Helm (1998). In our model, the benefits and costs of an international treaty are borne not only by the underlying constituencies, but also by the elected agent. Hence, the agent will be very cautious before committing his/her country to an agreement because imposing even minor costs on a constituency, or constituencies, without commensurate benefits may lead to large defections in political support. The political problem faced by politicians is exacerbated if there is considerable uncertainty in estimating constituent benefits and costs from international action.

[4] Consider the problem of forming a cartel where there are large cost differences. It is difficult to get low cost producers to join if their share of excess profits do not reflect the cost differences. Also, having the largest producers under agreement is critical to the success of any production cartel.

agreement, the expected net benefit of agreement and the transfer payment that is conditional on agreement as follows.

$$u_i(agreement, \theta_i, t_i) = \theta_i + t_i,$$
$$u_i(noagreement, \theta_i, t_i) = 0 \tag{2.1}$$

Given individually rational representative agents, no country will participate in the international agreement unless the sum of expected benefits of the treaty, including transfer payments, is greater than its value of non-agreement.[5]

Realizing that success in negotiations may depend upon the viability of transfer payments, we employ the well-known AGV mechanism suggested by d'Aspremont and Gérard-Varet (1979) and Arrow (1979). The mechanism provides a convenient transfer payment formula that guarantees "budget balance," or that any monies transferred to one country are paid in full by other countries.[6] More importantly, the approach illustrates the importance of transfer payments in international negotiations, and how small changes in the transfer scheme, such as defection by one of the international players, or alternatively, a reduction in uncertainty following new information, can have large consequences for the overall agreement.

Initially, countries' agents do not know each other's expected net benefits, θ_i. Because of uncertainty regarding constituent benefits and costs, agents may be unsure of the nature of domestic support for the treaty, and hence their own θ_i. Camps in support and in opposition to international treaties may solidify only after actual negotiations begin. We assume, however, that the distribution of each θ_i, $(i = 1, \ldots, I)$, is common knowledge.[7]

Each negotiator aggregates constituency preferences, according to a procedure described below, to determine the expected net benefit to report during the international negotiations. Let θ_i denote the expected net benefit reported by the agent for country i. The AGV mechanism's decision rule is to implement the treaty if and only if the aggregated reported expected net benefit is positive, or

[5]Note that we have assumed without loss of generality that each country's expected net benefit is reported in relation to a non-agreement value of zero.

[6]Essentially, the AGV transfer payments compensate "losers" by paying them the expected net benefit of all the others conditional on their own report. Although there is a strong sense in which the AGV mechanism is incentive compatible, to avoid technical issues that distract from the focus of this paper, we assume that negotiators truthfully report to the negotiations an expected net benefit that derives from a vote maximization calculation described below. This assumption is consistent with our view that the agent acts in his/her own political interest because the agents future political support depends on how faithful the constituency believes the agent is in representing their views. It is likely that, where the benefits of an agreement accrue over a long period of time, as is typically the case in environmental treaties, it may be very difficult for the agent to persuade a constituency that acting in violation of their stated views is, in fact, in their best interest.

[7]A regularity condition required for the AGV mechanism is that each country's expected net benefit distribution is common knowledge (d'Aspremont and Gérard-Varet 1979, p.38).

$$\sum_{i=1}^{I} > 0. \tag{2.2}$$

The transfer payments implied by the AGV mechanism if the treaty is implemented are given by d'Aspremont and Gérard-Varet (1979).

$$t_i(\theta) = E_{\theta_{-i}}\left[\left(\sum_{j\neq i}\theta_j\right) I\left(\left\{\sum_{j\neq i}\theta_j\right\} + \theta_i > 0\right)\right] + \tau(\theta_{-i}), i = 1\ldots, I \tag{2.3}$$

where $\theta = \{\theta_1, \ldots, \theta_I\}$, $\theta_{-i} = \{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_I\}$, $\tau(\theta_{-i})$ does not depend on θ_i and is chosen to ensure budget balance, and $I(\dot{})$ is and indicator function that takes value one if $\sum_{j\neq i}\theta_j + \theta_i > 0$ and is zero otherwise.

The first term on the right hand side of Eq. 2.3 is the sum of other countries' expected surplus conditional on the reported net benefit of country i. The second term in Eq. 2.3 is a function whose value is independent of country i's report and which is chosen to ensure budget balance, or that

$$\sum_{i=1,I}(\theta_i) = 0. \tag{2.4}$$

Before moving on, it is worth noting that the model thus far yields several well-known results from the collective-action literature. First, the reported expected net benefits $\{\theta_i\}_{i=1,I}$ in conjunction with Eq. 2.2 will determine the initial feasibility of successful collective action. If the $\{\theta_i\}_{i=1,I}$ are highly heterogenous and include both positive and negative values, then implementing the international treaty requires transfer payments.[8] Furthermore, if parties view the status quo as more attractive than the expected utility of agreement (net of transfers), a collective agreement is not likely. Indeed, where the agreement value is close to the status quo, a transfer payment may still be needed to move a party from its initial position.

It would be possible to proceed relying solely on the above framework, but such an approach would ignore important public choice issues that are central to our analysis, and which we believe will have a sizeable impact on any attempt to negotiate an international global warming treaty. For instance, the empirical analysis described later in the paper suggests that successful negotiations must account for the potentially highly heterogeneous preferences of underlying constituencies within a country. The way the country's agent fixes the negotiation position depends on these preferences,

[8]Heterogeneous in this context refers to the difference across countries of net benefits versus the status quo. For example, if half of the countries expect positive gains from agreement and the other half expects losses, agreement is far less likely (without transfer payments) than when all countries receive positive benefits from agreement. Instances where side payments have been used successfully are often characterized by differences that are calculable. For a different view on the motivation for countries to take action, see Gruber (2000).

and as political economy research has made clear, agents must be responsive to important interest groups in order to secure and maintain political office.

Suppose that there are N_i constituencies to which the agent for country i must respond, and denote the expected net benefit of the treaty to group n_i by B_{n_i}. For now we take B_{n_i} as given. In the sequel we discuss the way in which its value is determined. Our model captures the simple idea that a constituent's expected net treaty benefits affects the negotiation position they prefer their agent to adopt in the international treaty negotiations. Let $V - n - i(B_{n_i})$ denote this preferred position. That is, constituency n_i with expected benefits B_{n_i} prefers that the agent report $\theta_i = V_{n_i}(B_{n_i})$ at the treaty negotiations. It is natural to assume that $V - n_i$ is increasing. We assume that the constituency wants neither a report that is too far above nor too far below its preferred point. Extremely high reports at the international negotiations raise the possibility that agent *n's* country, and potentially constituency *i*, might be asked to fund large transfer payments to other countries (see Eq. 2.3). Lower reports than preferred, on the other hand, leave it relatively too likely that the treaty will not be implemented (see Eq. 2.2).

The agent for country i is interested in maintaining political office and, consequently, in maximizing political support.[9] In the spirit of Peltzman (1976), suppose that the likelihood that any member of group n will vote for agent i is a positive and differentiable function of their preferences, $p_n(V_n(B_n) - \theta_i)$, where p achieves its maximum at zero, is increasing to the left of zero, is decreasing to the right of zero, and takes values in the interval [0, 1]. It follows that the number of votes the agent expects to obtain from group n if there is an international agreement is $m_n p(V_n)$, where m_n is the number of voters in constituency n.

The agent's goal, therefore is to choose an expected net benefit position to report at the international treaty negotiations that solves the following simple vote maximization problem:

$$max_{\theta_i} \sum_{i=1,N} m_n p_n(V_n(B_n) - \theta_i). \tag{2.5}$$

It is worth emphasizing that the agents must choose the expected net benefits, θ_i, to report at the international negotiations based only on their knowledge of the expected values that their constituencies place on the treaty. A feature of many actual treaty negotiations, including those that we discuss below, is that these political calculations must be revisited as more information becomes available to the agent and his/her constituencies. In some cases, particularly in the event of noncompliance, this can lead initially promising treaty negotiations to break down, or alternatively, in the case of important uncertainty in constituent net benefit calculations, new information may make a treaty more politically feasible.

An interior solution to the agent's vote maximization problem Eq. 2.5 is characterized by the following easily derived first-order condition:

[9] To avoid cumbersome repetition of the *i* double-script, we will suppress this notation whenever it is clear from the context that we are discussing features of a country *i's* domestic political setting.

$$max_{\theta_i} \sum_{i=1,N} m_n p'_n(V_n(B_n) - \theta_i) = 0. \tag{2.6}$$

This has the immediate implication that larger (more powerful) constituencies will have a greater influence over the position taken by the agent than will smaller (weaker) constituencies. On the other hand, it also follows that even very large and powerful constituent groups that are highly in favor of a treaty might not have this message perfectly conveyed at international negotiations, particularly if there are several opposition constituencies. The reason is that opposing constituencies always have some voice, and at some point reporting excessively high θ_i could begin to erode the agents political support substantially. This type of tradeoff reflects the underlying public choice concerns that we find in all of the treaties examined in the empirical part of this paper.

Our model is closed by specifying the way in which the agents constituents determine their expected net benefit, B_n. The benefit that any group derives from a particular treaty depends upon that treatys outcome. Imprecise information about the nature of the environmental problem, the effects of a successful treaty, and the likelihood of international compliance, will generate uncertainty over a treatys outcome. This uncertainty will affect the expected net benefits that constituencies report to their political agents.

Let x be a scalar index of a treatys efficacy, with higher values of x indicating a more efficacious result. The advantage of this abstract index is that the treaty negotiations discussed below can each be mapped into this framework. For example, the efficacy of the Montreal Protocol might include some measure of the amount of CFC reduction. From the point of view of any constituent group, before the treaty is signed, x is a random variable whose value has not yet been realized. We assume that each constituency n can associate a net benefit $b_n(x)$ to any treaty outcome x. Moreover, it is natural to assume that $b_n(x)$ is increasing and concave. In the case of CFCs, this implies that the marginal benefit of reduction is non-increasing.

This formulation means that the uncertainty is over the outcome of the treaty, and that the constituents must be able to assign a net benefit to any realized outcome. It follows that the expected net treaty benefit that the constituent reports to its political agent is simply

$$B_n = \int_x b_n(x) dF_n(x) \tag{2.7}$$

where $F_n()\dot{}$ represents group n's subjective beliefs about the likelihood of various treaty outcomes.

This framework provides predictions about the effect of increased uncertainty on constituency n's reported expected net benefit. Following the classic treatment of Rothschild and Stiglitz (1970), define the distribution of a second index y by

$$y = x + \varepsilon \tag{2.8}$$

where $e[\varepsilon|x] = 0$ Hence, the distribution of y is derived by adding noise to the distribution of x. In relation to the distribution of x, the distribution of y provides a natural way to represent a situation where the constituents face more uncertainty about a treatys efficacy, stemming from greater risk of noncompliance, less precise information about the extent and distribution of the environmental problem and hence, the treatys effects. It then follows immediately from the results of Rothschild and Stiglitz (1970) that, as long as b_n is strictly concave, the expected benefit that the constituent will report when it faces y will be less than its report under the situation characterized by x.

Typically, one would expect changes in uncertainty to be roughly the same for all constituents. If all constituents face greater uncertainty and report lower expected net benefits, then the agent will report a lower expected net benefit at the international negotiations (this comparative static is easy to derive from the first order condition Eq. 2.6. Hence, an international agreement becomes less likely. Conversely, reducing the uncertainty surrounding a treatys outcome can improve the chance of its implementation even if this information does not indicate a more efficacious outcome (that is, the expected efficacy is the same under x and y, yet the situation characterized by x receives greater constituent support.)

It is worthwhile to point out that new information about the environmental problem, particularly that which affects the uncertainty surrounding the costs and benefits of addressing it, is effectively a public good. Once available, it allows for more precise determination of all individual constituents benefits and costs of international action, in that it affects their expected net benefits B_n associated with taking particular actions. Moreover, as a practical matter, less uncertainty about outcomes likely makes it easier to arrange transfer payments and other treaty provisions. Viewed in this way, one might expect information about treaty outcomes to be underprovided, and for there to be a role for a central authority to coordinate information accumulation. Pursuing this point rigorously is important, but beyond the scope of this paper.

Important implications of our model, most of which are intuitive, are summarized as follows. Successful collective action is most likely when each country reports an expected net benefit that is large and positive (see Eq. 2.3). In this case the sum of the reported net benefits will be positive, and it is efficient to implement the treaty. A country will provide a large, positive report when all of its constituencies expect the net benefits of the international action to be substantial (see Eq. 2.5). Increased uncertainty within countries over the aggregate gains from collective action, however, can reduce the likelihood of international cooperation. This uncertainty will be reflected in lower reported expected net benefits by each agents domestic constituencies, which in turn leads the agents to report lower expected benefits in international negotiations. Transfer payments from those countries expecting positive net benefits to those expecting zero or negative effects will be necessary to elicit cooperation. Finally, delaying immediate action until more information is available may make subsequent, more extensive international regulation politically feasible. We now apply this framework to the analysis of international environmental problems. Our empir-

ical focus is on constituencies within the U.S., although the approach applies to political agents and constituent groups in other countries as well.

2.3 The Montreal Protocol on Substances that Deplete the Ozone Layer of 1987

The experience of negotiating the Montreal Protocol illustrates how high levels of uncertainty regarding the expected net benefits of action impeded international action initially. As more information about the benefits of controlling emissions appeared, a consensus, particularly among industrialized countries, developed and the Protocol was drafted in September 1987. The Protocol limits the release of gases into the atmosphere that might damage the earths shield against Ultraviolet B (UVB) rays from the sun. Table 2.1 outlines the implications of the model described in Sect. 2 and what is actually observed regarding the progress and nature of international negotiations

Our theory speaks to three important features of this protocol. One is timing. Although the potential consequences of damage to the ozone layer were raised in the early 1970s, the impetus for collective action did not arise until the mid 1980s. The theory explains why the U.S. became the principal proponent, while Britain and France were skeptical. Initially, strong constituencies in those countries were very uncertain about the benefits of such action. A second feature of this protocol is the important role played by transfer payments from developed to developing countries to elicit international cooperation. The final feature is difficulty with compliance.

Table 2.1 Montreal protocol

Framework implications	Empirical observation
Collective action occurs when there are large expected net benefits among constituencies within negotiating countries, i.e., B_{n_i} high within negotiating countries	Collective action takes about 15 years of negotiation. U.S. constituencies eventually see benefits, those in Britain and France do not. Developing countries see few benefits
Uncertainty reduces expected net benefits and thereby reduces collective action. Reduction in uncertainty promotes collective action, i.e., large standard deviation of ε in treaty index y. Reduction in uncertainty promotes collective action, i.e., small ε in treaty index y	Limited information about the problem, alternative technologies, and commercial position limits action by developed and undeveloped countries. New information about ozone holes spurs action
Transfer payments $t_i(\theta_i)$, are necessary to offset differences in expected net benefits	Developing countries still anticipate few benefits, demand fund and technology transfers and lenient treaty provisions
Compliance problems raise uncertainty and reduces collective action	Cheating in transitional and developing economies lowers benefits to adhering constituents

We argue that noncompliance increases uncertainty over the outcome of the treaty, thus reducing developed countries constituents expected net benefits and reducing the likelihood of successful, long-term agreements.

2.3.1 Timing: New Information and the Reduction in Uncertainty

Depletion of the ozone layer emerged as a political concern in the U.S. in the early 1970s in debate over the SST (supersonic transport). While the U.S. had terminated investment in the SST in 1971, British, French, and Russian development of the Concorde and TU 144 continued. The U.S. sought to limit access of SSTs to American airports because of possible negative effects of exhaust emissions in the stratosphere. The Europeans saw this as a ploy to limit use of their planes, and generally dismissed the charges. Important to the continued political debate, however, was new scientific evidence released in the mid 1970s regarding potential damage to the ozone layer from ODS (ozone depleting substance) emissions. This evidence raised the expected benefits of ODS regulation. Studies by both the Department of Transportation and the National Science Foundation in 1974 and 1975 supported concerns about SST exhaust (Morrisette 1989). Additionally, (Molina and Rowland 1974) and Stolarski and Cicerone (1974) described the accumulation of chlorine in the upper atmosphere from CFC accumulation and associated potential deterioration in ozone levels. National Academies of Science (1976a) and National Academies of Science (1976b) outline additional negative environmental effects of CFCs. The EPA and FDA began regulatory proceedings on non-essential use of aerosol sprays under the 1976 Toxic Substance Control Act and the 1977 Clean Air Act Amendments. Public awareness of the issue rose, CFC aerosol sales dropped, and their use banned unilaterally in the U.S. in 1978.[10]

Despite this additional evidence, in the late 1970s and early 1980s there was still no domestic or international political consensus to halt all CFC use. Besides propellants, CFCs were used widely as low-cost refrigerants, solvents, and in the production and cleaning of electronics. The United States accounted for 30 percent of world production of CFCs in 1985, with most made by five firms, DuPont, Allied Signal, Pennwalt, Kaiser, and Racon (Sandler 1997, p. 111). The chemical industry was a politically powerful constituency with a vital interest in regulation, and it initially saw no benefits. The Chemical Manufacturers Association and the lobby group, Alliance for Responsible CFC policy, claimed that CFCs were "incorrectly

[10]See Morrisette (1989) and Litfin (1994). Canada, Sweden, Norway, and Denmark also banned CFCs in aerosols. Noll and Krier (1990) discuss public reaction to low probability catastrophic events.

being blamed for the alleged decreases in atmospheric ozone."[11] The industry argued that more information was needed before taking further action. The atmospheric mechanisms involved with CFCs were incompletely understood, the extent of ozone depletion was unclear, and substantial economic dislocation seemed possible from restricting an industry where the U.S. had a commercial advantage. Consequently, in 1983 the EPA advised Congress that no additional action should be taken until the relationship between CFCs and ozone depletion was better understood (Nangle 1988, p. 543, Hollick and Cooper 1997, p. 157).

By the mid 1980s, however, two events spurred international regulatory efforts. A key factor was the 1985 report of a British Antarctic Survey indicating a 40 percent drop in atmospheric ozone from 1964 levels during the period 1977 to 1984. Although not directly linked to CFC accumulations, the Antarctic ozone "hole" seemed vital new evidence about the problem. This information further increased the expected net benefits of international action among constituencies in developed countries. Second, domestic political opposition in the U.S. diminished with development of low-cost alternatives to CFCs. DuPont announced the company would no longer make CFCs, and lobbied Congress for international restrictions on CFC production and use. An important goal was to help ensure that the CFC-substitute customers, who would bear substantial costs in retrofitting to accommodate the new technology, could not shift to alternative foreign sources of CFCs (Wiener 1999b, p. 772). European firms, particularly British and French companies, however, remained more wary of regulation. They had increased their share of the CFC market and had not invested as much in substitutes (Scott et al. 1995). Agents from European governments initially took more cautious positions in international negotiations for CFC regulation.

The first international action was the 1985 Vienna Convention for the Protection of the Ozone Layer. With estimated benefits in the U.S. of controlling ODS emissions of over $3 trillion at a cost of around $21 billion (Barrett 1994), the U.S. was the major proponent, and it ratified the convention in August of 1986.[12] The convention established broad international objectives, but disagreements among agents of participating countries blocked any substantive CFC control measures. European and American negotiators could not agree on the extent of regulation, and representatives of developing countries did not see elimination of CFCs as beneficial. For constituents in developing countries, the net benefits of international action were assessed as either near zero or negative. CFCs were attractive to developing countries because they were refrigerants that did not require sophisticated technology. Additionally, developing countries objected to trade restrictions and other costs associated with banning CFC imports and exports. In response, developed countries sought to reduce treaty costs (increase side payments) to developing countries. The Vienna Conference established a special category for developing countries that had less than

[11] Comments by Elwood P. Blanchard, Group Vice President for Chemicals and Pigments, DuPont, and by the Chemical Manufacturers Association, May 1987 before the Senate Subcommittee on Stratospheric Ozone Depletion to the Committee on Environmental and Public Works.

[12] Vienna Convention for the Protection of the Ozone Layer, May 2, 1985, Treaty Doc. No. 9, 99th Congress 1st Session, 1985.

0.3 kg per capita consumption of CFCs. Initial international control efforts were to focus on developed countries with consumption levels above the threshold.

The divide among agents of developed countries was closed with additional information on CFC levels and the ozone layer coming in 1986 and 1987 (World Meteorological Organization 1986; Watson et al. 1986; Environmental Protection Agency 1987). At the same time both U.S. and European firms improved their technologies for substitutes. A second round of international negotiations led to the Montreal Protocol of 1987. The protocol defined more precise measures to reduce consumption and production of CFCs and related substances.[13] In Montreal, agents from developed countries were treaty advocates. Under the agreement, developed countries were to cut production and consumption of CFCs by 20 percent of 1986 levels by 1993 and by 50 percent by 1998. CFC trade with countries not adopting the restrictions was to be stopped. Developing countries, however, still required side payments as a condition for participation. Under the notion of "common but differentiated responsibilities," they were allowed an extra 10-years delay to reach reduced production targets and were authorized to exceed their 1986 levels of production by up to 10 percent to satisfy "basic domestic needs."

2.3.2 International Transfers to Induce Participation

Even with these concessions, there was a split between the two groups of countries. 22 of the 50 countries that participated in the Montreal Protocol were developed, and of those, 19 (86 %) signed the agreement. By contrast, of the 19 developing countries that participated, only 6 (34 %) signed (Ling 1992). Two years later, in 1989, just 14 of the worlds developing countries had ratified the Montreal Protocol, whereas most developed countries had. Further, China and India stated they would not participate in the agreement unless more technical and financial aid was forthcoming.

A Second Meeting of the Parties to the Montreal Protocol was held in June 1990 in London to devise additional side payments and to add other chemicals to the control list that had been found to be damaging to the ozone layer. With reduction in uncertainty about the effects of CFCs and related compounds on the ozone layer, developed countries consented to bear more of the costs of regulation. They agreed to end production and consumption of CFCs earlier, by the year 2000. In exchange, developing countries were to stop exporting CFCs to non-participating countries by 1993 (Nangle 1988, 531). A Multilateral Fund was established to provide developing countries with financial and technical assistance. The World Bank became the implementing agency in 1991, and by December 1998 had disbursed $156.2 million with additional commitments of $336.08 million.[14] Administration of the Fund was

[13] Montreal Protocol on Substances that Deplete the Ozone Layer Treaty Doc No. 10, 100th Congress, 1st Session, 1987.

[14] Data from World Bank. The United National Environmental Ozone Secretariat reports a larger disbursement of fund, $768 million to phase out CFCs.

criticized for a lack of accountability, and chemical companies in developed countries were reluctant to relinquish control over substitute technology (DeSombre and Kauffman 1996).

Developing countries still faced major uncertainties with respect to the substitution process (HCFCs, initial substitutes, were found also to be damaging to the ozone layer), costs of compliance, and the extent to which their incremental costs would be covered by the Fund. Hence, agents of developing countries like China and India were not been enthusiastic for the treaty. Only very general language stating that the parties must take "every practicable step" to control CFC emissions could be agreed to by all parties. Many developing countries still had not ratified the Amendments to the Montreal Protocol that placed new chemicals under control (United Nations, Ozone Secretariat 1998). Under the delays granted developing countries, their CFC production continued to rise through 1994 and Halon (another ODS) output increased, so that ozone layer depletion rose at least through 2000, despite strict controls in developed countries (United Nations, Ozone Secretariat 1998).

2.3.3 Enforcement and Uncertainty Treaty Benefits

In addition to resistance by key developing countries to proposed international controls on CFCs, enforcement became an issue in the 1990s. Enforcement was not a critical problem initially since the U.S. was responsible for a large share of total CFC production and the EPA could monitor U.S. firms compliance relatively easily. Throughout the 1990s, however, there were reports of rising production of CFCs in developing countries, the Russian Federation and other transitional economies. There was also evidence that CFCs were being smuggled into regulated areas (Benedictk 1998; Dorfman 1997; Sandler 1997). Cheating resulted in more chlorine in the atmosphere, and complying firms found that their substitute products were facing new competition, while being restricted in order to continue to reduce chlorine levels. Migration of CFC-intensive industries to less regulated countries also reduced the benefits of the agreement within countries that adhered (Chemical Marketing Reporter 1996). Representatives of DuPont complained before the Congressional Subcommittee on Stratospheric Ozone Depletion that developing countries were relying too much on CFCs rather than on substitutes. Company officials testified "at least six CFC plants have started up or are under construction in less-developed countries since the Montreal Protocol was available for ratification."[15]

Systematic cheating raises a significant possibility that the political coalitions that supported the Montreal Protocol could unravel. At a minimum, the presence of cheating increases uncertainty over the net benefits of cooperation. Moreover, a single non- cooperative country raises the uncertainty for all cooperators. The result,

[15]Testimony by Dwight Bedsole, business manager, DuPont Freon Products Division, U.S. House of Representatives Committee on Energy and Commerce, subcommittee on Stratospheric Ozone Depletion, January 25, 1990, CIS 90H361-38, 271–73.

as we showed in Sect. 2.2, can be a reduction in the aggregate expected net treaty benefits and, therefore, an increased likelihood that the treaty will end in failure. A natural way to prevent such cooperative decay is to increase monitoring and transfers to developing countries. Whether developed countries constituencies perceive the benefits of the Montreal Protocol as large enough to offset higher monitoring and transfer costs is an open question. Additionally, since international environmental treaties are voluntary, forcing the compliance of constituencies in sovereign countries may not be possible at any reasonable cost.[16]

2.4 The Kyoto Protocol to the United Nations Framework Convention on Climate Change of 1997

Like with the Montreal Protocol, the experience of the Kyoto Protocol illustrates how uncertainty in the expected net benefits of emission controls across countries has limited international action. As with the Montreal Protocol, it may be that as more information appears, an international consensus on regulation eventually will develop. This point underscores the importance of information generation as a public good in international collective action. Table 2.2 outlines the implications of the theoretical framework for the progress and content of international negotiations to control emission of greenhouse gases (GHG) and empirical observations up to this point.

The theory explains why uncertain constituent net benefits within developed and undeveloped countries that are major carbon emitters makes global warming regulation so politically controversial. It also suggests why delay in adopting significant international commitments is a reasonable position for political agents in those countries. Important differences in the anticipated benefits and costs of GHG regulation across countries means that significant transfer payments to build support for international actions are required. The theory indicates, however, that these will be politically difficult to design because of the uncertain constituent net benefits of regulation. Finally, because of the high costs of regulation in some countries treaty compliance is an issue. As with the Montreal Protocol, cheating adds more uncertainty for all countries in calculating the net benefits of international action.

[16]Chang (1995) presents a case for the use of trade sanctions by countries in support of international environmental treaties. In narrowly focused agreements trade sanctions might serve as an enforcement mechanism. In broad treaties, like the Kyoto Protocol where the range of industries and countries is much larger, trade sanctions are less likely to be effective. Barrett (1994) argues that the Montreal Protocol accomplished relatively little over a non-cooperative outcome.

Table 2.2 Kyoto protocol

Framework implications	Empirical observation
Collective action occurs when there are large expected net benefits among constituencies within negotiating countries, i.e., B_{n_i} high within negotiating countries	Collective action remains controversial. Key U.S. constituencies see few benefits, those in Britain and France anticipate gains. Constituencies in most developing countries do not
Uncertainty reduces expected net benefits and thereby reduces collective action. Reduction in uncertainty promotes collective action, i.e., large standard deviation of ε in treaty index y. Reduction in uncertainty promotes collective action, i.e., small ε in treaty index y	Limited information about the problem, costs, and regulatory approach limits action by developed and undeveloped countries. New information might promote action
Transfer payments $t_i(\theta_i)$, are necessary to offset differences in expected net benefits	Developing countries still anticipate few benefits, demand fund and technology transfers and lenient treaty provisions
Compliance problems raise uncertainty and reduces collective action	Lack of enforcement raises concern about benefits of collective action

2.4.1 Uncertainty and Calculation of Net Benefits from Political Action

International collective action to regulate general GHGs is much more complex than is effort to control CFCs and related chemicals. The number of gases involved, the constituencies affected, and the range of economic costs and benefits are far much larger, and the extent of uncertainty is greater. There are numerous sources of uncertainty that affect assessment of the benefits and costs of GHG abatement, hence affecting political stands by agents in negotiations. Despite the availability of new information about higher temperatures, the magnitude of global warming remains undetermined. The rate at which greenhouse gas concentrations will increase and the relationship between accumulation of GHG in the atmosphere and the extent of warming is not quantified. Offsetting effects of other factors, such as the oceans and forests are unknown. The human role is disputed, and the reaction of the oceans and ice caps to higher temperatures is difficult to gauge.[17] The global change models that are used to simulate possibilities are particularly imprecise about regional effects, masking regional variation. These regional patterns, however, are crucial for motivating country participation and adherence to international efforts.

There are substantial heterogeneities within and among countries in the anticipated effects of global warming and in the costs of reducing greenhouse gas emissions

[17] For disputes of GHG effects, see the testimony of Patrick J. Michaels of the University of Virginia before the U.S. House Committee on Small Business, July 29, 1998. For summary discussion of the many issues and uncertainties involved see Paterson (1996); Hollick and Cooper (1997); Shogren and Toman (2000). (Houghton 2009, pp. 1–8) seems more confident in the consistency of the patterns.

(Holtz-Eakin and Selden 1995). Current global circulation models (GCMs) indicate that some areas might benefit from moderate global warming, others might be moderately affected, and some might be seriously harmed. Those countries vulnerable to a rise in sea levels seem to have most at stake, including small island states, Bangladesh and the Netherlands. China, Russia, other Northern European countries, and Canada might benefit through increased agricultural production. Studies indicate a possible increase in agricultural production in the U.S. (Kane et al. 1992; Mendelsohn et al. 1994; Mendelsohn and Neumann 1999). In the tropics, predictions are less clear, but there may be little change.

Estimates of the costs of GHG abatement vary widely, according to assumptions used about the speed and amount of reduction, adjustment flexibility, advent of new energy technology, and the whether emission-permit trading or other market mechanisms are allowed. To meet Kyoto objectives, the U.S. will have to reduce emissions by 30 % of their 1990 level by 2020, a large amount in a short time.[18] U.S. GDP losses by 2010 range from 1 to 4.2 %.[19] These differences add uncertainty for constituents and politicians in calculating the net benefits of global warming regulation. If regulations were gradually put into place, world GDP growth over the next 50 years would decline from a projected 2.3 to 2.25 % annually. More abruptly implemented controls, however, could cost 2.5 % of world GDP by 2043 or $2.25 trillion (Burniaux et al. 1992; Weyant 1993).

The costs will be the greatest for the countries that produce the most CO2. In 1996 the U.S. and Canada, both of which rely on coal-based electricity production, accounted for about 25 percent of global CO2. Because of subsidized reliance on coal, China is projected to be the largest producer of CO2 by 2015 and India the second largest (Burniaux et al. 1992; Poterba 1993). Among countries, Canada, U.S., Italy, Japan, France, and Australia will bear the greatest costs. Among U.S. states, Alaska, Montana, New Jersey, Florida, Texas Louisiana, and Wyoming will be the hardest hit (WEFA 1998). And among industries, coal and energy-intensive sectors, such as steel, aluminum, paper, chemicals, and transport will incur the greatest costs. More broadly, consumer prices will rise with higher energy costs.

Differences in anticipated net benefits of regulation create conflicting stands among political constituencies. Proponents of regulation are environmental groups and firms like BP and Royal/Dutch Shell with large holdings of natural gas and investment in alternative energy sources, such as solar and wind.[20] Within the U.S. and other countries, such as Australia, another large and powerful group of constituents

[18] Weyant (1993), Hollick and Cooper (1997), National Academies of Science (1992), and Manne and Richels (1990) estimate that reducing CO2 emissions by 20 % would cost the U.S. between $800 billion and $3 trillion between 1990–2010, or about 5 % of total macroeconomic consumption.

[19] The U.S. Energy Information Agency compared the cost estimates provided by WEFA, Charles River Associates, Pacific Northwest National Laboratory, MIT, the Electric Power Research Institute, and DRI, Inc. See also Kirova (1999).

[20] For the position of some groups, see US House of Representatives (1998). BP and Shell have large holdings of natural gas that would be in greater demand with restrictions on other fossil fuels. BP also has invested in alternative energy sources and the value of the investment would rise with restrictions on carbon use (Murphy 2002).

that anticipate harm have mobilized to oppose GHG regulation. For example, during 1998 Congressional hearings, representatives of the American Petroleum Institute and the American Council for Capital Formation presented treaty cost estimates that were much higher than those presented by the Clinton Administration

(U.S. House, 1998d, 53–78). Also, powerful small business and farm groups have voiced concerns about substantially higher energy costs (U.S. House, 1998a, 4–37; 1998b, 3–40).[21] As outlined in Sect. 2.2 above, these heterogeneous, uncertain net benefits make it difficult for political agents to support rapid, major cuts in GHG emissions or transfers to other countries to facilitate international action.

2.4.2 Transfer Payment Demands

With existing information, agents of developing countries have taken the position that a global warming treaty would provide few benefits, but have high domestic economic costs. China and India have claimed that they would not participate in international emissions reductions unless there were large compensating transfers from developed countries (Agarwal and Narain 1991; Hollick and Cooper 1997). In response, agents from developed countries initially proposed more lenient treaty provisions for developing countries similar to those granted in the Montreal Protocol. In the 1992 United Nations Framework Convention on Climate Change (FCCC), only Annex 1 countries, which included developed countries in the OECD and transitional economies in Eastern Europe and the former Soviet Union, were called upon to voluntarily reduce GHG emissions. Developing countries were exempted from taking direct action: the FCCC explicitly recognized that these countries had common, but differentiated responsibilities.

Given the continued buildup of gases that damage the ozone layer due to developing country exemptions to the Montreal Protocol, there was similar concern that GHG abatement goals could not be met. The issue was raised during the Conference of Parties of the FCCC in Berlin in April 1995, but no new commitments were requested of developing countries. In this spirit, the 1997 Kyoto Protocol to the Framework Convention adopted binding emission reduction targets only for Annex B or industrialized countries.[22] Even within that group, differences were allowed. Total CO2 emissions were to be reduced by 5.2 percent of 1990 levels. European Union countries were allowed to follow an inclusive or bubble reduction target of eight percent and the U.S. a seven percent reduction by the period 2008–2012. Transition economies were allowed to use a different base year, and developing countries were not required to take any action. Abatement exemptions, however, add more

[21] For example fear about possible job losses in the US led the Byrd-Hagel Resolution to pass 95 to 0 in July 1997 that insisted that developing countries participate in any global warming effort, 143 Congressional Record S8113-05, daily edition, July 25, 1997.

[22] United Nations Framework Convention on Climate Change S. Treaty Doc No. 102-38, 31 ILM 849. Kyoto Protocol to the FCCC, FCCC Conference of the Parties, 3d Sess, UN Doc.

uncertainty to estimations of the effects of any global warming treaty. With current rates of economic development and fossil fuel use in developing countries, net emissions by participating countries would have to become negative by the middle of the 21st century in order to lower GHG accumulations. The associated higher costs could stimulate political opposition to international cooperation.

Accordingly, the magnitude of international financial and technical transfers to secure participation by developing countries is far larger than in the past (Jacoby et al. 1998, p. 90). Such transfers are implicit in the Clean Development Mechanism (CDM) and Joint Implementation projects outlined in the Kyoto Protocol, whereby developed countries obtain emission abatement credits for investing in carbon reduction in developing countries. Neither mechanism, however, has been defined. They are tied up with international debate over the nature and extent of emission rights trading. The much larger amounts that would be transferred from developed countries to any new international GHG fund, similar to the Montreal Protocols Multilateral Fund, remain unresolved.

2.4.3 Compliance and Enforcement

There is no underlying enforcement mechanism within the Kyoto Protocol. No consequences of noncompliance could be agreed upon, and the compliance provisions that are included apply only to Annex 1 or industrialized countries (Breidenich et al. 1998). Monitoring depends on annual self-reports by countries using ostensibly comparable methodologies. Absent effective enforcement, there will be incentives for countries to defect whenever the internal political costs of regulation become too high.[23] While there are growing enforcement problems with the Montreal Protocol, compliance will be a much greater issue with GHG regulation. Controls must be more sweeping across sectors, involving higher economic costs. With heterogeneous, uncertain net benefits of regulation, there will be differential incentives to comply even with transfers. But widespread cheating will only add to uncertainty regarding the returns to international cooperation.

2.5 Conclusion

Theory and research regarding collective action addressing local open-access resource problems indicates that success in controlling externalities comes when: (a). There is a consensus on the aggregate benefits to be gained, (b). The parties perceive positive net gains from agreement, and (c). They are homogeneous with respect to bargaining

[23] See Chang (1995) for use of trade sanctions as a means of enforcement. Few international environmental agreements contain substantive commitments (US General Accounting Office 1999). Bac (1996) discusses free riding.

objectives and in the distribution of the costs and benefits to be incurred. Agreements reached under these conditions tend to be self-enforcing because it is in the interest of all parties to insure success (Ostrom 1990). Collective action may also achieve its objectives if the parties are heterogeneous with respect to the net gains from cooperation if: (a). The spread is not too great, (b). There is little uncertainty as to the consequences of agreement, and (c). There are bases for constructing side payments to compensate those parties that may bear more costs or receive fewer gains. The resulting property rights structure must be secure so that the side payments are long term and predictable. These conditions require an enforcement arrangement that is binding for all parties. Negotiating international agreements for collective action regarding the control of environmental externalities confronts the same requirements for success. But the challenges are much more formidable.

The experience of the Montreal Protocol to Control Substances that Damage the Ozone Layer is insightful for understanding the issues raised by the Kyoto Protocol on Global Warming. The magnitudes of the problems faced by political constituencies in assessing net benefits and by politicians in assembling domestic coalitions for international action are much larger. As suggested by our framework, as new information emerges and uncertainty is reduced, political agreement may be forthcoming. In the meantime, strict GHG regulations are unlikely to gather much political support in countries that anticipate high costs of regulation. Moderate R&D and information development objectives also avoid premature adoption of long-term, irreversible abatement technologies where the opportunity costs exceed those of GHG stock irreversibilities (Kolstad 1996; Pindyck 2000).

Acknowledgements We have benefited from comments by Doug Allen, Robert Fischman, Ron Johnson, Thomas Lyon, John McGinnis, Gordon Tullock, members of the Law and Economics Workshop, University of Pennsylvania, and participants at the Conference of the International Association for the Study of Common Property, Bloomington, June 2000 and at the Western Economics Association Conference, Vancouver, June 2000. The authors also gratefully acknowledge the support of the International Center for Economic Research (ICER), Turin, Italy.

References

Agarwal A, Narain S (1991) Global warming in an unequal world: a case of environmental colonialism. Centre for Science and Environment, New Delhi, India

Arrow K (1979) The property rights doctrine and demand revelation under incomplete information. In: Boskin M (ed) Economics and human welfare. Academic Press, New York, pp 23–40

Bac M (1996) Incomplete information and incentives to free ride on international environmental resources. J Env Econ Manag 30(3):301–315

Barrett S (1994) Self-enforcing international environmental agreements. Oxford Econ Pap 46:878–894

Benedictk RE (1998) Ozone diplomacy: new directions in safeguarding the planet. Harvard University Press, Cambridge

Breidenich C, Magraw D, Rowley A, Rubin JW (1998) The kyoto protocol to the united nations framework convention on climate change. Am J Int Law 92(2):315–331

Burniaux JM, Martin JP, Nicoletti G, Martins JO (1992) The costs of reducing CO_2 emissions. OECD, Paris, France
Chang HF (1995) An economic analysis of trade measures to protect the global environment. Georgetown Law J 83(6):2131–2214
Chemical Marketing Reporter (1996) Europeans calling for CFC trade ban. Chem Mark Rep 250(13):9
d'Aspremont C, Gérard-Varet LA (1979) Incentives and incomplete information. J Pub Econ 11(1):25–45
DeSombre ER, Kauffman J (1996) The montreal protocol multilateral fund: Partial success story. Institutions for Environmental Aid: Pitfalls and Promise, MIT Press, Cambridge (USA) and London pp 89–126
Dorfman R (1997) Protecting the transnational commons. In: Dasgupta P, Mäler KG, Vercelli A (eds) The economics of transnational commons. Clarendon Press, Oxford, UK, pp 210–219
Environmental Protection Agency (1987) Assessing the risk of trace gases that can modify the stratosphere. Environmental Protection Agency, Washington
Gruber L (2000) Ruling the world: power politics and the rise of supranational institutions. Princeton University Press, Princeton
Helm C (1998) International cooperation behind the veil of uncertainty: the case of transboundary acidification. Env Res Econ 12(2):185–201
Hoel M (1997) How should international greenhouse gas agreements be designed? In: Dasgupta P, Mäler KG, Vercelli A (eds) The economics of transnational commons. Clarendon Press, Oxford, UK, pp 172–191
Hollick AL, Cooper RN (1997) Global commons: can they be managed? In: Dasgupta P, Mäler KG, Vercelli A (eds) The economics of transnational commons. Clarendon Press, Oxford, UK, pp 141–171
Holtz-Eakin D, Selden TM (1995) Stoking the fires? CO_2 emissions and economic growth. J Pub Econ 57(1):85–101
Houghton J (2009) Global warming: the complete briefing. Cambridge University Press, Cambridge, UK
Houghton J, Jenkins G, Ephrams J (1990) Climate Change:The IPCC Scientific Assessment. Cambridge University Press, Cambridge
Houghton JT (1995) Climate change 1995: the science of climate change: contribution of working group i to the second assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, UK
Houghton JT, Callander BA (1992) Climate change 1992: the supplementary report to the ipcc scientific assessment. Cambridge University Press, Cambridge, UK
Jacoby HD, Prinn RG, Schmalensee R (1998) Kyoto's unfinished business. Foreign Aff 77(4):54–66
Kane S, Reilly J, Tobey J (1992) An empirical study of the economic effects of climate change on world agriculture. Clim Change 21(1):17–35
Keohane RO, Levy MA et al (1996) Institutions for environmental aid: pitfalls and promise. MIT Press, Cambridge
Kirova MS (1999) Estimating the costs of Kyoto: uncertainities and assumptions driving the model results/c Milka S. Center for the Study of American Business, Washington University, Kirova
Kolstad CD (1996) Learning and stock effects in environmental regulation: the case of greenhouse gas emissions. J Env Econ Manag 31(1):1–18
Ling B (1992) Developing countries and ozone layer protection: Issues, principles and implications. Tulane Env Law J 6:91–126
Litfin K (1994) Ozone discourses: science and politics in global environmental cooperation. Columbia University Press, New York
Manne AS, Richels RG (1990) CO_2 emission limits: an economic cost analysis for the usa. Energy J 11(2):51–74
Mendelsohn R, Neumann JE (1999) The impact of climate change on the United States economy. Cambridge University Press, Cambridge, UK

Mendelsohn R, Nordhaus WD, Shaw D (1994) The impact of global warming on agriculture: a ricardian analysis. Amer Econ Rev 84(4):753–771
Molina MJ, Rowland FS (1974) Stratospheric sink for chlorofluoromethanes: chlorine atom-catalysed destruction of ozone. Nature 249(28):810–812
Moore TG (1998) Climate of fear: why we shouldn't worry about global warming. Cato Institute, Washington
Morrisette PM (1989) Evolution of policy responses to stratospheric ozone depletion, the. Nat Res J 29:793–820
Murphy C (2002) Is bp beyond petroleum? hardly. Fortune 146(6):44–45
Nangle OE (1988) Stratospheric ozone: United states regulation of chlorofluorocarbons. Boston College Env Aff Leg Rev 16:531–580
National Academies of Science (1976a) Halocarbons: effects on chlorofluoromethane release. NAS Press, Washington
National Academies of Science (1976b) Halocarbons: effects on stratospheric ozone. NAS Press, Washington
National Academies of Science (1992) Policy implications of greenhouse warming: mitigation, adaptation, and the science base. NAS Press, Washington
Noll RG, Krier JE (1990) Some implications of cognitive psychology for risk regulation. J Leg Stud 19(2):747–779
Olson M (1971) The logic of collective action. Harvard University Press, Cambridge
Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge
Paterson M (1996) Global warming and global politics. Routledge, New York
Peltzman S (1976) Toward a more general theory of regulation. J Law Econ 19(2):211–240
Pindyck RS (2000) Irreversibilities and the timing of environmental policy. Res Energy Econ 22(3):233–259
Poterba JM (1993) Global warming policy: a public finance perspective. J Econ Perspect 7(4):47–63
Putnam RD (1988) Diplomacy and domestic politics: the logic of two-level games. Int Org 42(03):427–460
Rothschild M, Stiglitz JE (1970) Increasing risk: I. a definition. J Econ Theory 2(3):225–243
Sandler T (1997) Global challenges: an approach to environmental, political, and economic problems. Cambridge University Press, Cambridge, UK
Scott GL, Reynolds GM, Lott AD (1995) Success and failure components of global environmental cooperation: the making of international environmental law. J Int Comp Law 2:23–60
Shogren J, Toman M (2000) Climate change policy. In: Portney PR, Stavins RN (eds) Public policies for environmental protection. Resources for the Future, Washington, pp 125–168
Sparber PG, O'Rourke PE, Landrith GC (1998) Understanding the Kyoto protocol: a comprehensive citizen's guide to the scientific and political issues surrounding the new United Nations treaty and global warming. Legal Center for the Public Interest, Washington
Stolarski RS, Cicerone RJ (1974) Stratospheric chlorine: a possible sink for ozone. Can J Chem 52(8):1610–1615
United Nations, Ozone Secretariat (1998) Control measures under the montreal protocol. United Nations Ozone Secretariat, New York
US General Accounting Office (1999) International environment: literature on the effectiveness of international environmental agreements. US General Accounting Office, Washington
US House of Representatives (1998) Oversight hearing on theKyoto Protocol: The undermining of american prosperity. In: Hearing before the Committee on Small Business, US House of Representatives
Watson RT, Geller M, Stolarski RS, Hampson R (1986) Present state of knowledge of the upper atmosphere: an assessment report. National Aeronautics and Space Administration, Washington
Wefa I (1998) Global warming: the high costs of the Kyoto protool. National and State Impacts, WEFA
Weyant JP (1993) Costs of reducing global carbon emissions. J Econ Perspect 7(4):27–46

Wiener JB (1999a) Global environmental regulation: Instrument choice in legal context. Yale Law J 108:677–800
Wiener JB (1999b) On the political economy of global environmental regulation. Georgetown Law J 87:749–794
World Meteorological Organization (1986) Atmospheric ozone, 1985. World Meteorological Organization, Geneva, Switzerland

Chapter 3
Too Inexpensive to Be Inexpensive: How Government Censorship Increases Costs by Disguising Them

J.R. Clark and Dwight R. Lee

Abstract Politicians often see price ceilings, subsidies and third-party payments as effective ways of reducing the amount consumers pay directly for goods and services and take credit for reducing their costs. While these policies may reduce prices, they are a form of censorship that invariably increases costs. Politically inspired interference in the communication that takes place through market prices reduces the information and discipline required to control costs. The most notable recent example of politicians trying to take credit for reducing costs with policies that increase them is found in their recommendations to reform health care. There are unfortunately a number of other examples such as price controls on apartment rents and subsidies to agriculture and education.

3.1 Introduction

Controlling costs requires conveying information on what costs are and then motivating people to consider them when making decisions. This is not easily accomplished and no economic system performs this task perfectly. Every decision one person makes imposes costs on countless others by using products or resources from which others could have benefited. Ideally, no one will make a decision to use something that is worth less to him than it costs-i.e., is worth to others. While no economic system achieves this ideal perfectly, market economies based on private property and voluntary exchanges are far more effective than any other process at communicating information on costs and benefits in ways that motivate appropriate responses to this information. In markets, this communication takes place through prices.

J.R. Clark (✉)
The University of Tennessee at Chattanooga, 615 McCallie Avenue, Chattanooga, TN 37403, USA
e-mail: j-clark@utc.edu

D.R. Lee
Southern Methodist University, 6212 Bishop Blvd., Dallas, TX 75275, USA
e-mail: leed@cox.smu.edu

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_3

Market prices are constantly adjusting to changing conditions and preferences to reflect the marginal costs of making goods and services available and the marginal value consumers realize from them. These prices inform consumers of the marginal cost of consuming a good, both in terms of additional production and sacrificed consumption by others. In addition, market prices motivate suppliers to expand production of goods as long as their marginal value is worth at least as much as the marginal cost of making them available. Again, no real-world market functions perfectly. When markets fail, or are seen to fail, the common response is that government policies should correct the perceived failure by altering prices with controls, subsidies or taxes. The problem is that government policies do not work perfectly either and seldom improve upon market prices at communicating dispersed information to those best able to use it in ways that keep costs as low as possible.[1]

But, no matter how well markets are controlling costs, consumers would like them to be lower. This creates opportunities for politicians to gain support by promising, if not a free lunch, a cheaper lunch with policies that give the appearance of lowering the costs of a variety of goods and services. Despite appearances, however, these policies almost always increase costs by censoring the price information that is required to keep people informed on, and sensitive to, the real costs of their decisions.[2] By concealing costs with policies that outlaw or distort the price communication that reveal those costs, politicians receive gratitude from people who, while believing they are paying less, are actually paying more.

In this paper, we consider several ways politicians consistently convince people that their costs are being reduced with policies that, by censoring communication through market prices, increase them. Price ceilings, subsidies, and third-party payments (including the problems caused by government policies on private medical insurance that are being used to justify medical-care reforms) will be examined in Sects. 3.2, 3.3, and 3.4 respectively. Concluding comments, with a brief discussion of why past health-care reforms have increased health-care costs, will be offered in Sect. 3.5.

3.2 Price Ceilings

Despite the long lines for gasoline caused by federal price ceilings in the 1970s and 80s, a Gallop Poll conducted during May 2008 found that a majority of Americans

[1] Given that costs are foregone benefits, keeping costs as low as possible is equivalent to increasing benefits as much as possible.

[2] Censorship is not too strong a word for government actions that alter market prices for political purposes. As we have argued in a previous article (Clark and Lee 2008), the information communicated through market prices is every bit as important to our prosperity, liberty and general wellbeing as the information communicated in written and verbal forms that is protected by the first amendment to the United States Constitution.

Fig. 3.1 Price ceilings

favored price controls on gas because of its high cost.[3] The federal government did not impose a price ceiling on gas in 2008. In recent years, however, state governments have imposed such controls for the stated purpose of protecting consumers against the high cost of gasoline after natural disasters and price spikes. Also, a number of municipal governments have rent ceilings on apartments which are justified as necessary to keep down the cost of housing. And in the early 1970s, the federal government imposed price ceilings on literally thousands of goods and services in the name of fighting inflationary increases in the cost of living. In each of these cases, and many more, the effect of a price ceiling is the opposite of what politicians claim and the public seems to believe. As can easily be seen from demand and supply curves, price ceilings increase costs instead of reducing them.

Consider Fig. 3.1 with the demand curve D and supply curve S for a product. Without any restrictions on price communication, the price will reach an equilibrium given by P_E and the amount demanded and supplied will be equal at the equilibrium rate given by Q_E. Assume, however, politicians decide the price P_E is too high and promise to reduce its costs to consumers by imposing a price ceiling given by P_C. As seen at price P_C, consumers want Q_C units of the product, which is where their marginal consumption value equals P_C, and suppliers are willing to supply only Q_S units, which is where their marginal production cost is also P_C.[4]

[3] Fifty three percent wanted the government to impose price controls on gasoline and 45 % were opposed (Jacobe 2008).

[4] The exception to such a shortage being created by a price ceiling below the equilibrium price occurs if the good is being produced by a monopolist and the price ceiling is set at the price where that demand curve intersects the monopolists marginal cost curve. We ignore this possibility here.

Assuming that the price ceiling is strictly enforced, making it impossible for consumers to pay suppliers or for suppliers to receive from consumers a price higher than P_C, the marginal value of the good to consumers will increase to P' (the height of the demand curve at Q_S). Given a marginal value of P', consumers are willing to pay $P' - P_C$ more than is legal to pay with money. So, they will pay in other ways. In the case of price ceilings on gasoline, for example, the most obvious ways to pay more are by waiting in lines, driving around looking for shorter lines, or carrying full gasoline containers in the car trunk. Consumers would pursue some combination of these activities to get another gallon of gas until they are paying approximately $P' - P_C$ per gallon in terms of the opportunity cost of their time, convenience and safety. So their total cost ends up being P' per gallon, which is obviously more than the P_E it would cost without the price ceiling.

Price ceilings are commonly accompanied by non-price rationing schemes enforced by government. But, these schemes also lead to costly adjustments as the choices consumers would make (given their particular preferences and circumstances) are replaced by arbitrary and uniform restrictions that ignore the diversity of preferences and circumstances. In the case of price ceilings on gasoline, governments have rationed it with restrictions such as how much can be purchased at one time, how often it can be purchased (e.g., on odd or even dates, depending on the last number on ones license plates), or coupons. Coupon rationing is potentially the least costly non-price rationing approach. Gas coupons are distributed that allow a specified amount of gas to be purchased at the controlled price of P_C per gallon, with the coupons restricting the total of gas consumed to the amount supplied at the price ceiling-Q_S gallons in Fig. 3.1. Of course, coupon rationing also results in costly inconvenience for almost everyone, with the distribution of coupons having little to do with how much gasoline is worth to different people.[5] In the most efficient variation of coupon rationing, the coupons can be bought and sold at an unregulated price resulting in them being reallocated until all consumers realize the same marginal value from gas, which maximizes the value of the available gas (which is the same as minimizing the cost of using the available gas). Of course, the price of the coupons will be bid up to $P' - P_C$ per gallon. So even with the best government rationing approach, the price ceiling has still increased the cost of gas, or any other good subject to a price ceiling, from P_E to P'.[6]

The increased cost caused by a price ceiling is directly related to the censorship of effective communication between consumers and suppliers. Long lines of inconvenienced and frustrated consumers do communicate to suppliers that consumers want more of the good being supplied. But, those lines do not provide infor-

[5] Some groups will be favored over others, but largely on the basis of the relative political influence of different groups. This creates incentives for groups to lobby political authorities, which is another cost associated with government interference with price communication. See the discussion on rent seeking in Sect. 3.3.

[6] In most cases, it is illegal to buy and sell rationing coupons since the price paid for the coupons makes it obvious that the price (and costs) has increased. Despite the law, however, markets for coupons invariably materialize because people benefit from exchanges by transferring the rationed good from those who value it less (at the margin) to those who value it more.

mation on consumer preferences as clearly and concisely as would market prices, and they provide no motivation for suppliers to respond to the desire for more of the good. Without the price ceiling, market prices would provide suppliers with both the information and motivation needed to expand production to Q_E, where consumers no longer valued another unit of the good by more than the cost of producing it.

Of course, enforcement of price ceilings is never as strict as we have assumed, and despite the penalties on those caught buying and selling at a price greater than the legal price P_C, illegal exchanges take place. These exchanges, by communicating price information from consumers to producers, will lower the cost of the product to consumers below P', but the cost will remain higher than the market price of P_E. This result is shown in Fig. 3.1 with shifts in the demand and supply curves that reflect the expected penalties from buying and selling at an illegal price.

Assume that the expected marginal penalty imposed for violating the price ceiling is a constant amount for each unit bought and sold, and it is half as much for buyers as for sellers.[7] The effect is to shift the demand curve down by the buyers expected marginal penalty (from D to D_P) and to shift the supply curve up by the sellers expected marginal penalty (from S to S_P). The result is an illegal equilibrium at price $P_{E'}$ and quantity $Q_{E'}$. Both buyers and sellers are better off by violating the price ceiling, even with the penalties for doing so, since it results in more output with a marginal value greater than marginal cost (including the penalty cost) being made available. And, the per-unit cost to consumers has declined below P', but that cost is still greater than it would be without a price ceiling. The cost is now equal to P'' in Fig. 3.1, which equals the price paid for the product, $P_{E'}$, plus the expected marginal penalty (given by the vertical distance between the demand curves D and D_P).

The cost to consumers is the same no matter how the expected marginal penalty is split between buyers and sellers. For example, if the penalty were imposed entirely on buyers, the supply curve would remain at S and the demand curve would have shifted down by an amount equal to the entire penalty, shown as $D_{P'}$ in Fig. 3.1. The illegal equilibrium would occur now at price $P_{E''}$ and quantity $Q_{E'}$. The new price is less than consumers were paying without the price controls, P_E, but the cost of the good to consumers is still P'' when the expected marginal penalty is added to the price $P_{E''}$.

A price ceiling that is below the free-market equilibrium price, even if imperfectly enforced, increases the cost consumers pay to acquire the good. The increased cost considered here considers only the cost of what is purchased and ignores the loss to consumers from not being able to buy as much of the good as they would like at prices that would motivate suppliers to make the additional amount available. We now consider government policies that claim to lower the cost of goods to consumers by subsidizing them. Again, the effect is contrary to the political claims, with subsidies

[7] If the same marginal penalty is imposed on sellers and buyers, the *expected* marginal penalty on sellers will typically be higher because they are easier to catch (sellers have to make information available to potential customers that can be intercepted by the police). It will be clear that our conclusion is the same for a given expected marginal penalty no matter how it is split between buyers and sellers.

increasing the cost of the goods to consumers. The difference between subsidies and price ceilings is that subsidies increase the cost of the goods by making it too easy for consumers to buy more of them.

3.3 Subsidies

It is not uncommon for governments to subsidize a good and claim that doing so will lower its cost and increase the amount consumed. Subsidies do increase the amount consumed, as advertised, but they increase costs to consumers. There are two general ways for government to subsidize a good: (1) subsidize its production by paying suppliers a given amount for each unit produced, and (2) subsidize its consumption by paying consumers a given amount for each unit consumed. We assume that the subsidies are given in direct money payments.[8] Although there can be political reasons from preferring subsidies to suppliers, or subsidies to consumers, or some combination of the two, for a given per-unit subsidy the effects on output, consumption and costs are the same for any split between producers and suppliers. We first consider producer subsidies, then consumer subsidies, and finally we discuss the costs of subsidizing a good in addition to those reflected in the direct cost of the subsidy.

3.3.1 Production Subsidies

In Fig. 3.2, the unsubsidized demand and supply curves for a good are shown as D and S respectively, with the equilibrium price and quantity given by P_E and Q_E. If it is determined through the political process that more of the product should be consumed, one way of accomplishing this is by lowering the price of the good by subsidizing its production. Assuming that producers are paid a given amount, M, for each unit produced, the private marginal cost of production will drop by M, which shifts the supply curve down by the same amount. This new supply curve is shown in Fig. 3.2 as S_S. The demand curve D is unaffected by the production subsidy and the subsidized equilibrium is determined by the intersection of D and S_S, and shown by price P_S and quantity Q_S. The subsidy reduces the price consumers pay for the good, but increases its cost. The money needed to pay for the subsidy has to be paid for through some combination of current taxes, future taxes, or inflation, and it is

[8] Subsidies are often paid to producers in less obvious ways for political reasons, as when farmers in arid areas receive water from expensive water diversion projects at a small fraction of the costs, or when governments guaranty loans to producers which allow them to pay lower interest rates. Although such subsidies are more convoluted than direct cash subsidies, our analysis of the latter applies to any subsidy that lowers the private marginal cost of production.

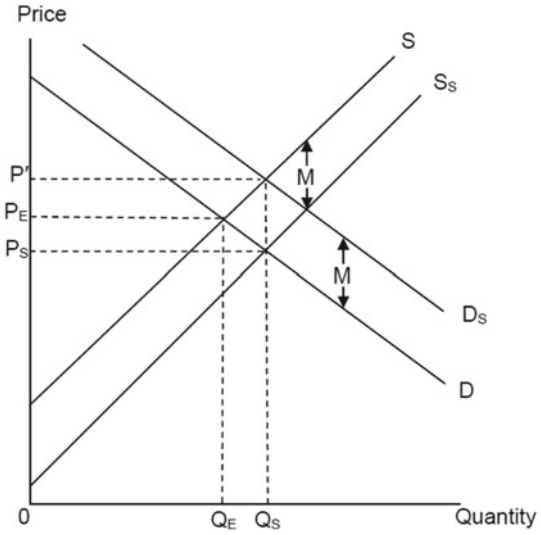

Fig. 3.2 Subsidies

consumers (all of us) who will do the paying.[9] So the cost of each unit of the good is given by P', the price P_S plus M, the per-unit cost of subsidy-which is greater than the cost, given by P_E, which is the full per-unit cost without the subsidy.[10]

The political advantage from the subsidy is that its effect on supply is easily noticed and appreciated by producers, while its cost is typically ignored by consumers and has no effect on the amount demanded. There are fewer producers of a subsidized good than there are consumers of it, so the subsidy to each producer is much larger than the taxes each consumer pays to finance it. Also, few of the taxes people pay come as itemized bills. Taxpayers often have no idea if a good they are consuming is being subsidized, and almost never know the per-unit cost of the subsidy or how much of that cost they are paying with their taxes. Also, since the total amount a producer receives from the subsidy increases as he produces more, the subsidy reduces the private marginal cost of production, as seen by the downward shift in the supply from S to S_S. It is this increase in supply that explains the decrease in price from P_E to P_S.

[9] Another possible way to pay for the subsidy is to reduce government spending on something else. This might seem to be an attractive possibility since it would make the subsidy costless to consumers if it were paid for by reducing government spending that is wasteful already. But despite ample examples of completely wasteful government spending, eliminating such spending is not likely to be very popular to politicians. If politicians were looking for ways to eliminate wasteful spending, they would not be looking for ways to finance government subsidies that increase the cost of goods and services.

[10] It is noted that this conclusion depends on the supply, or marginal cost, curve being upward sloping at the equilibrium. This is almost always the case, although it is possible for marginal cost to be declining at the equilibrium as a result of a positive externality in production. We ignore positive production externalities in this paper, although we do consider arguments for subsidies based on a positive externality in consumption in Sect. 3.3.1.

This lower price is what consumers notice, not the cost of the subsidy they pay in higher taxes. And from each consumers perspective, it is completely rational to ignore the cost of a subsidy, even if she knows what it is, because it has effectively zero effect on the marginal cost she is paying. No consumer will see any increase in her total taxes if she buys another unit of a subsidized good. All she would accomplish by reducing her consumption of the good below the level where marginal value equals price would be to sacrifice some of her consumer's surplus to subsidize the consumption of others. This is a sacrifice few can be expected to make, which is why consumption is given by Q_S in Fig. 3.2, where the marginal value of the good, as measured by the height of the demand curve, equals the good's price. Each person's consumption of the good shifts almost all of the cost of the subsidy to others. The subsidy can be thought of as an invitation for people to benefit at the expense of others.[11]

Because of the subsidy, the price of the good no longer does what market prices typically do, which is communicate to producers and consumers the real costs of consumption decisions and motivate them to take these costs into consideration. Not surprisingly, when not all costs are communicated through prices that are paid directly, producers and consumers cease to respond to those costs in ways that keep them under control.

3.3.2 Consumption Subsidies

Instead of subsidizing producers for supplying a good, government can subsidize consumers for buying it. To examine this case, we first go back to demand curve D and supply curve S in Fig. 3.2 and the unsubsidized equilibrium price P_E and quantity Q_E. We next consider a consumption subsidy that takes a form similar to the production subsidy just examined – a specified payment to consumers for each unit of a good they purchase. This has the effect of shifting the demand curve up by the amount of the payment since the marginal value of the good to each consumer is increased by that amount. Assuming that the per-unit subsidy is the same as the production subsidy (M), the new subsidized demand curve is given by D_S, as shown in Fig. 3.2. Since the supply curve is not affected by the subsidy, the subsidized equilibrium in this case is given by price P' and quantity Q_S. Compared to the equilibrium with the production subsidy, the quantity produced and consumed remains the same, but the price has increased from P_S to P', an increase equal to the per-unit subsidy.

Although the price has increased, the marginal cost to each consumer is still given by P_S because of the subsidy received on every unit purchased. But the average

[11] Although all consumers pay the same price for the product, the per-unit cost each consumer pays for the good varies. Those who pay little in taxes and consume lots of the subsidized good shift much of the per-unit cost of the subsidy to those who pay lots in taxes and consume little of the good.

amount consumers pay for each unit of the good continues to be P' once the tax-cost of the subsidy is considered. As in the previous case, some consumers will pay more of the tax cost for the subsidy they receive than others. But again, the tax cost has no effect on any consumers marginal cost, and therefore no effect on the amount consumed. The subsidy remains an invitation for people to gain at the expense of each other since the price fails to motivate consumers to take all the costs of their decisions into consideration.

So whether the per-unit subsidy is paid entirely to the producer or entirely to the consumer, the effect on consumption and production is the same. And in both cases, the governments attempt to lower the cost of a good, or give consumers the impression that it is lowering the cost, actually increases the cost. It is easily shown that if the subsidy were split between producers and consumers the effect on consumption, production and cost would remain the same regardless of the relative amount given to each.

One can argue that a subsidy can be justified despite the result of our analysis. It might be that consuming the good creates a positive externality with efficiency being increased by motivating more consumption of the good with a subsidy. Education, for example, is often cited as a good that should be subsidized because education is thought to provide benefits to society in addition to those captured by the direct consumer.[12] There is nothing in our argument that denies the existence of such positive consumption externalities or the possibility that when they exist a government subsidy can, if correctly sized and targeted, add value that exceeds the additional cost that results.[13] Our point is that subsidizing a good almost always increases its costs.

3.3.3 Additional Costs

The political motivation for providing subsidies is typically influenced significantly by the political influence of those receiving them, particularly when the recipients are well-organized interest groups.[14] The exercise of this influence can generate additional costs in addition to the direct tax costs discussed in the two previous

[12] See Hall (2000, 2006) for viewpoints on this argument.

[13] Unfortunately, the information necessary to know the size of a subsidy that will efficiently internalize an externality is rarely, if ever, available. And even if it were available, the political considerations that determine what goods are subsidized, and by how much, seldom have much to do with economic efficiency.

[14] In general, producers are better organized than consumers, and so producers can be expected to have more influence on the type and size of subsidies than consumers. This does not imply that political influence will favor producer over consumer subsidies. Producers can benefit from consumer subsidies as much as they do from producer subsidies and may favor the former because the benefit they receive from consumer subsidies is less direct and obvious than it is from producer subsidies.

subsections and result in a larger subsidy than warranted by any positive externality that may exist.

Political benefits seldom come free of charge. A group seeking benefits from a government subsidy has to compete for the attention and favor of politicians against many other groups who want political benefits. This "rent seeking" is costly, requiring access to key politicians and the ability to make a plausible case that the subsidy is in the public interest and a compelling case that it is in the political interest of those best able to get the necessary legislation passed.[15] The rent-seeking cost takes the form of campaign contributions, hiring well-connected lobbyists and recognized experts, generating favorable publicity and mobilizing at least the appearance of public support for the subsidy. And, much of this expense has to continue after the legislation is passed if the political support necessary to protect the subsidy against ongoing competition for government largesse is to be maintained. This rent-seeking cost is mostly a real cost-not just a transfer-since rent seeking employs resources with valuable alternative uses. The size of the per-unit subsidy may be positively related to the amount spent on rent seeking, but once the subsidy is "purchased" and the shifts in the demand and/or supply curves have taken place, the rent-seeking cost is independent of how many units of the good are produced and consumed. So the rent-seeking cost of the subsidy is a fixed (as opposed to a marginal) cost that, at least in the short-run, has no effect on production and consumption decisions.[16] It is a cost, however, that the subsidy adds to the production and consumption of the good over and above the direct tax cost. And, although it is consumers who will ultimately pay this rent-seeking cost, they completely ignore this cost in their consumption decisions.

Another cost of subsidizing a good that is invariably disregarded by consumers results from the fact that it always cost more than a dollar to raise a dollar in tax revenue. A significant portion of this extra cost is referred to by economists as the excess burden of taxation. People respond to taxes by making choices that create less value than those they would have made without taxes. For example, an employee may be willing to move to take a more productive job for the extra income her employer is willing to pay her, but not for the additional after-tax income. The net value that fails to be created is the excess burden of the income tax in this case. There have been many studies attempting to estimate the size of the excess burden of taxation. The estimates vary depending on the study and the tax being considered. In the case of the income tax, the tax that is probably the most commonly studied, it is routinely found that the excess burden of raising another dollar is $.25 (an excess burden of

[15]Tullock (1967) provided the first systematic analysis of rent seeking, although he did not use that term (Tullock 2003). The term was coined by Krueger (1974) in a paper that considered examples of the competition for political influence. Tollison (1982) surveyed the main themes and implications of the rent seeking literature.

[16]If the subsidy results in more output, and therefore profits, than anticipated, the politicians may decide to share in the unanticipated bounty of the suppliers by increasing the rent-seeking payments for maintaining the subsidy. So in the long run, the extra output might result in higher costs.

25 %) or higher.[17] There is little evidence that politicians consider the excess burden of taxation when making spending decisions. Doing so would require, for example, rejecting a project that would create $1.2 million in value and could be financed with $1 million in taxes since, with an excess burden of taxation of 25 %, the cost of the project is $1.25 million. Neither do consumers consider the excess burden of taxation when considering how much the subsidized good actually costs or how much of it to purchase. As discussed previously, none of the tax cost of a subsidy affects the marginal cost to the consumer.

Another obvious, but indirect, cost of a subsidy is the cost the government incurs to collect taxes and taxpayers incur to keep the records and prepare (or pay someone else to prepare) the forms required when paying taxes. Both of these tax-related costs are borne by taxpayers. As before, however, these costs are not seen by consumers as being connected to the total or marginal cost of a subsidized good. Consumers see the subsidy as lowering the cost of the good, but fail to see that this reduction is being more than offset by the costs of the subsidy, both direct and indirect.

3.4 Third-Party Payments and Insurance

Subsidizing a good with tax revenues to make consumers believe that its cost is being reduced can be thought of as an example of a third-party payment. Instead of each person paying the entire cost of his consumption directly, much of the costs are covered by taxes paid by others (third parties), with there being no relationship between the amount of the subsidy an individual pays and the amount he consumes. So the previous analysis of government subsidies is completely applicable to some third-party payments, but not all arrangements involving third-party payments are the same. For example, private medical insurance is an arrangement where the cost of medical care is not paid entirely by direct payments from the person receiving the care. The care each person receives is being subsidized in part by the payments of others-third-party payments-in the form of insurance premiums.[18] Although publicly subsidized medical care can be considered a form of insurance, and is often justified as such, there are important differences between it and privately provided medical insurance, even though both increase the cost of the care.

Consider first the similar effects both have on costs. The considerations discussed in Sect. 3.3 explain why subsidizing medical care with taxation increases the cost of that care. For some of the same reasons, privately provided medical insurance

[17] See Vedder and Gallaway (1999) for a discussion of different estimates.

[18] Of course, each person is also contributing to the care of everyone else with his own premium payment. As with government subsidies considered in Sect. 3.3, some people will end up paying more in premiums than they receive, and others will end up paying less. One can argue that it is only the latter that are being subsidized. But the important point is that everyone purchasing a good subsidized in part by insurance premiums, or taxes, will ignore the amount the subsidy is costing him when deciding how much of the good to consume. Premium payments for insurance are not marginal costs.

also increases the cost of medical care. Because the premium payments are not marginal cost, those payments are not considered in the decisions on how much care to purchase. The result is that medical care for the average consumer ends up costing more as consumption is expanded beyond the point where its marginal value is equal to its increasing marginal cost, because much of the cost is being paid for indirectly through insurance premiums. And, insurance premiums, as with the cost of taxes, are greater than the amount returned to consumers in the form of lower direct health-care payments. Much of the premium payments go to covering the cost of the personnel required by the insurance companies to collect the premiums, distribute the payments and keep track of it all. Furthermore, hospitals and doctors are also responsible for much of the cost of all this recordkeeping, and this cost also gets passed on to consumers.

Paper work is commonly required when transactions are made, but it is less than it would be otherwise when people are disciplined by spending their own money. Spending discipline is reduced when other peoples money is being spent, and restoring some of this lost discipline with cumbersome forms, regulations and aggravating red tape seems inevitable. Indeed, red tape can be justified as a way of moderating the moral hazards that result when people are able to shift the costs of their decisions onto others, as they do when not exercising proper care to avoid costs being subsidized or insured against. Moral hazards are an inherent feature of third-party payments, whether they result from public subsidies or private insurance, and provide another explanation for why subsidies and insurance increase the cost of goods by substituting non-marginal costs for marginal costs.[19]

In the case of most private insurance, the cost of moral hazards and red tape are reduced with insurance policies that require policy holders to pay a significant amount of the loss that is being insured against, with the insurance covering only large losses. This is referred to as high-deductible insurance. Fire insurance on a house is an example. Obviously, people are going to be more responsible in avoiding fire hazards (be less prone to moral hazards) if their fire insurance has a high deductible than if it has a low, or no, deductible. By keeping fire costs down and insurance premiums low, high-deductible fire insurance is more attractive than low-deductible insurance to most insurance customers, which explains why it is readily available from private insurance companies. The third-party payments that result from all insurance clearly have the effect of increasing the cost of that care. But, competition in the provision of private insurance generally moderates this effect by limiting the subsidy to unlikely, unpredictable, and relatively costly occurrences.

Interestingly, however, high-deductible medical insurance is not very popular, even though it has the same advantages as other types of high-deductible insurance. Most medical insurance is low-deductible/low-co-pay insurance, where once the medical cost reaches a relatively small threshold of a few hundred dollars (the deductible) the insurance pays most of the additional cost, leaving a small co-pay for

[19]The extra cost resulting from moral hazards is often more than justified in the case of insurance, because of the value people receive from replacing the low risk of a large and unpredictable cost with a the certain cost of small and predictable payments for insurance premiums.

the insured (commonly 20 %). This insurance is significantly more expensive than high-deductible medical insurance since it creates little incentive for the insured to make benefit-cost comparisons between different medical options, which biases decisions toward more costly choices, reduces the price competition faced by suppliers, and increases insurance premiums.[20] The question is: why is the type of medical insurance that results in higher medical cost and higher insurance premiums so popular? The answer is: government policy is disguising the cost of most medical insurance with a tax subsidy.

Most medical insurance is provided through businesses as part of employee compensation. Employees are in effect purchasing medical insurance from their employers and paying for it with lower salaries and wages than they would receive otherwise. This is an arrangement that can make both sides of the exchange better off when, as is often the case, the employer can buy group medical insurance for less than employees would pay individually, and for less than employees are willing to pay. The problem results from the fact that the value of employer-provided medical insurance is not taxed while monetary compensation is, so the more workers pay their medical bills with medical insurance premiums (which are part of their compensation, but not taxed) instead of directly in the form of copayments (which are taxed), the more they save in taxes.[21] So the limits on third-party payments that most private insurance contains are relaxed on most medical insurance because of the tax subsidy, which leads to more medical care being consumed, higher costs for medical care and for the insurance premiums that pay for much of it. As argued in Sect. 3.3, government subsidies almost always increase the cost of what is being subsidized.

The discussion in this section touches on some of the issues that are critical to any meaningful healthcare reform. The demand for reform is being driven by concern over the high cost of medical care and medical insurance, with proponents of different approaches to reform all claiming that their approach will make medical care available to more people by lowering its cost. Our prediction is that any reform legislation that satisfies the political requirements for passage will continue the trend that medical-care reform has long taken-substituting yet more subsidies and third-party payments for direct payments to give the impression that medical-care costs have been reduced, or have at least been shifted to someone else. If this prediction is correct, we are convinced that the result will be higher medical-care cost, and continued demand for reform. We are reminded of the joke: If you think medical care is costly now, just wait until its free.

[20] As reported in Cogan et al. (2005, p. 40) the average family medical insurance policy cost about $7,000 per year in the early 2000s, which reflected a high percentage of low-deductible/low-co-pay policies. At the same time, the median annual premium payment for medical insurance for a family of four with a $3,000 deductible was $2,683.

[21] This also makes providing low-deductible/low-co-pay medical insurance more attractive as a way to pay workers.

3.5 Conclusions

When politicians recommend policies to reduce the costs of goods, they invariably have in mind policies that censor the communication of information transmitted through market prices. This censorship reduces the coordination between suppliers and consumers needed to produce goods as cheaply as possible and supply them to the point where their marginal production cost equals their marginal value. The result is that government policies aimed at reducing costs almost always increase them.

The most blatant way governments censor price information to lower the cost of a good is by imposing a price ceiling below the equilibrium price. This has the effect of increasing cost by reducing the amount supplied and therefore increasing the amount people are willing to pay for the marginal unit. The competition between consumers for the good that follows can take many forms, but always include incurring significant costs associated with the inconvenience of dealing with an artificial shortage. This competition increases the non-price cost enough to elevate the cost of acquiring the good above what its price would be without the price ceiling. Also, government commonly steps in with restrictions to ration the goods that motivate other costly adjustments on consumers. Without being permitted to pay more than the legal price ceiling for the good, none of the costs consumers incur do anything to motivate changes in the amount supplied, which is the only thing that would reduce the price, and cost, of the good to its free-market level.

Another way governments distort price information in efforts to reduce, or appear to reduce, the costs of goods is by directly subsidizing their production or consumption with tax revenues. Once the tax-costs of the subsidies, and the related rent-seeking and excess burdens, are considered, the costs of the subsidized goods are higher than they would be without a subsidy. Consumers of publicly subsidized goods do not connect the tax they pay to support the subsidy with the goods' cost. Even if they did, the tax cost of the subsidy is ignored in consumption decisions because it does not affect the marginal cost of consumption.

The third-party payments inherent in insurance also lead to increased costs by lowering the marginal costs of choices that are subsidized by insurance premiums which each of the insured see as a fixed cost. The moral hazards that result (from government subsidies as well as from insurance) are typically moderated in the case of private insurance by high deductibles requiring direct payments on routine and predictable expenses, with the insurance reimbursing only relatively costly expenditures that are unusual and unforeseeable. But, in the case of medical insurance, a tax subsidy exists that biases all employer-provided policies toward low deductibles and low copayments which exacerbates the moral hazard and significantly increases the cost of medical care and medical insurance.

Although price ceilings increase costs by reducing the amount supplied and consumed, and public subsidies increase costs by increasing the amount supplied and consumed, there are similarities between the two. The increased costs caused by governments often result in pressures on governments to do more to reduce the costs that their cost-reducing policies have increased. In the case of price ceilings, the

response is often to increase penalties on those engaged in black market activities that are actually reducing costs. Also, price ceilings are commonly accompanied by government imposed rationing schemes that increase costs with one-size-fits-all rules that make it even more difficult to direct available goods to go to those who value them most.

In the case of public subsidies, the political response to complaints about escalating costs is typically to further reduce the amount consumers are paying directly by increasing the government subsidy. This has clearly been the history of government efforts to contain the rising cost of medical care.[22] The result is always more cost escalation caused by increased demand, and more political pressure for even larger subsidies. This process eventually leads to a government attempt to suppress the amount demanded with government rationing. This is true even though the original rational for medical-care subsidies was to increase the consumption of medical care. Elderly people are justifiably concerned that more of this reform will make it more difficult for them to receive care that could prolong their lives. As opposed to the economic stimulus plan, health-care reform might really be "shovel ready."

We do not want to leave the impression that government is always inept in its attempts to alter costs. When it sets out to increase costs, it can be depended upon to do an excellent job. Of course, politicians never brag about their ability to increase costs, or admit that is what they are doing. When enacting policies that increase cost, the claim is always that some noble purpose is being achieved such as saving American jobs (import restrictions); protecting family farms (agricultural price supports); achieving energy independence (mandated use of corn-based ethanol and tougher CAFE standards); improving education (limitations on the competition faced by public schools); or protecting the environment (mandated use of renewable energy sources). But no matter what politicians are claiming to do, when they are substituting political choices for those individuals would make in response to uncensored market information, what they are almost always doing is increasing costs.

References

Clark JR, Lee DR (2008) Censoring and destroying information in the information age. Cato J 28(3):421–434

Cogan JF, Hubbard RG, Kessler DP (2005) Healthy, wealthy, and wise: five steps to a better health care system. AEI Press, Washington

Goodman JC, Musgrave G (1992) Patient power: solving America's health care crisis. Cato Institute, Washington

Goodman JC, Gorman L, Herrick D, Sade RM (2009) Health care reform: do other countries have the answers? National center for policy analysis, Dallas TX

[22] According to Goodman and Musgrave (1992, p. 232), 51.6 % of all personal medical expenses in the U.S. in 1965 were paid directly by those receiving the care. In a recent paper, Goodman et al. (2009) reports that the amount of personal medical expenses paid directly for health care in the United States was only 13 %, while the average for OEDC countries was 20 %.

Hall JC (2000) Investment in education: public and private returns. Joint economic committee, Washington DC
Hall JC (2006) Positive externalities and government involvement in education. J Priv Enterp 21(2):165–175
Jacobe D (2008) Majority of Americans support price controls on gas. Gallup.com
Krueger AO (1974) The political economy of the rent-seeking society. Am Econ Rev 64(3):291–303
Tollison RD (1982) Rent seeking: a survey. Kyklos 35(4):575–602
Tullock G (1967) The welfare costs of tariffs, monopolies, and theft. West Econ J 5(3):224–232
Tullock G (2003) The origin of the rent-seeking concept. Int J Bus Econ 2(1):1–8
Vedder RK, Gallaway LE (1999) Tax reduction and economic welfare. Joint economic committee, Washington DC

Chapter 4
The Great Depression: A Tale of Three Paradigms

Lowell E. Gallaway and Richard K. Vedder

Abstract In this article we articulate three distinct paradigms about the cause of the Great Depression. These three paradigms are: the neoclassical view epitomized by Murray Rothbard and William Hutt, the intermediate view of Peter Temin, and the underconsumptionist views expressed by Herbert Hoover and Henry Ford. We present empirical evidence in favor of the Rothbard–Hutt view. The neoclassical labor market adjustment mechanism that effectively reversed the downturn of 1920–1921 did not operate in the mid-to-late 1930s, perhaps because of underconsumptionist ideology.

4.1 Introduction

> Nearly everything in this country is too high priced. The only thing that should be high priced in this country is the man who works. Wages must not come down, they must not even stay on their present level; they must go up. And even that is not sufficient of itself–we must see to it that the increased wages are not taken away from our people by increased prices that do not represent increased values.

With that statement reported in the *New York Times* (Gallaway and Vedder 1997, p. 92), Henry Ford set forth his formula for dealing with what was perceived to be rising economic distress following the stock market crash of the Autumn of 1929. Ford's remarks are reminiscent of the economics of John A. Hobson, who in the early 1920s disagreed sharply with what he felt was the conventional orthodoxy of the day, viz., that, "In depressed trade, with general unemployment, business men have considerable support from economists in calling for cuts in real wages..." (Hobson 1923, p. 84). Hobson's counterarguments bear a striking similarity to some of the reasoning presented over a decade later in Chap. 19 of Keynes' *General Theory*, employing the interrelationships between wage levels, consumption (i.e., effective demand), and prices (Keynes 1936). This underconsumptionist view was fairly widespread. Herbert Hoover, with the support of business leaders (including Ford) in attendance at

L.E. Gallaway · R.K. Vedder (✉)
Ohio University, Bentley Hall Annex, Athens, OH 45701, USA
e-mail: vedder@ohiou.edu

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_4

a November 21, 1929, White House conference espoused it, as represented by his statement summarizing that conference (New York Times 1929, p. 1):

> The President was authorized by the employers who were present at this morning's conference to state on their individual behalf that they will not initiate any movement for wage reduction, and it was their strong recommendation that this attitude should be pursued by the country as a whole.
>
> They considered that, aside from the human consideration involved, the consuming power of the country will thereby be maintained.

The support for an underconsumptionist oriented policy approach would be understandable coming from either the political or trade union sectors of the economy.[1] From the business sector, though, it may seem somewhat surprising.[2] Rothboard (1963), p. 45 offers an explanation as to why business leaders would support such a policy:

> As early as the 1920s, 'big' businessmen were swayed by 'enlightened' and 'progressive' ideas, one of which was mistakenly held that American prosperity was caused by the payment of high wages (rates) instead of the other way around .... By the time of the depression ... businessmen were ripe for believing that lowering wage rates would cut 'purchasing power' (consumption) and worsen the depression....

As to Hoover, his memoirs and papers suggest that he found morally and intellectually unacceptable the means employed in dealing with earlier incidents of depressed economic conditions. He termed it the "liquidation" of labor and he opposed it on two grounds. First, "labor was not a commodity: it represented human bones". Second, the underconsumptionist doctrines already discussed had clearly captured Hoover's mind (Rothbard 1963).[3]

Of course, the critical question in this respect is whether the formally stated policy positions of some business leaders and the President had any actual impact on the behavior of wage rates. Rothbard believes they did, but at least two chroniclers of the history of this era suggest that wage stability was not maintained. Broadus Mitchell (1947, p. 84) claims that, "The obligation [of industry] not to cut wages was ... widely dishonored" and Arthur Schlesinger Jr. states that, "The entire wage structure was apparently condemned to disintegration (Schlesinger Jr 1957, p. 249)". The whole thrust of Rothbard's argument is that the underconsumptionist posture of Hoover and the business leaders interfered with the normal processes of labor market adjustment, a view that is concurred in by W.H. Hutt.[4] Interestingly, an intermediate view is that

---

[1] For example, the position of the American Federation of Labor (AFL) at this time is suggested by a statement by the AFL's John P. Frey in 1929 relating to a public works scheme of Hoover's. Frey's statement was to the effect that the President was in agreement with the AFL's position that depressions were the result of underconsumption and low wages. See Dorfman (1959), pp. 349–50.

[2] Not to be ignored in this respect is the fact that ideas such as Hobson's and Hoover's were not as unorthodox among professional economists as sometimes claimed. See Davis (1971) pp. 94–99, who presents an interesting array of statements by economists and other academics relating to the issue of the impact of wage reductions on the economy.

[3] For more on the underconsumptionist doctrine and Hoover, see Chap. 8 of Rothbard (1963).

[4] Hutt's views on the general subject of the underconsumptionist position are well summarized in his essay "Illustration of Keynesianism" from Hutt (1971).

of Peter Temin, who offers certain statistical data to suggest that the real wage rate did not rise during the Great Depression (Temin 1976). Temin's purpose in presenting these data is to demonstrate that the neoclassical argument that unemployment was caused by real wage rates being too high is invalid.

Evaluating whether the announced goal of maintaining relatively high real wage levels was achieved during the Great Depression is complicated by the dilemma of determining what is the "normal" pattern of behavior of real wage rates during a downturn in the business cycle. Normal in the neoclassical theory that Hobson attacked would be the high levels of real wage rates advocated by the underconsumptionists, while normal for the underconsumptionists would be the low levels desired by the neoclassicists. To at least partially deal with this problem, we will compare the early period of the Great Depression with a similar interval during the economic crisis that marked the immediate post-World War I period. That economic downturn is chosen for comparison purposes because it initially shows an almost identical degree of decline in employment levels. Measured by the Federal Reserve Board series on factory employment, the post-World War I downturn begins in the second quarter of 1920. By fourth quarter 1921 (a period of seven quarters into the cycle), employment levels have fallen to 72.9 % of what they were in the first quarter of 1920. The downturn in factory employment in the Great Depression begins in the fourth quarter of 1929. By the second quarter of 1931 (seven quarters into the depression) employment is at 73.6 % of the third quarter 1929 level.

A further problem is the choice of a suitable real wage measure. At the heart of the underconsumptionist position are certain notions about the impact of the distribution of income on levels of aggregate effective demand. Hobson argues that a ceteris paribus increase in the real wage rate will redistribute income from the propertied to the laboring classes with an appropriate stimulus to consumption (Hobson 1923). If this is the underconsumptionist position, and we believe it is, the critical real wage measure is one that controls for changing levels of productivity (i.e., alterations in the per capita level of real income brought about by increases in total factor productivity).[5] Thus, what is important to the underconsumptionists is a change in the real wage rate which results in its being higher than would be expected given the change in productivity per unit of labor that occurred in the same period.

4.2 The Adjusted Real Wage

Such a wage measure, which we shall call the adjusted real wage, can be calculated from data available in the various issues of the *Federal Reserve Bulletin*. The specific measure of the we have calculated is given by

[5] All other variants of the underconsumption argument that we are familiar with come back to the same basic position, viz., there is a positive relationship between the level of wage rates and employment. See, for example, Foster and Catchings (1925) or (1927).

Table 4.1 Comparison of employment and adjusted real wage

Cycle referenced on first quarter 1920			Cycle referenced on third quarter 1929	
Cycle in quarter	Employment	Adjusted real wage	Employment	Adjusted real wage
1	100.0	100.0	100.0	100.0
2	97.6	94.6	96.2	112.5
3	93.5	104.8	90.1	115.7
4	83.2	122.1	87.7	109.3
5	71.2	133.1	82.9	120.1
6	71.6	150.3	78.7	126.7
7	71.1	139.9	74.1	117.0
8	72.9	128.8	73.6	119.8
9	73.6	120.9	71.6	125.8
10	76.7	118.3	67.4	130.5
11	81.2	114.8	64.6	126.7
12	85.6	107.4	60.1	127.2

Note Comparison is for first 12 quarters of the 1920–1922 business cycle and the Great Depression

$$R = \frac{W}{OP} \tag{4.1}$$

where R denotes the adjusted real wage in manufacturing, W is the manufacturing wage bill (given by the data series describing total factory payrolls), O is the Federal Reserve Board index of industrial production, and P is the wholesale price level.

While this expression may seem unusual, the real wage (w_r) equals the money wage (W_m) divided by the price level (P), i.e., $w_r = W_m/P$. The average productivity of labor equals the total real output (O) divided by the quantity of labor employed (L), i.e., $\pi = (O/L)$. Dividing the real wage by the average productivity of labor to obtain the adjusted real wage (R) gives $(W_m/P)/(O/L)$, which simplifies to $(W_m/L)/(OP)$. The numerator is the total wage bill and the denominator is the money value of total output and income.

The calculated values of R and the level of employment during the downturns that began in the second quarter 1920 and fourth quarter 1929 are shown in Table 4.1. All data series are in index number form, with the quarter before beginning of the downturn set equal to 100. The pattern of movement in the real wage rate statistic in the 1920 cycle is one of constant upward movement after the second quarter of the cycle through the sixth quarter and then a decline through the twelfth quarter. At that point, the wage variable is only 7.4 % higher than it was in the first quarter of 1920. In the cycle beginning in 1929, the movement is more consistently upward with the peak coming in the tenth quarter of the cycle. By the twelfth quarter, the wage variable is still 27.2 % greater than it was in the third quarter of 1929.

The adjusted real wage rate data of Table 4.1 are suggestive that real wage levels were maintained, and in fact increased, during the Great Depression to a far greater

extent than they were during the 1920–1921 depression in economic activity.[6] This would seem to reflect adversely on the Hoover–Hobson–Ford version of underconsumptionism while being supportive of the Rothbard–Hutt view of the events subsequent to 1929. However, it does not constitute a formal test of the opposing views nor does it come to grips with the intermediate position taken by Temin and supported by certain of Keynes' theoretical notions. The Keynes–Temin view is that wage changes are basically irrelevant, being neither the cause nor the cure for the unemployment of the 1930s.[7]

It should be noted that in some respects it is perhaps unfair to ascribe fully to Keynes this intermediate view that wage changes are irrelevant. Keynes agreed with the basic neoclassical premise that unemployment was associated with the real wage rate being out of equilibrium on the high side (Keynes 1936). It is only after one reaches such a position that Keynes argues that wage changes will have no effect. His view on the role of the wage rate, though, as a fundamental source of unemployment is essentially the same as that taken by Rothbard and Hutt. Thus, assigning Keynes solely to one of these three positions is something of an exaggeration. However, on balance, he is probably closer to the intermediate position.

4.3 Testing the Three Paradigms

To more fully evaluate the various paradigms that have been advanced to explain the events of the 1930s, we will begin by defining a model that is essentially neoclassical in character, i.e., of the Rothbard–Hutt type, and attempt to empirically evaluate its validity. The neoclassical view of the labor market adjustment mechanism can be described through a basic supply and demand model. Define the demand for labor as

$$D = f(R), f'(R)\ 0 \tag{4.2}$$

where D denotes the demand for labor per unit of the population in the economy. Similarly, the supply of labor per unit of population (S) is given by

$$S = \phi(R), \phi'(R) > 0 \tag{4.3}$$

Equations 4.2 and 4.3 can be combined to provide an unemployment rate (U) equation

$$U = \frac{(S - D)}{S} = \frac{[\phi(R) - f(R)]}{\phi(R)} \tag{4.4}$$

[6] Others have made this observation. For example, Wolman (1931) observed, "It is indeed impossible to recall any past depression of similar intensity and duration in which the wages of prosperity were sustained as long as they have been during the depression of 1930–1931".

[7] See Temin (1976), p. 140. Specifically, Temin remarks ".... in the post war debate over the Keynesian system, one of the dominant questions was whether an unemployment equilibrium was possible. The consensus now seems to be accepted that in the long run it is not".

Differentiating 4.4 and keeping in mind that $\phi(R)$ and $f(R)$ are always positive, it can be shown that

$$\frac{dU}{dR} > 0 \tag{4.5}$$

Thus we may write a first approximation of 4.4:

$$U = a + bR \tag{4.6}$$

This may be further expanded by defining R in the current period as follows:

$$R_t = R_{t-1} + \dot{w}_t - \dot{p}_t - \dot{\pi}_t \tag{4.7}$$

where t and t_1 represent different time periods, $\dot{w}_t$ is the rate of change in money wage rates, $\dot{p}_t$ is the rate of change in the price level, and $\dot{\pi}_t$ is the rate of change in productivity. All rates of change are between times t and t_1.

Combining 4.6 and 4.7 yields the following reduced form equation for the full model:

$$U_t = a + bR_{t-1} + c\dot{w}_t - d\dot{p}_t - e\dot{\pi}_t \tag{4.8}$$

This is the basic model we will test using annual data for the period 1901–1941. The use of annual data is necessitated because of a lack of information on unemployment rates for periods of less than a year. For this analysis we have used the standard data series for unemployment presented in *Historical Statistics of the United States* (Series D-86) (US Bureau of the Census 1975). Price change data are also taken from *Historical Statistics*, with the choice being Series E-135 for the consumer price index. Several options are available with respect to wage rate and productivity series. We have chosen to use two wage series, Lebergott's annual earnings of workers while employed (Historical D-724) and the David-Solar index of unskilled hourly wage rates (David and Solar 1977).

For productivity measures, we use John Kendrick's estimates, as reported in *Historical Statistics*, employing an annual output series with Lebergott's earnings series and an hourly series with the David-Solar wage measure (Historical D-724). Both sets of measures are included in the analysis at the same time. The rationale for this is that the Lebergott series measures average wage levels while the David-Solar series captures any differential movements in unskilled wage levels compared to the average. All data series are indexed with 1929 = 100. With all the variables included, the basic regression model (with the expected signs indicated) is:

$$\begin{aligned} U_t = a + b(R_{t-1})_L + c(R_{t-1})_{DS} + d(\dot{w}_t)_L + e(\dot{w}_t)_{DS} \\ -f\dot{p}_t - g(\dot{\pi}_t)_L - h(\dot{\pi}_t)_{DS} + u \end{aligned} \tag{4.9}$$

where the subscripts L and DS denote the Lebergott and David-Solar data series (or the productivity series used with them), respectively, and u is a random error term. The estimation yields the results in Table 4.2

Table 4.2 Adjusted real wage regression

Variable	U_t
$(R_{t-1})_L$	0.7314 (9.55)
$(R_{t-1})_{DS}$	0.3312 (7.94)
$(\dot{w}_t)_L$	0.689 (0.40)
$(\dot{w}_t)_{DS}$	0.2788 (2.76)
$\dot{p}_t$	−0.8772 (5.18)
$(\dot{\pi}_t)_L$	−0.6205 (3.46)
$(\dot{\pi}_t)_{DS}$	−0.0314 (0.18)
R-squared	0.9217
Adj. R-squared	0.9079
D-W	1.25

Note Absolute value of t-statistics in parentheses

The results are quite consistent with the basic hypotheses underlying the model. All variables have the expected signs and all but two (one money wage change and one productivity change variable) are significant at the one percent level or beyond. Collectively, the variables in the model explain over 90 % of the variation in the unemployment rate over the interval 1901–1941, a period that embraces such diverse events as the Panic of 1907, World War I, the prosperity of the 1920s, and the Great Depression.[8]

Of particular interest for our purposes is the performance of the model during the years 1929–1941, i.e., during the Great Depression and the approach to full recovery. Table 4.3 shows the actual unemployment rates for these years as well as the rates predicted by the model. An examination of the information shown there suggests that the neoclassical labor market adjustment model does a remarkably good job of explaining the behavior of the unemployment rate during the 1930s. This is a result that is quite consistent with the Rothbard–Hutt position and remarkably inconsistent with the Hoover–Hobson–Ford or Keynes–Temin paradigm of the Great Depression.

The earlier evidence comparing the first twelve quarters of the post-World War I business cycle and the Great Depression tends to confirm the results reported in Table 4.2. A diagrammatic representation of the data presented in Table 4.1 is given in Fig. 4.1. It shows that an economic downturn of greater initial severity than that beginning in 1929 was stemmed and then reversed by the operation of the neoclassical

[8]The Durbin–Watson statistic for this regression equation is in the indeterminate range.

Table 4.3 Actual and estimated unemployment rates

Unemployment rate		
Year	Actual (%)	Estimated (%)
1929	3.2	3.4
1930	8.7	9.2
1931	15.9	15.4
1932	23.6	24.8
1933	24.9	23.3
1934	21.7	24.3
1935	20.1	22.4
1936	16.9	20.2
1937	14.3	16.9
1938	19	20.7
1939	17.2	17.6
1940	14.6	14.1
1941	9.9	8.6

Note Estimated unemployment rates obtained from the model in Table 4.2

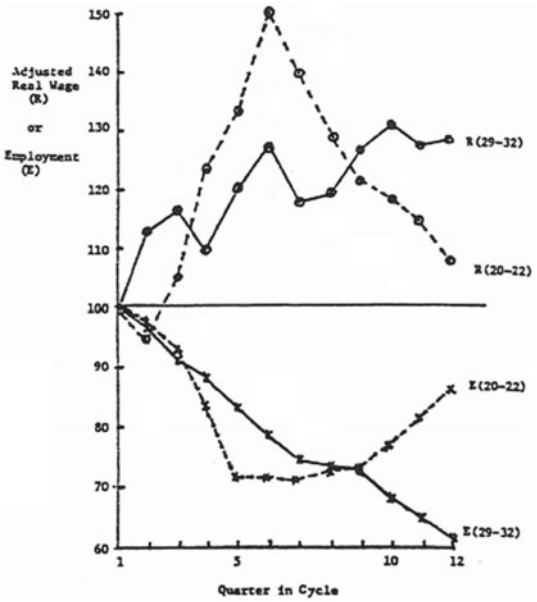

Fig. 4.1 Comparison of adjusted real wage and employment, 1920–1922 business cycle and great depression

adjustment mechanism, which took hold some five to six quarters into the downturn, producing a reversal in the rise of the adjusted real wage.

By contrast, ten quarters into the Great Depression, the adjusted real wage peaked and declined only slightly in the next two quarters. It is worth noting that the patterns

of these two cyclical swings in economic activity are quite similar for the first six quarters. Thereafter, though, the adjusted real wage rate continues to rise during the Great Depression and employment continues to decline, just as predicted by the neoclassical model.

4.4 Conclusion

What can we conclude from this brief inquiry into the causes of the Great Depression? Basically, we have articulated three distinct paradigms, ranging from the neoclassical view epitomized by Rothbard–Hutt to the underconsumptionist position advocated by Hoover–Hobson–Ford. Of the three explanatory schemata, the Rothbard–Hutt argument clearly has the better of the empirical evidence.[9] A strong argument can be made that the Great Depression acquired the adjective "Great" precisely because the neoclassical labor market adjustment mechanism that so effectively reversed the rather severe economic downturn of 1920–1921 did not operate after mid-to-late 1930. To the extent that business leaders and political figures allowed their judgment to be clouded by the siren call of underconsumptionist ideology, they must shoulder a substantial part of the blame for the escalation of human misery that resulted from the intensifying of the Great Depression. Herbert Hoover may not have wished to "liquidate" labor, but the achieving of the policies he advocated did just that, reminding one of Chesterton's (1908) incisive commentary from page 31 of *Robert Browning*.

> It is when men begin to grow desperate in their love for the people, when they are overwhelmed with the difficulties and blunders of humanity, that they fall back on a wild desire to manage everything themselves. ... [the] belief that all would go right if we could only get the strings into our own hands is a fallacy, almost without exception, but nobody can justly say that it is not public spirited.

References

Chesterton GK (1908) Robert browning. Grosset & Dunlap, New York
David PA, Solar P (1977) A bicentenary contribution to the history of the cost of living in America. Res Econ Hist 2(1):59–60
Davis JR (1971) The new economists and the old economists. Lowa State University Press, Ames, IA
Dorfman J (1959) The economic mind in American civilization. Viking, New York
Foster WT, Catchings W (1925) Profits. Houghton Mifflin Company, Boston
Foster WT, Catchings W (1927) Business without a buyer. Houghton Mifflin Company, Boston

[9] It is somewhat ironic to present empirical evidence to support a position advocated by Rothbard in that he would reject, on methodological grounds, the act of empirically testing propositions. So be it.

Gallaway LE, Vedder RK (1997) Out of work: unemployment and government in twentieth-century America. NYU Press, New York
Hobson JA (1923) Economics of unemployment. Macmillan, New York
Hutt WH (1971) Politically impossible. Institute for Economic Affairs, London
Keynes JM (1936) General theory of employment, interest and money. Macmillan, London
Mitchell B (1947) Depression decade. Holt, Rinehart, and Winston, New York
New York Times (1929) Hoover to set up business council. New York Times
Rothbard MN (1963) America's great depression. Van Nostrand, Princeton NJ
Schlesinger A Jr (1957) The age of Roosevelt: the crisis of the old order, 1919–1933. Houghton Mifflin, Boston, MA
Temin P (1976) Did monetary forces cause the great depression?. Norton, New York
US Bureau of the Census (1975) Historical statistics of the United States, colonial times to 1970. US Bureau of the Census, Washington DC
Wolman L (1931) Wages in relation to economic recovery. University of Chicago Press, Chicago

Chapter 5
Bad Economics, Good Law: The Concept of Externality

Roger E. Meiners

Abstract The term "externality" is pervasive in modern economics. Most microeconomic theory textbooks have a chapter devoted to the topic as do texts covering public economics. This chapter argues that law deals with the matter of externality in an economically efficient manner. Courts largely ignore the term externality despite its common use in economics and, more importantly, law has changed little to incorporate the now-common economic meaning of externality. Law, especially tort law, often deals with what economists would call relevant externalities. Economists often fail to understand what constitutes a relevant externality, resulting in the term being operationally meaningless.

5.1 Introduction

The term "externality" is pervasive in modern economics. For decades, beginning at the principles level, most microeconomic theory textbooks have a chapter devoted to the topic, as do texts covering public sector economics. In general, externality means the imposition of a cost on another party without consent, or the provision of a benefit without prior agreement.[1] It is the basis of a huge literature in economics, especially in the environmental area, that is used to justify a wide range of policy proposals to correct alleged defects in market-based arrangements that do not account for such costs.

This chapter argues that law deals with the matter of externality in an economically efficient manner. Courts largely ignore the term externality despite its common use in economics and, more importantly, law has changed little to incorporate the now-common economic meaning of externality. Law, especially tort law, often deals with what economists would call relevant externalities. As will be discussed, economists

[1]The latter is rarely of concern even though it is common. Details of the economic definition of externality will be reviewed later in the chapter.

R.E. Meiners (✉)
University of Texas at Arlington, 701 West Street, Arlington, TX 76019, USA
e-mail: meiners@uta.edu

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_5

often fail to understand what constitutes a relevant externality, resulting in the term being operationally meaningless in much discussion in economics. It has become a straw man that justifies nearly any policy prescription one wishes to advocate. Economic theory has gone astray, but law has not.

This chapter reviews the use of the term externality in law from its first appearance in a reported case up through 2012. We begin with a review of older cases, where the meaning of the term was not much related to its modern economic and policy use. We then look at its usage in the more modern content as it appears in case law. Then we turn to economic theory and contrast how externality is often presented to how law deals with relevant externalities.

5.2 First References in Law to Externality

The first published decision using the word externality was by the Supreme Court of Vermont in 1928. The court noted that Justice Holmes wrote that the words used in a will should be taken in the sense in which they would have been used by the testator in usual circumstances: "The normal speaker of English is merely a special variety, in literary form, so to speak, of our old friend the prudent man. He is external to the particular writer, and reference to him as the criterion is simply another instance of the externality of the law."[2] What Holmes meant was elaborated by the Mississippi high court several decades later.[3] If there is conflict about the presumed "internal" meaning of words, such as those in a will or a contract, one should look to the "external" or common meaning of the words in their normal context, not divine some peculiar construction. As we will see, this usage will appear in some cases in recent years.

The first use to the word externality in a reported federal case involved a land dispute by the Seneca Nation and the City of Salamanca, New York, when the Second Circuit noted, in 1942, that the law in medieval Germany suffered from "a multitude of unimportant externalities" meaning extreme formalism or "form-rigorism."[4] Judge Frank must have been enamored with the term as he used it again the next year in a contracts case. In discussing mutual assent in contracts, the court said: "We now speak

[2] *Middlebury College v. Central Power Corp. of Vermont*, 143 A. 384, 390 (Sup. Ct., Vt. 1928), citing Holmes (1899).

[3] In a contested will, the court looked to effectuate the textratrix's intention. "The criterion designated by Mr. Justice Holmes as 'the externality of the law' simply reflects that we must ask what the words used in the instrument would mean 'in the mouth of a normal speaker of English using them in the circumstances in which they were used…' (Holmes 1899, p. 417)" *Hemphill v. Mississippi State Highway Comm.*, 145 So.2d 455, 459 (Sup. Ct., Miss. 1962). The same court used the word in that context a few years later in *Yates v. State*, 189 So.2d 917 (Sup. Ct., Miss. 1966), stating, at 921, "The infeasibility of getting into the mind of a person to determine what he thought or believed is a sound reason for the principle of externality that requires judgments be based on something more than subjective statements of what one believes or thinks."

[4] *U.S. v. Forness*, 125 F.2d 928, 935 (2nd Cir. 1942).

of 'externality,' insisting on judicial consideration of only those manifestations of intention which are public ('open to the scrutiny and knowledge of the community') and not private ('secreted in the heart' of a person)."[5] Perhaps because the Supreme Court reversed that decision in one terse paragraph,[6] although for reasons not related to externality, the term did not reappear in a reported federal case for almost three decades.[7]

The term appears in a few other state court decisions. In a child custody dispute, the court said that the religious training given a child was one of the externalities (outside influences) that impacted the child.[8] In an insurance death benefits case, the issue was whether the cause of death was "external, violent and accidental," which the court refers to as an "externality," as opposed to death from an internal cause such as disease.[9] In a shareholder derivative suit, the court noted that when it stripped "the matter of all externalities" and just considered "the fundamental aspects of the transaction" the essence of the dispute was clear.[10]

For our discussion the cutoff point for "old cases" is 1972. The nine cases prior to 1973 used the term externality in a non-economic sense. After 1972 the term is used in a non-economic sense at times, but more commonly it is used in some economic sense. Sometimes the term is used carelessly, but generally it is used with a particular meaning. As we will see, in some areas of law, the term has come to have a particular meaning, hence the little areas of concentrations in the reporting of cases here.

Beginning in 1973, when the term began to be used in its modern sense, it appeared 225 times in reported decisions over 40 years (through 2012), for an average of about 5.5 uses per year in federal and state cases. Again, some of the uses are not in an economic sense, which would reduce the count further if our interest is in the economic meaning and application of externality. The usage in all cases is categorized by area of law in Table 5.1. The areas were chosen because of the number of times the term appeared. Areas of law with few references to externality, such as child custody and litigation procedure, are lumped together as "random."

The non-economic sense of externality is, as Justice Holmes (1899) explained, events or definitions that are external to the immediate legal issues or documents. There is no economic cost issue implied in such uses. When the term is used in an economic sense, meaning costs (or benefits) not accounted for by the decision maker, the usage is correct in most cases, but not all, as will be discussed. However, overall courts use the term judiciously. We review the use of externality in many cases to understand better how the concept is employed in law.

---

[5] *Zell v. American Seating Co.*, 138 F.2d 641, 646 (2nd Cir. 1943). The court noted in footnote 20a accompanying that sentence that Williston and Wigmore may not always agree with its interpretation.

[6] *American Seating Co. v. Zell*, 322 U.S. 709, 64 S.Ct. 1053 (1944).

[7] *City of Burlington v. Turner*, 356 F.Supp. 594 (S.D. Iowa 1972). The court discussed "external costs" of a bridge, meaning costs related to operating, as opposed to constructing, a bridge, at 608.

[8] *Boerger v. Boerger*, 97 A.2d 419 (Super. Ct., Chan. Div., N.J. 1953).

[9] *Towner v. Prudential Insurance Company of America*, 137 So.2d 449, 451 (Ct. App., La. 1962).

[10] *Urnest v. Forged Tooth Gear Co.*, 243 N.E.2d 596, 601. (Ct. App., Ill., 1968).

Table 5.1 Number of cases in which "Externality" appears through 2012

Classification		Number
Old cases (pre-1973)		9
Antitrust		17
Bankruptcy		9
Constitutional		16
Contracts		6
Crime		13
Environmental		
	NEPA	2
	Air	11
	Land	11
	Water	6
FCC, ICC, and PUC		10
Intellectual property		14
Labor		16
Takings		9
Tax		14
Torts		19
Zoning		25
Random		27
Total		234

Notes: Search also includes "externalities" and is limited to reported decisions

5.3 Antitrust

The first "externality case" post-1972 was an antitrust suit by the state of California contending that the automakers conspired to eliminate competition in research and development of air pollution equipment. The appeals court noted that the automakers considered antipollution devices to be "externalities, whose development would increase price without concomitant spur to consumer interest."[11] The use is vague; the court seemed to be saying that the automakers did not care for the costs they incurred by having to invest in pollution control devices, especially because consumers were not interested in paying for such devices. Not liking the fact of a cost does not make it an externality in the formal economic sense, but the intended use appears to be economic. The next antitrust case also used the term in the sense of an external force that impacted the market.[12] That is a common way in which courts use the term in decisions.

[11] *In re Multidistrict Vehicle Air Pollution*, 481 F.2d 122, 124 (9th Cir. 1973).

[12] *Knutson v. Daily Review, Inc.*, 383 F.Supp. 1346 (N.D. Calif., 1974), the court noted that a restriction on maximum resale prices were "an externality" imposed on the market (at 1383), meaning an external force that impacted the market.

The concept of network externality has been mentioned in several antitrust cases. This comes from the notion of network effects, a reference to the fact that when many people use the same "network," such as the telephone or Internet, it will become more useful. Network externality derives from network effects, meaning the change in benefits from a good or service that one derives when others are consuming the same good (Liebowitz and Margolis 1998; Metcalfe 2007). The decisions of users are made independently, but benefits are mutual. Those may include cost savings by sharing network costs among a group and the informational or use benefits that arise from an expanded network. External costs imposed by other network participants may include clogging the system (hogging bandwidth) and imposing junk (spam) seen as a bad by most participants.

Network externality first appeared in a decision in 1996. The operator of an ATM network challenged an order by the Federal Reserve System allowing a merger of several ATM systems. Upholding the decision, the court quoted the Board of Governors, which said "Network externalities… tend to promote the consolidation of regional ATM networks."[13] That is, ATM systems are more efficient when in larger networks.

Some see benefits from larger networks, others see threats to monopolization by a dominant network, including dominant software. Hence the issue of network effects has arisen in some antitrust suits claiming network effects tended to lead to monopolies, a view rejected in most decisions.[14]

Externality has appeared in a few other antitrust cases. One concerned alleged monopolization of the broadcast of scores from golf tournaments. Plaintiff argued that release of the scores was a "positive externality" that benefited all who wanted to know the scores. The court agreed, but held that the Professional Golf Association has a valid business reason for restricting access to proprietary score information.[15] In a suit involving tortilla makers paying for shelf space, the court discussed "market realities and other externalities," seeming to mean external effects in general.[16] In another case, the court rejected a monopolization claim made against a college by a fraternity. The court noted that efforts by a college to enroll the "best" students make a school more attractive, which is a "positive externality."[17] In sum, the use of

[13] *Money Station, Inc. v. Board of Governors of the Federal Reserve System*, 81 F.3d 1128, 1133 (D.C. Cir. 1996).

[14] For specific references to network externalities see *U.S. v. Microsoft Corp.*, 147 F.3d 935, (D.C. Cir. 1998); *California Dental Assn. v. F.T.C.*, 224 F.3d 942 (9th Cir. 2000); *U.S. v. Microsoft Corp.*, 253 F.3d 34 (D.C. Cir., 2001); *Freeman v. San Diego Association of Realtors*, 321 F.3d 1133 (9th Cir. 2003) where Judge Kozinski noted, at 1153 fn. 28, that those who worry about the anti-competitive consequences of network externalities have made it the market "*defect du jour;*" *Broadcom Corp. v. Qualcomm, Inc.*, 501 F.3d 297 (3rd Cir. 2007); *Novell, Inc. v. Microsoft Corp.*, 505 F.3d (4th Cir. 2007); *Sprint Nextel Corp. v. AT&T, Inc.*, 821 F.Supp.2d 308 (2011).

[15] *Morris Communications Corp. v. PGA Tour, Inc.*, 235 F.Supp.2d 1269, 1280 (M.D. Fla. 2002).

[16] *El Aguila Food Products, Inc. v. Gruma Corp.*, 301 F.Supp.2d 612 (S.D. Tex. 2003).

[17] *Delta Kappa Epsilon (DKE) Alumni Corp. v. Colgate University*, 492 F.Supp.2d 106, 110 (N.D. N.Y. 2007), citing the expert testimony of Prof. Jerry A. Hausman from the economics department at MIT.

externality in reported antitrust cases is consistent with common economic usage. A decision is made by a party that has a cost impact on other parties.

5.4 Bankruptcy

In two cases externality is used in the sense from Justice Holmes, referring to external meanings of words and the law.[18] Another case has a nearly incomprehensible discussion of "externality-creating religious conduct" and "free exercise externalities" that appear to mean external factors in determining if a religion qualifies as a valid charity.[19] It is not an economic use of the term.

Other cases refer to events outside (external) to a business that impacted the business.[20] That is, bad economic conditions, which are external to the firm, may help drive it into bankruptcy. One case clearly attempts to use externality in an economic sense. It discusses the problem of debtors or creditors trying to shift costs to other parties.[21] While it is true that can be a costly exercise, it is not a proper economic use of the term. Such cost shifting, if legal, would not qualify as an externality in its economic sense because it would be a cost that should or should not be borne by a party to the transaction. It is presumed in economics that one role of law is for judges to resolve liability disputes. Hence most bankruptcy cases do not use externality in an economic sense but in a more general sense of events outside the firm that impacted it.

5.5 Constitutional Law

Externality appears in cases that deal primarily with a constitutional issue. In most, the intent was to use the term in its economic sense. The first case involved the striking down of an Illinois statute intended to force coal-fired utilities to use more Illinois coal.[22] The State made the argument that the use of Illinois coal was an economic benefit to Illinois because it forced expenditures inside the state and that the state would suffer an externality (economic loss) if utilities were allowed to buy non-Illinois coal. Striking down the statute as a violation of the Commerce Clause, the

[18] *Matter of Sinclair*, 870 F.2d 1340 (7th Cir. 1989) and *In re Walsh*, 260 B.R. 142 (Bkrtcy. D. Minn. 2001).

[19] *In re Saunders*, 215 B.R. 800, 804 (Bkrtcy. D. Mass. 1997). The usage seems to mean that certain religious activities can impose costs on others outside of the religion.

[20] For example, *In re 523 East Fifth Street Housing Preservation Development Fund Corp.*, 79 B.R. 568 (Bkrtcy. D.D. N.Y. 1987).

[21] *In re An-Tze Cheng*, 308 B.R. 448 (9th Cir. BAP 2004).

[22] *Alliance for Clean Coal v. Miller*, 44 F.3d 591 (7th Cir. 1995).

concurring opinion noted that the external damage suffered by Illinois coal producers was offset by the external benefits of being part of a free-trade nation.[23] The fact that the law does not recognize the kind of externality costs argued by Illinois is consistent with the economic concept of pecuniary externalities not being relevant. That is, if a Caribou Coffee opens a store across the street from a Starbucks, thereby causing the Starbucks to lose revenue, the Starbucks has incurred a real cost imposed upon it by the decision of Caribou. The fact that income shifts due to changes in competitive conditions is beneficial to society as a whole. The economic view of pecuniary externalities, like the rule of law explained in the opinion, is that such costs are irrelevant or will be ignored.

Some cases used externality in its economic sense as a technical externality, where a cost is involuntarily imposed that may be worthy of legal consideration so as to improve market efficiency. In one case, the court explained that health benefits for coal workers were an attempt by Congress to force the employers of miners to include health costs in the production equation.[24] This is a common economic view, that there is some sort of defect or negative externality in the market (miners are underpaid) that can be addressed by legislative action. Similarly, plaintiffs in another case argued that certain expenditures on public health programs are justified as producing a positive externality for society at large by reducing the spread of disease and reducing the cost of treatment.[25]

Other cases were challenges to statutes in New Jersey and Pennsylvania that required candidates for local office or public-sector employees to have resided in a specific local jurisdiction for at least one year before eligible for office or for employment. In the New Jersey case the court held that the election-qualification statute imposed an unconstitutional burden on the right to intrastate travel.[26] It noted that the state argued in favor of the statute as a way to reduce externalities. If office holders were required to live in the jurisdiction in which they wished to hold office, then they would incur the impact of taxes imposed.[27] The subsequent case upheld the right of jurisdictions to impose residency requirements on prospective employees; it cited the previous case discussion of supposed externalities as being in favor of such a rule.[28]

Other cases apply externality in what is intended to be a conventional economic sense, contending that certain restrictions have the benefit of reducing externalities

[23] *Id.* at 598.

[24] *Unity Real Estate Co. v. Hudson*, 889 F.Supp. 818 (W.D. Pa. 1995).

[25] *Reynolds v. Wagner*, 128 F.3d 166 (3rd Cir. 1997).

[26] *Callaway v. Samson*, 193 F.Supp.2d 783 (D. N.J. 2002).

[27] *Id.* at 788. This is not a proper economic application of the notion. Public servants need not be taxpayers of specific jurisdictions any more than company employees must consume products made by their employer for them to be able to provide good value in employment. The argument is political, not economic.

[28] *McCool v. City of Philadelphia*, 494 F.Supp.2d 307 (E.D. Pa. 2007).

suffered by other due to the actions of certain parties. In one case, concerning regulations that restrict placement of the Confederate flag, a dissenting judge argued that such time, place, and manner restrictions deal with "a negative externality of what otherwise may be protected symbolic speech."[29] That is, seeing the flag imposes a cost on some observers, so restrictions would reduce such costs. The judge ignored the similar costs borne by those who are denied the possibility of seeing the flag in a particular location.

More importantly, it is generally presumed that constitutional rights are not distributed based on a weighing of the costs and benefits of certain parties holding certain rights in contrast to other parties. Markets will function under whatever set of rights have been established. In another case, a First Amendment claim against restrictions on sexually-oriented businesses was rejected. Such businesses impose "negative externalities" on neighbors, so limitations are permissible.[30] Similarly, certain Rhode Island liquor regulations were upheld. The court cited a law review article to the effect that states may impose laws to correct for problems, including externalities.[31]

The term was raised in cases challenging the constitutionality of the Affordable Care Act (ACA) (aka Obamacare). One dissenting judge, arguing for the constitutionality of a challenged provision of the Act, said that it corrected "a massive market failure caused by tremendous negative externalities" (i.e., the lack of health insurance for some, which justified a policy that requires obtaining insurance).[32] In another case that found a portion of the ACA to be unconstitutional, the appeals court noted that "under the government's theory, Congress can enlarge its own powers under the Commerce Clause by legislating a market externality into existence, and then claiming an extra-constitutional fix is required."[33]

In sum, in most cases judges intend to use externalities in its economic sense of a cost visited upon other parties. The courts simply recognized the fact of costs being involved. Costs were never quantified, just recognized as a fact.

---

[29] *Sons of Confederate Veterans, Inc. v. Commissioner of the Virginia Dept. of Motor Vehicles*, 305 F.3d 241, 252 (4th Cir. 2002).

[30] *Center for Fair Public Policy v. Maricopa County, Ariz.*, 336 F.3d 1153, 1162 (9th Cir. 2003). Other courts have ruled similarly, holding that regulation of "adult" business, such as strip clubs, is a permissible activity for the state as it protects parties from external effects created by the presence of such businesses; see *Entertainment Productions, Inc. v. Shelby County, Tenn.*, 588 F.3d 372 (6th Cir., 2009).

[31] *Wine and Spirits Retailers, Inc. v. Rhode Island*, 418 F.3d 36 (1st Cir. 2005).

[32] *Liberty University, Inc. v. Geithner*, 671 F.3d 391 (4th Cir. 2011). The discussion by the judge seems misplaced; if people do not buy insurance it may not be due to a market failure but due to low income or lack of interest in the product.

[33] *Florida v. U.S. Dept. of Health and Human Services*, 648 F.3d 1235 (11th Cir. 2011) at fn. 101. That is, the court said Congress declared the fact that some persons do not have health insurance to be a market defect that gives it the constitutional basis for correcting a market defect. No doubt the supporters of the ACA see it that way.

5.6 Contracts

Six contract cases have used the e-word. Five cases use the word in a Holmes sense in reference to external events.[34] For example, concerning payment of property insurance benefits, a dissenting judge argued that the collateral source rule should apply in contract cases to avoid allowing parties to double collect in some instances, in which case they can profit from "externalities," an unfortunate external event such as, in this case, a hurricane.[35]

The sixth case, a dealership dispute, noted: "Not every conferral of a benefit creates an implied contract (consider gifts, third-party beneficiaries, parent-child relationships and the vast array of indirect benefits the economists call positive externalities). Nor do such benefits automatically give rise to claims for unjust enrichment."[36] This is an economically consistent and disciplined use of the concept. The world is full of externalities-costs and benefits we visit upon each other-but most are not relevant legally or we would all live in courtrooms.

5.7 Criminal Law

Criminal law cases often use externalities to refer to factors outside of the evidence-external events or facts.[37] A few cases use it in an economic context. One refers to "interstate externalities" that may justify federal criminal jurisdiction in what might normally be thought of as a matter for state concern.[38] Judge Posners *Economic Analysis of the Law* is cited for that proposition at that point (Posner 1992). While this is not incorrect usage of the term, it is part of the *carte blanche* problem of externality-it is a concept that can be used to justify nearly any interference since costs are ubiquitous.

Taking the other tack based on the same generic notion of externality was a case in which the court struck down a requirement that a certain party must provide a DNA sample to go into the national DNA database. The government argued that such databases create "positive externalities" as they help reduce crime.[39] The judge noted that there were benefits from such things but that did not justify a violation

[34] *Superior Oil Co. v. Western Slope Gas Co.*, 604 F.2d 1281 (Ct. App. Colo. 1979); *Court Street Steak House, Inc. v. County of Tazewell*, 643 N.E.2d 781 (Sup. Ct. Ill. 1994); *Smelkinson Sysco v. Harrell*, 875 A.2d 188 (Ct. Spec. App. Mary. 2003); *Willard Packaging Company, Inc. v. Javier*, 899 A.2d 940 (Ct. Spec. App. Mary. 2006).

[35] *Citizens Property Ins, Corp. v. Ashe*, 50 So.3d 645, 658 (Fla. App., 1 Dist., 2010).

[36] *Motorsport Engineering, Inc. v. Maserati SPA*, 316 F.3d 26, 31 (1st Cir., 2002).

[37] For example, *People v. Pate*, 310 N.W.2d 883 (Ct. App., Mich. 1981); *Zettlemoyer v. Fulcomer*, 923 F.2d 284 (dissent) (3rd Cir. 1991); *U.S. v. Meyers*, 906 F.Supp. 1494 (D. Wyo. 1995); *U.S. v. Bin Laden*, 132 F.Supp.2d 168 (S.D. N.Y. 2001); *People v. Huston*, 802 N.W.2d 261 (Mich. S.Ct., 2011).

[38] *U.S. v. Lipscomb*, 299 F.3d 303, 332 (5th Cir. 2002).

[39] *U.S. v. Miles*, 228 F.Supp.2d 1130, 1139 (E.D. Cal. 2002.

of Fourth Amendment rights as part of a broader criminal matter. Another case also used externality in a generic cost sense, discussing "negative externalities" that would result from two parents being in prison at the same time away from their children.[40] Again, this is not a market transaction; it is a matter to be determined by the rule of law, not some judicial weighing of costs and benefits that cannot be measured.

5.8 Environmental Law

To economists, the word externality is most commonly linked to environmental problems. Most textbooks use pollution as examples to explain the concept. Polluters do not bear all costs of production when they do not prevent waste from being strewn, thereby imposing costs on other people. Those who bear the cost of pollution subsidize the producer. In case law we find externality used a bit more frequently in this area of law than any other area, but they are only about 13 % of the cases. Most of the cases are under various statutes, but some are common law; we first cover National Environmental Policy Act (NEPA) cases, then air, land and water.

5.8.1 NEPA

Two cases involved a claim that an agency violated the National Environmental Policy Act. The first asserted that the ICC had failed to require a ferry service provider from Long Island to provide an environmental impact statement as part of rate deregulation.[41] The court held that under NEPA, Congress requires agencies to include environmental considerations, including "the full range of possible externalities, including environmental costs and benefits," but the court asserted NEPA did not apply to a decision to exempt ferry service from ICC rate regulation.[42]

A later case challenged the adequacy of an Environmental Impact Statement prepared by the Army Corps in its decision to dredge the shipping channel in the Columbia River. The appeals court affirmed summary judgment for the government, dismissing the plaintiff's claim that "the agency understated the costs associated with the project by failing to consider environmental externalities associated with channel deepening…."[43] The court held that the Corps satisfied NEPA by conducting

[40] *Smith v. U.S.*, 277 F.Supp.2d 100, 114 (D. D.C. 2003).

[41] *Cross-Sound Ferry Services, Inc. v. Interstate Commerce Comm.*, 934 F.3d 327 (D.C. Cir. 1991).

[42] *Id.* at 333; Judge Clarence Thomas, in a concurring opinion, said the Transportation Act of 1940 "did not mean to give the ICC power to regulate ferries in order to promote ecological consciousness-raisingor any other 'externalities' unconnected to [the narrow focus on transportation]." At 338.

[43] *Northwest Environmental Advocates v. National Marine Fisheries Service*, 460 F.3d 1125, 1147 (9th Cir. 2006).

extensive economic and environmental analyses. So both NEPA cases use externality in the environmental economic sense.

5.8.2 Air Pollution

The first federal case to use the term externality for environmental damage was in 1980. The case was one in a 20-year marathon involving aluminum smelters in Oregon along the Columbia River.[44] The court awarded compensatory and punitive damages to an orchard owner who suffered crop damage from fluoride emissions from a smelter. "Our society has not demanded that such externalized costs of production be completely eliminated. Instead, we tolerate externalities such as pollution as long as the enterprise remains productive: that is, producing greater value than the total of its internalized and externalized costs of production."[45]

The other air cases that mention externality were statute based. For example, when a challenge arose to the issuance of a permit for a new coal-fired electric generation plant in Alaska, the state high court, upholding the Alaska Public Utility Commission (APUC) decision, noted that the definition of externalities was unclear.[46] "The Federation [the group opposing the permit] uses the term 'environmental externalities' interchangeably with the terms 'environmental impacts' and 'environmental costs.' GVEA [the permit holder] argues that 'environmental externalities' are not the equivalent of 'environmental costs' and 'impacts.' It contends that environmental externalities encompass only environmental impacts that are 'not internalized elsewhere in the permitting process.'"[47] It claimed that definition was consistent with NEPA procedure. The court noted that the APUC "defines 'environmental externalities' as those impacts on the environment caused by the production of electricity 'which have not historically been reflected in the costs of electricity.' We find APUC's definition the most accurate."[48] Having approved an expansive definition of the term, the court held that "APUC is not required to consider costs associated with environmental externalities… in its inquiry concerning whether a service is required for the public convenience and necessity" and if the applicant is "fit, willing and able."[49] The dissenters protested on that point, contending that part of being fit to provide service meant being able to provide an explanation of "environmental externalities" that

[44] The first suit was filed in 1961; see *Renken v. Harvey Aluminum, Inc.*, 226 F.Supp. 169 (D. Ore. 1963). It did not use the term externality.

[45] *Orchard View Farms, Inc. v. Martin Marietta Aluminum, Inc.*, 500 F.Supp. 984, 989 (D. Ore. 1980).

[46] *Alaska Federation for Community Self-Reliance v. Alaska Public Utilities Comm.*, 879 P.2d 1015 (Sup. Ct. Alak. 1994).

[47] *Id.* at 1018, fn. 2.

[48] Id. at 1018.

[49] *Id.* at 1022.

APUC should evaluate.[50] Both opinions indicate an understanding of the economic meaning of externality in an environmental context, but there was no discussion of specifically how such costs should be addressed, the issue was to follow procedure properly.

The same year a similar case was decided by the Massachusetts high court. The Department of Public Utilities (DPU) ruled that electric utilities must consider environmental externalities when planning facilities. As the court explained: "The department recognized 'that including environmental externalities in resource selection decisions may result in some higher direct costs in the short-term' but that 'environmental externalities are real costs borne by ratepayers and the rest of society in the form of increased health care expenses, economic impacts on material and agricultural resources, and a reduced quality of life.'"[51] To measure externalities, the DPU looked at "the implied valuation method" as a "proxy" for damages that reflect "what society as a whole is willing to pay to avoid damages from pollutant emissions."[52]

The court held that the DPU could consider a utility's pollution and consider mitigation costs against long-term economic benefits. "Where we disagree with the department (as a matter of legal principle, but not as a matter of environmental policy) is in the department's conclusion that increased costs (and hence higher rates) are justified solely because of the potential or real effect of pollution on other than ratepayers [i.e., society as a whole]…. These are important subjects, but they lie in the jurisdiction of legislatures…."[53] DPU was not allowed to impose expansive economic measures of environmental costs.

Some cases are more mundane in that they use externality to refer generically to problems imposed by air pollution and regulations adopted to deal with the problems, but there were no unusual legal issues concerning how costs might be determined.[54] Two cases were more contentious. In one, the Washington state high court was "asked to decide whether a municipal utility may mitigate the effects of its greenhouse gas emissions by paying public and private entities to reduce those entities' emissions. We hold that combating global warming is a general government purpose, albeit a meritorious one, and not a proprietary utility purpose."[55] The four dissenting justices would have permitted the expenses, asserting that "GHGs and anthropogenic climate change are externalities of electricity generation-they are costs borne from the activity

[50] *Id.* at 1024.

[51] *Massachusetts Electric Co. v. Dept. of Public Utilities*, 643 N.E.2d 1029, 1032 (Sup. Jud. Ct. Mass. 1994).

[52] Id. at 1032.

[53] *Id.* at 1034.

[54] *Alliance for Clean Coal v. Miller*, 44 F.3d 591 (7th Cir. 1995); *U.S. v. Marine Shale Processors*, 81 F.3d 1329 (5th Cir. 1996); *LaFleur v. Whitman*, 300 F.3d 256 (2nd Cir. 2002); *Glustrom v. Colorado Public Utilities Comn.*, 280 P.3d 662 (Colo. S.Ct., 2012); *American Coatings Assn., Inc. v. South Coast Air Quality Dist.*, 278 P.3d 838 (Calif. S.Ct., 2012).

[55] *Okeson v. City of Seattle*, 150 P.3d 556, 558 (Sup. Ct. Wash. 2007). For example, the utility paid for buses and ferries to burn cleaner fuels and paid DuPont $650,000 to buy 300,000 tons of emission offsets from a DuPont plant in Kentucky.

which are not reflected in electricity rates."[56] In another case, states and public interest organizations petitioned for review of the National Highway Traffic Safety Administration (NHTSA) rule regarding corporate average fuel economy (CAFE) standards for light trucks. The Ninth Circuit struck the regulations because the agency was arbitrary and capricious for failing to monetize the benefits of greenhouse gas emissions.[57] The court noted that a committee of the National Academy of Sciences "found 'externalities of about $0.30/gal of gasoline associated with the combined impacts of fuel consumption on greenhouse gas emissions and on world oil market conditions.'"[58] NHTSA's contended that this number was too speculative, unlike other externality costs associated with driving, such as emissions from gasoline refining and noise from driving. The agency claimed that the "value of reducing emissions of CO2 and other greenhouse gases [is] too uncertain to support their explicit valuation and inclusion among the saving in environmental externalities from reducing gasoline production and use."[59] The court rejected that position, thereby highlighting a key issue in the economic concept of policies that can be justified based on the notion of externality.

In sum, the economic concept of externality appeared not to be central to most decisions that recognize the authority of regulators, under statutes, to impose regulations. In the Alaska and Massachusetts cases the economic notion of external costs played a large role as the courts discussed which costs could be "internalized" in the regulatory process. The courts stuck with the legislatively defined parameters. In the last two cases noted, the economic cost concept of externalities played a large role, with the dissenters in the Washington state case arguing that the costs provided justification for greenhouse gas emission side payments and with the Ninth Circuit stating that NHTSA must monetize externalities in its fuel economy standards. These positions come close to what is often called judicial activism, but one was a minority view and the other is a view commonly adopted in air pollution regulatory matters in recent years-monetizing non-market environmental costs.

5.8.3 Land Pollution

Of the eleven land pollution cases that use the term externality, four were common law actions. The first was a nuisance suit against a feedlot; the court held that its ruling for plaintiffs was based "not simply upon general notions of fairness; it is also grounded in economics… the problems of 'externalities'…"[60] The two cases

[56] *Id.* at 566. The dissenters also argued that the "program internalizes the externalities associated with electricity generation in the most efficient manner, thus benefiting the ratepayers."

[57] *Center for Biological Diversity v. National Highway Traffic Safety Admin.*, 508 F.3rd 508 (9th Cir. 2007).

[58] *Id.* at 517.

[59] *Id.* at 524.

[60] *Carpenter v. Double R Cattle Co.*, 669 P.2d 643, 653 (Ct. App. Idaho 1983). "Externalities distort the price signals essential to the proper functioning of the market." The court cited Coase (1960),

were for contamination of property that spilled over to neighboring property; one was a nuisance action; the later one billed as a toxic tort matter.[61] The last case was a successful nuisance action against a hog operation.[62]

Most of the other land contamination cases, in which externality is mentioned, were scraps about remediation costs under the Comprehensive Environmental Response, Compensation and Liability Act (CERCLA; aka Superfund). The discussion of environmental externalities was perfunctory and the economic issues did not appear to be relevant to the decisions.[63] The other two cases were similar clean up cases, one under a Michigan statute, the other under the Energy Policy Act that requires uranium purchasers to contribute to remediation; the term externality was used in reference to costs spilling over on other parties.[64] In sum, of the eleven cases, externality was always used in reference to pollution costs spilling over on other parties. The discussion was brief except in the first case mentioned here, where the economic argument played a role in the Idaho courts adoption of the *Restatement* rule regarding nuisances that invade neighboring property.

5.8.4 Water Pollution

In one water pollution case, externality refers to external substances that caused contamination,[65] but in the other four cases it is used in the sense of environmental costs being imposed on other parties. In three cases, the mention is brief and it has no impact on the logic of the decision.[66] In another case, involving environmental standards for a mining operation, the court notes that "The externalities produced by

(Footnote 60 continued)
and argued that intervention was required because there are "impediments to changes of property" that might otherwise solve problems such as the one illustrated in this case.

[61] *Philadelphia Electric Co. v. Hercules, Inc.*, 762 F.2d 303 (3rd. Cir. 1985); the court cited, at 314, an article with the word externalities in the title. *Cottle v. Superior Court*, 3 Cal.App.4th 1367 (Ct. App., 2 Dist., Cal. 1992); the court briefly discussed the divergence of private costs and social costs, citing an article on externality at 1402. In neither case did brief discussion of externality appear to have any effect on the outcome.

[62] *Tetzlaff v. Camp*, 715 N.W.2d 256 (Sup. Ct. Iowa 2006). The court noted at 261 that the property owner had to know of the externalities that flowed from the hog operation.

[63] *Lincoln v. Republic Ecology Corp.*, 765 F.Supp. 633 (C.D. Cal. 1991); *Transportation Leasing Co. v. State of California*, 861 F.Supp. 931 (C.D. Cal. 1993); *Louisiana Pacific Corp. v. Beazer Materials & Services, Inc*., 842 F.Supp. 1243 (E.D. Cal. 1994); *Westfarm Associates LP. v. Washington Suburban Sanitary Comm.*, 66 F.3d 669 (4th Cir. 1995); *Acushnet Co. v. Coaters, Inc.*, 948 F.Supp. 128 (D. Mass. 1996).

[64] *Aetna Casualty & Surety Co. v. Dow Chemical Co.*, 28 F.Supp.2d 448 (E.D. Mich. 1998); *PSI Energy, Inc. v. U.S.*, 59 Fed.Cl. 590 (Fed. Cl. 2004).

[65] *U.S. v. Massachusetts Water Resources Authority*, 97 F.Supp.2d 155 (D. Mass. 2000).

[66] *McGowan v. Mississippi State Oil & Gas Bd.*, 604 So.2d 312 (Sup. Ct., Miss. 1992); *In re Water Use Permit Applications*, 9 P.3d 409 (Sup. Ct. Haw. 2000); *U.S. v. Wayne County, Mich.*, 369 F.3d 508 (6th Cir. 2004).

a mining operation including pollution, traffic, and the aesthetic harms created by having a large mining operation nearby-also affect the surrounding community."[67] This statement does not appear to have impacted the decision, it reflects the way environmental costs are commonly discussed.

One case marks the only Supreme Court decision in which externality is used in reference to environmental problems and is posited as a justification for federal intervention.[68] The Court held, 5–4, that abandoned gravel pits planned for use as dumps for non-hazardous solid waste disposal, were not under federal jurisdiction under the Clean Water Act because water in the pits was not navigable. In dissent, Stevens contended that migratory birds used the pits, which should be sufficient to create national issues. "In such situations, described by economists as involving 'externalities,' federal regulation is both appropriate and necessary."[69]

Stevens then cited an article (Revesz 1992) in support of this point, but mischaracterized its argument on this very point. What Revesz argued can be summarized by this line from the article: "the race-to-the-bottom hypothesis, though influential, lacks a sound theoretical basis." (Revesz 1992, p. 1244). That is, externality is a common argument made in favor of federal regulation, but that presumption must be taken carefully; competition among states tends to provide many protections for the environment. Further, market forces, under a rule of law, also works to resolve many problems. We do not end up in an environmental cesspool in the absence of federal regulations that are alleged to internalize all environmental externalities.

In sum, of the water cases that mention externality; only Stevens in dissent expressly found it to be an argument in favor of expanded regulatory control. The other cases used the term in its common meaning of an external cost imposed on others, now generally subject to regulatory controls.

5.9 FCC, ICC, and PUC Cases

Thirteen cases concern regulatory matters determined under Federal Communication Commission (FCC), Interstate Commerce Commission (ICC), or state Public Utility Commission (PUC) rules.[70] In most cases the issue was if an external cost could be internalized-that is, brought into the rate that could be charged by the regulated utility. In some cases the courts held that the cost matters were external to what they were allowed to consider. In other cases the courts held that firms could not be exempt from rate regulations or that certain costs that had previously been held to

[67] *Hydro Resources, Inc. v. U.S. E.P.A.*, 608 F.3d 1131 (10th Cir., 2010).

[68] *Solid Waste Agency of Northern Cook Co. v. U.S. Army Corps of Engineers*, 531 U.S. 159, 121 S.Ct. 675 (2001).

[69] *Id.* as 195, 695.

[70] Public Utility Commissions have different names in different states; the term is generic here.

be external could be brought into the rate base.[71] This use of the term externality is economic but is a matter of statutory interpretation of what costs are costs legally. It is not a matter about which economic analysis has much to offer. In passing, a court mentioned network externalities.[72]

Two cases were claims by environmental groups that PUCs must take into account environmental externalities not considered in granting a permit to build a new generating plant[73] or not considered when electricity was purchased from out-of-state.[74] This would be the kind of external cost often discussed in economics. Environmental concerns have real value. But in both cases the courts remained within the statutory definitions of costs to be considered. The objections raised by outside groups were not, by law, to be considered, so were not relevant.

This did not mean the courts did not comprehend that real costs may be at issue. They treated environmental costs like any other cost petitioned for review. Under regulatory schemes, some costs are counted, some are not. Since the law is used to clarify the treatment of such costs, it allows parties to adjust to the rule. Those unhappy with the allocation of costs can appeal to the legislature for a revision of the law or can deal with the utility accused of polluting by offering payments to change or cease certain activities. The use of externality in all cases was straightforward and correct.

5.10 Intellectual Property

Most of the intellectual property cases are copyright cases that use externality use it in a non-economic sense. The first case in this area to use the term referred to "the externalities of the cotton market" referring to facts that cannot be captured in copyright.[75] The court cited *Plains Cotton* several years later, noting that *scenes a faire*, common knowledge or situations, such as the idea of a secret Swiss bank account used in a spy novel, cannot be captured by a copyright. That was one of

[71] *Brae Corp. v. U.S.*, 740 F.2d 1023, 1057 (D.C. Cir. 1984); *Colorado Office of Consumer Counsel v. Public Utilities Comn.*, 786 P.2d 1086 (Sup. Ct. Colo. 1990); *CF&I Steel, L.P. v. Public Utilities Comm.*, 949 P.2d 577 (Sup. Ct. Colo. 1997); *GTE Southwest Inc. v. Public Utility Comm.*, 10 S.W.3d 7 (Ct. App. Tex. 1999); *New York State Electric & Gas Corp. v. Public Service Comm.*, 753 N.Y.S.2d 332 (Sup. Ct. N.Y. 2002); *Commonwealth Edison Co. v. Illinois Commerce Comn.*, 937 N.E.2d 685 (Ill. App., 2nd Dist., 2011); *People ex rel. Madigan v. Illinois Commerce Comn.*, 958 N.E.2d 405 (Ill. App., 1st Dist., 2011).

[72] *Rural Cellular Assn. v. F.C.C.*, 685 F.3d 1083 (D.C. Cir. 2012).

[73] *Texas Utilities Elec. Co. v. Public Citizen, Inc.*, 897 S.W.2d 443 (Ct. App. Tex. 1995).

[74] *In re Northern States Power Co.*, 676 N.W.2d 326 (Ct. App. Minn. 2004).

[75] *Plains Cotton Cooperative Assn. of Lubbock, Texas v. Goodpasture Computer Service, Inc.*, 807 F.2d 1256, 1262 (5th Cir. 1987).

several decisions that cite the word externality from Plains Cotton and use it in the same non-economic context.[76]

Other cases also used a non-economic meaning of externality. In one the court referred to the limited capacity of a computer as an externality a programmer would take into account.[77] Another case used what it called the "externalities doctrine" concerning *scenes a faire*.[78] In a patent case concerning an "externally mounted intercooler" the term externality was referring to the device.[79] The only copyright case in which externality was used in an economic sense was one that referred to network externalities-as a reason why copyright protection should not be too inclusive.[80] Hence, most intellectual property cases do not use externality in an economic meaning.

5.11 Labor Law

Most labor cases, which cover a wide variety of labor issues, use externality in a legal sense only. Whether ERISA benefits, workers' compensation benefits, a dispute over Davis-Bacon wages, collective bargaining agreements or Title VII application, the cases use externality in reference to external conditions of employment or matters external to the law itself.[81] There is an economic-type usage in a case concerning a

[76] *Engineering Dynamics, Inc. v. Structural Software, Inc.*, 26 F.3d 1335 (5th Cir. 1994). See also *Autoskill, Inc. v. National Educational Support Systems Inc.*, 793 F.Supp. 1557 (D. N.M. 1992); *Kepner-Tregoe, Inc. v. Leadership Software, Inc.*, 12 F.3d 527 (5th Cir. 1994); *Mitel, Inc. v. Iqtel, Inc.*, 124 F.3d 1366 (10th Cir. 1997); *Torah Soft Ltd. v. Drosnin*, 136 F.Supp.2d 276 (S.D. N.Y. 2001). The *Mitel* case was cited, in the same externality context, in *Dun & Bradstreet Software Services, Inc. v. Grace Consulting, Inc.*, 307 F.3d 197 (10th Cir. 2002).

[77] *Computer Associates International, Inc. v. Altai, Inc.*, 982 F.2d 693 (2nd Cir. 1992). Similarly, another court referred to a network externality caused by too much demand on computer user bases; *Free FreeHand Corp. v. Adobe Systems Inc.*, 852 F.Supp.2d 1171 (N.D. Cal. 2012).

[78] *Control Data Systems, Inc. v. Infoware, Inc.*, 903 F.Supp. 1316, 1323 (D. Minn. 1995).

[79] *Rice v. U.S.*, 84 Fed.Cl. 575 (Ct. Fed. Cl. 2008).

[80] *Apple Computer, Inc. v. Microsoft Corp.*, 799 F.Supp. 1006 (N.D. Cal. 1992). Another case mentioned network externality in referring to a publication with the term in the title; *DocMagic, Inc. v. Ellie Mae, Inc.*, 745 F.Supp.2d 1119 (N.D. Cal. 2010).

[81] *DeArmond v. Sommer*, 348 N.E.2d 378 (Ct. App. Ohio 1975); *Dickerson v. U.S. Steel Corp.*, 472 F.Supp. 1304 (D.C. Pa. 1978); *Martin v. Sullivan*, 932 F.2d 1273 (9th Cir. 1991); *Westinghouse Hanford Co. v. Hanford Atomic Metal Trades Council*, 940 F.2d 513 (9th Cir. 1991); *Martin Marietta Corp. v. Lorenz*, 823 P.2d 100 (Sup. Ct. Colo. 1992); *Davis v. Portline Transportes Maritime Intl.*, 16 F.3d 532 (3rd Cir. 1994); *Northern California Drywall Contractors Assn. v. Dist. Council of Painters No. 8*, 879 F.Supp. 96 (N.D. Cal. 1995); *Carollo v. Cement and Concrete Workers Dist. Council Pension Plan*, 964 F.Supp. 677 (E.D. N.Y. 1997); *Collette v. St. Luke's Roosevelt Hospital*, 132 F.Supp.2d 256 (S.D. N.Y. 2001); *Melvin v. US Local 13 Pension Plan*, 202 F.Supp.2d 564 (W.D. N.Y. 2002); *City of Long Beach v. Dept. of Industrial Relations*, 1 Cal.Rptr.3d 837 (Ct. App. 2 Dist., Cal. 2003); *Keenan v. Director for Benefits Review Board*, 392 F.3d 1041 (9th Cir. 2004); *Committee of Concerned Midwest Flight Attendants for Fair and Equitable Seniority Integration v. International Broth. of Teamsters Airline Div.*, 662 F.3d 954 (7th Cir. 2011).

claim by public employees of violation of First Amendment rights.[82] The decision noted that the unconstitutional conditions doctrine can be justified as dealing with market failures, including negative externalities. This includes the problems that arise from having many government sector jobs subject to patronage.[83] In another case the court mentioned the "positive externality" that may arise from health insurance coverage that may obviate the need for worker's compensation.[84] That is, only three of sixteen cases used externality in any economic sense.

5.12 Takings

Takings cases are generally constitutional law cases but there were enough to put them in a separate category for our purposes. In the first takings case to use the term, a California court affirmed the constitutionality of the Coastal Conservation Act of 1972 that created the California Coastal Commission. Plaintiffs contended the law allowed the Commission to impose costs (externalities) on them. The court held that "it can be safely said that where [an] activity, whether municipal or private, is one which can affect persons… the state is empowered to 'prohibit or regulate the externalities.'"[85] This is an elastic view of externalities that could seemingly be used to justify most government actions because externalities are pervasive; most uses of the concept are not so expansive.

The next case was the first Supreme Court case in which the term externalities appeared. A cable television company attached a cable box to the exterior of rental property. That was done under a New York statute that allowed the cable attachment for payment of $1. The landlord sued, contending the cable company, with state backing, was engaged in a taking. The landlord demanded the cable company negotiate with her rather than take her property by attaching a box to the exterior of the building under a permit from the state. The New York courts upheld the New York law and the cable company action. Writing for the majority, Justice Marshall held the placement of the cable box to be an unconstitutional taking.[86]

In dissent, Justice Blackman, joined by Brennan and White, argued that a physical action should not be the basis for the definition of a taking. A small box attached to the exterior of a building, with state permit, was less invasive to the value of property than many regulations that restrict property usage but are not physically invasive. "Modern government regulation exudes intangible 'externalities' that may diminish

[82] *McCloud v. Testa*, 97 F.3d 1536 (6th Cir. 1996).

[83] *Id.* at 1551, citing Epstein (1987).

[84] *Joseph M. Still Burn Centers, Inc. v. AmFed Nat. Ins. Co.*, 702 F.Supp.2d 1371 (S.D., Ga. 2010). Similarly, in another worker's compensation case the court noted that payment from another source was an externality (a benefit to the worker); *In re Wadsworth's Case*, 935 NE.2d 333 (Mass. App. Ct. 2010).

[85] *Creed v. California Coastal Zone Conservation Comm.*, 118 Cal.Rptr. 315, 321 (Ct. App., 4th Dist., Cal., 1974), citing Sato (1972).

[86] *Loretto v. Teleprompter Manhattan CATV Corp.*, 458 U.S. 419, 102 S.Ct. 3164 (1982).

the value of private property far more than minor physical touchings."[87] The use of externality appears to be in the sense of government actions, such as zoning laws, that impact property value.

The issue next arose in a case in which a city denied a gas station owner permission to add a convenience store to his station. The property owner protested that this was a taking and the district court agreed: "the City may constitutionally 'tax' plaintiff to recoup the costs of the negative externalities that its increased business activities cause: Without a showing of such externalities, the condition which the City attached to building permits is simple extortion."[88] This use of externality is consistent with the economic use meaning costs being imposed involuntarily on others. The gas station would increase traffic that would impact neighbors.

The next case involved a federal government taking, by legislative action, of 550 acres of land next to the Manassas Battlefield Park that the owner planned to convert to housing and retail development. Before the development occurred, the government paid the owner for the property, but the county that expected economic benefits from the development sued for an uncompensated taking. The appeals court rejected the claim.[89] In its analysis, it cited economic analysis about externalities. The court noted, from the perspective of economic analysis, that there cannot be an externality if a right does not exist; that is, the county had no right to the future benefits it hoped to accrue from the planned development.[90] The county can claim it suffered an externality, but since it had no legal right to a gain that never came into existence, it was not actionable. This is much like court reasoning in many zoning cases.

Other cases mention externalities briefly, referring to some activity not liked by some people, or liked by some people.[91] The use of the term is in its normal economic meaning of spillovers incurred by some due to the actions of others. One case refers to externality in the legal sense of external events that courts should or should not take into account in triggering certain rights.[92] Hence, in most taking cases the discussion of externality was economic in nature. Some rather vague, others quite precise.

---

[87] *Id.* at 447, 3182.

[88] *William J. (Jack) Jones Insurance Trust v. City of Fort Smith, Ark.*, 731 F.Supp. 912 (W.D. Ark. 1990).

[89] *Board of County Supervisors of Prince William County, Virginia v. U.S.*, 48 F.3d 520 (Fed. Cir. 1995).

[90] *Id.* at 525; the court cited an original economics paper on the role of transaction costs, externality, and property rights citepdemsetz1967.

[91] *Melillo v. City of New Haven*, 732 A.2d 133 (Sup. Ct. Conn. 1999), the court rejected that a takings occurred due to airport noise, which the plaintiff claimed to be externalities; *District Intown Properties LP v. Dist. Columbia*, 198 F.3d 874 (D.C. Cir. 1999), the concurring opinion asserted that historic preservation laws resulted in positive externalities by protecting old buildings that some people enjoy; *R&Y, Inc. v. Municipality of Anchorage*, 34 P.3d 289 (Sup. Ct. Alaska 2001), holding that a construction setback requirement was not a taking. Such land-use restrictions impose externalities no worse than the traffic suffered from increased commercial activity; it is non-actionable.

[92] *Cashman v. City of Cotati*, 374 F.3d 887 (9th Cir. 2004).

5.13 Taxes

Externalities as used in tax often refer to activities outside of a business-external factors that happen to affect business value-including the impact of weather on crops and changes in market conditions that impact values.[93] In two cases the courts refer to positive externalities that might be taken into consideration.[94] In an explicit use of the economic notion, three federal appeals court cases the courts noted that a tax may be imposed on those who create negative externalities to compensate those who suffer the costs.[95] In some cases the use was so imprecise as to be meaningless, although it appears refer to externalities in the economic sense.[96] In sum, most tax cases externality is intended to be economically meaningful and, in some, was used as partial justification for a tax.

5.14 Torts

The tort cases that refer to externality provide a good vehicle to explain the problem with the concept applied to law. What is an externality-and when it should be

[93] *International-Stanley Corp. v. Dept. of Revenue*, 352 N.E.2d 272 (Ct. App. Ill. 1976); *Nunes Turfgrass, Inc. v. County of Kern*, 111 Cal.App.3d 855 (Ct. App. Cal. 1980); *Michigan Assn. of Counties v. Dept. of Management and Budget*, 345 N.W.2d 584 (Sup. Ct. Mich. 1984); *Chevron U.S.A., Inc. v. City of Perth Amboy*, 10 N.J.Tax 114 (Tax Ct. N.J. 1988); *Lampy Ready Mix, Inc. v. County of Otter Tail*, 1991 WL 44882 at 4 (Tax Ct. Minn. 1991); *Vermont Soc. of Assn. Executives v. Milne*, 779 A.2d 20 (Sup. Ct. Vt. 2001).

[94] *IHC Health Plans, Inc. v. Commissioner of Internal Revenue*, 325 F.3d 1188 (10th Cir. 2003), referring to the "positive externalities" generated by certain public goods; *PSI Energy, Inc. v. U.S.*, 59 Fed.Cl. 590 (Fed. Cl. 2004), referring to the "positive externality" received by free riders who get public benefits without paying for the benefits.

[95] In one case the court euphemistically called the taxes a "compensation charge" or "user fee" where property owners were assessed a fee when they demolished a residential building. The fee could be avoided if the property owner constructed "affordable" housing. The fee was to compensate other city residents for the impact of the demolition. *Kathrein v. City of Evanston, Ill.*, 636 F.3d 906 (7th Cir 2011); then cited in *Empress Casino Joliet Corp. v. Balmoral Racing Club, Inc.*, 651 F.3d 722 (7th Cir. 2011). In this case the court was not ruling on the constitutionality of the taxes but on the right to contest the taxes. In *Rincon Band of Luiseno Mission Indians of Rincon Reservation v. Schwarzenegger*, 602 F.3d 1019 (9th Cir. 2010), the court struck down a state tax on gambling on an Indian reservation but noted that in general gambling taxes may be imposed to help offset the negative externalities caused by gambling.

[96] *Holmdel Builders Assn. v. Township of Holmdel*, 583 A.2d 277 (Sup. Ct. N.J. 1990), referring to "unfettered non-residential development" that has caused externalities; *Burlington Northern Santa Fe Railroad Co. v. The Assiniboine and Sioux Tribes of the Fort Peck Reservation*, 323 F.3d 767 (9th Cir. 2003), referring to "nonmember activities that produce externalities for tribes but do not rise to the level [required for legal action or for compensation to be required]" as the tribes wanted to be paid for costs suffered by trains crossing the reservation; *Carter v. Carolina Tobacco Co.*, 873 N.E.2d 611 (Ct. App. Ind. 2007), regarding externalities asserted to exist, by a consultant to states' attorney generals for the Master Settlement Agreement on tobacco, if any tobacco sellers could avoid being part of the settlement.

actionable-can be in the mind of the beholder; the job of the courts is, regardless of what term is employed, to follow the rules of law. Externality could allow anything and everything to be actionable.

The first tort case to use the term externality was a wrong death suit based on failure to warn for a child killed by a B-B gun. The Pennsylvania high court agreed that the suit should be dismissed.[97] The dissent argued that "the trial court should direct the jury's attention to… physical properties as well as externalities such as marketing, promotional activities, labels, logo" that could affect how a consumer views a product.[98] Advertising materials can be relevant in tort suits, but the externality argument open the barn door to throw in anything that anyone would assert could be possibly relevant. Where should the line be drawn in law? The economic concept of externality as any cost imposed on, or suffered by, anyone gives no clue.

The next case was a claim of tortious interference with business relationships.[99] Affirming the denial of an injunction requested by a party, the court noted that judgments about matters in equity, like many other things, "can be arrived at only through a subjective quantification because of the subjective values, externalities, and effects on the public interest that may be involved in an injunction case."[100] That is, what is an externality is highly subjective. They are real, but hard to pin down and courts are normally wont to engage in such slippery relationships.

Another case concerned a defamation claim brought by an employee against his employing brokerage firm that would proceed to arbitration.[101] Discussing how the actions of some employees at the firm resulted in damage to the reputation of many other employees, the court noted that "any one member's reputation tends to reflect on the others; this externality gives each an interest in the other's standards of conduct."[102] That is, if a member of a firm acts badly and negative press follows, that impacts other employees who are tarnished by the bad actor. This is a real externality, but generally not one the law would recognize as giving rise to a cause of action. There would be no limits to such matters.

One court discussed the desirability of imposing liability on negligent parties who impose costs on innocent parties. A patient contracted AIDS from tainted blood he received during surgery. Allowing certain claims to proceed, the court explained defective products contain hidden costs or "what economists refer to as 'externalities.'"[103] The judge opined that if costs of externalities (defective products) were not imposed on manufacturers, then goods would be too inexpensive and members of society would be encouraged to buy more of the low-price goods, thereby suffering even more damage and imposing more costs on society. Absent strict liability "the

---

[97] *Sherk v. Daisy-Heddon, a Division of Victor Comptometer Corp.*, 450 A.2d 615 (Sup. Ct. Pa. 1982).

[98] *Id.* at 633.

[99] *Lawson Products, Inc. v. Avnet, Inc.*, 782 F.2d 1429 (7th Cir. 1986).

[100] *Id.* at 1434.

[101] *Pearce v. E.F. Hutton Group, Inc.*, 828 F.2d 826 (D.C. Cir. 1987).

[102] *Id.* at 830.

[103] *Doe v. Miles Laboratories, Inc.*, 675 F.Supp. 1466, 1471 (D. Md. 1987).

costs of externalities are thrust upon victims or upon society…."[104] Even if one takes the assertion as true, the judge gives no clue as to where to draw the line. Should the maker of the tainted blood have to compensate everyone who worries about the safety of transfusions?

The next case concerned injuries caused by use of a medical device used on a mother that caused her child, conceived after the medical procedure, to suffer injuries *in utero*. The appeals court affirmed dismissal of the suit, holding the negligent party in the accident had no obligation to an as-yet not conceived child.[105] The dissent quoted (Posner 1972, p. 47): "we want the total liability of negligent injurers to equal the total cost of their accidents."[106] Hence, the dissent argued, "In economic terms, all externalities must be internalized."[107] Liability might be endless in such a formulation.

The court in another decision recognized that problem. Suit was brought for misappropriation of likeness, violation of the false light doctrine, and related claims. The district court granted summary judgment for defendants. The appeals court affirmed.[108] It noted that protecting one's name or likeness "is socially beneficial because it encourages people to develop special skills, which then can be used for commercial advantage."[109] That is, when property rights are secure, there will be greater investment in the development of property. However, there are limits on rights. The claim here reached too far. Protection is not given to general incidents from a person's life or to material in the public record. The court stayed with the rules regarding what specific information is protected by the tort of misappropriation. The fact that the plaintiff felt wronged by an externality, a taking of value from him that went beyond what is protected by law, does not matter legally. The rule of law, not the personal beliefs of an individual as to what is valued, set the bounds of protected interests.

In the next case an insurer, Erie, was sued for denying it had a duty to defend an insured business, Alliance, accused of "advertising injury" inflicted on another company. Advertising injury was covered by the insurance policy. Alliance had been hired by LSC to review the quality of work performed by Sear. Alliance gave LSC a negative report about Sear's work. Sear then sued Alliance for defamation. Erie refused to defend Alliance as the policy did not cover defamation. The trial court and appeals court held for the insurer, saying that the negative information provided to LSC was not advertising.[110] Discussing the matter, the appeals court held: "We

---

[104] *Id.*.

[105] *Hegyes v. Unjian Enterprises, Inc.*, 286 Cal.Rptr. 85 (Ct. App., 2 Dist., Cal. 1991).

[106] *Id.* at 111.

[107] *Id.*.

[108] *Matthews v. Wozencraft*, 15 F.3d 432 (5th Cir. 1994).

[109] *Id.* at 437. Posner (1992) is cited in support.

[110] *Erie Insurance Group v. Sear Corp.*, 102 F.3d 889 (7th Cir. 1996).

refuse to hold that every activity which produces the positive externality of increasing business, especially those activities requisite to basic job performance, constitutes 'advertising' as intended in the [policy here]."[111] That is, Alliance was required to report its findings to LSC as a part of its obligation. The fact that Alliance did a good job, presumably pleasing LSC, created positive spillovers or positive externalities. While good job performance is "good advertising" that may lead to more business, it does not qualify as advertising as it is generally understood. Actionable externalities are kept within the bounds of accepted definitions of actions; individual parties cannot expand the definition to suit their purpose or externalities are ever actionable.

Every loss may be called an externality by the person suffering the loss, but every loss is not actionable. For example, Monroe properly submitted his name to be a teacher for a school district. The district negligently left his name off the list, so he was not considered for employment. He sued but the appeals courts agreed with the trial court that Monroe had no cause of action.[112] "From an economic perspective, traditional common law judges decided that… purely intangible economic risks were matters that should be left as externalities borne by the party that experience them rather than as costs internalized into the social contract of safety."[113] The fact that some employee of the school board carelessly omitted a name did not give rise to a tort of negligence, regardless of the fact that Monroe may indeed have failed to obtain employment he otherwise may have had. No doubt he was aggrieved, but the externality he suffered is not actionable.

One court used the looseness of externality as a justification for allowing a novel cause of action to proceed. The Mayor of Cleveland sued various firearms makers for public nuisance, unreasonably dangerous design and unjust enrichment for making and selling firearms. The trial court rejected the motions to dismiss, holding that suit could proceed.[114] The court reasoned "that the City has paid for what may be called the Defendants' externalities-the costs of the harm caused by Defendants' failure to incorporate safety devices into their handguns and negligent marketing practices."[115] By this logic, automakers could be responsible for the clean up costs of car accidents that cities incur; everything is an externality as one cost is linked to another. As the next case illustrates, this can be so even if one voluntarily assumes the risk.

---

[111] *Id.* at 895.

[112] *Monroe v. Sarasota County School Bd.*, 746 So.2d 530 (Ct. App., 2 Dist., Fla. 1999). This language was quoted later in *Virgilio v. Ryland Group, Inc.*, 695 F.Supp.2d 1276 (M.D. Fla., 2010). The court rejected a claim by homeowners that it was negligent for a developer not to reveal to them that the location of their homes had once been a bombing range used by the military.

[113] *Id.* at 535.

[114] *White v. Smith & Wesson Corp.*, 97 F.Supp.2d 816 (N.D. Ohio 2000). The decision has generally been ignored by other courts.

[115] *Id.* at 829.

A health insurer sued the tobacco companies for causing increased cost of medical services due to smoking. The jury awarded $17 million in compensatory damages.[116] The judge upheld the award: "leading commentators and economists accept the significant 'insurance externalities'-i.e., real world impacts-that result from the presence of first party insurance in consumer markets, and the tobacco market in particular."[117] Providing insurance to smokers, who presumably have revealed the fact of smoking to insurer, who adjusts premiums accordingly, is ordinarily presumed a matter of contract, but by employing the endless cost notion of externality, liability extends.

Similarly, a case involving voluntary sports activity may import liability by notions of externality. A guest at a resort was injured when he fell off a horse provided for a trail ride. The resort provided the novice rider with a gentle horse, but still the accident occurred. The guest had been briefed about the dangers of horseback riding and had signed a liability release, but the appeals court rejected that and allowed suit to proceed.[118] "The effect of the Release is to require society so subsidize Defendants' negligent operation of their business…. Nationwide…'Riding horses may involve greater risk of fatal injury than most other sports'…. There can be no doubt that equine activities expose substantial numbers of consumers to risks of serious physical harm."[119] So externality was cited as a basis for liability not only for costs unwillingly borne that are imposed by others, but for costs suffered in accidents from activities willingly undertaken.

Some tort cases used externality in passive ways not relevant to the outcome, so are not discussed.[120] Most cases discussed here indicate the sloppiness of the meaning of externality. It refers to costs, which are ubiquitous. As some court noted, not all costs are actionable because there is an externality, but some courts seem less sure. If all external costs are actionable, the law has no bounds. Some courts expressly recognize that the externality concept opens the door to unlimited liability. "The theory of using the law to internalize the externality of a business is a well-discussed idea among those who study law and economics. However, it is particularly difficult for a common law court to create a carefully tailored and limited theory of recovery for a special group… without creating more problems than it solves."[121]

---

[116] *Blue Cross and Blue Shield of New Jersey, Inc. v. Philip Morris, Inc.*, 178 F.Supp.2d 198 (E.D. N.Y. 2001).

[117] *Id.* at 235. The decision cited several articles that argued that liability should be imposed on tobacco companies due to externalities.

[118] *Berlangieri v. Running Elk Corp.*, 48 P.3d 70 (Ct. App. N.M. 2002). The decision has not been cited favorably by any other court.

[119] *Id.* at 75. The court cited the *Miles Labs* case, *supra* n. 107 for the proposition that "the costs or externalities are thrust upon victims or upon society," at 1471.

[120] *LaFleur v. Shoney's, Inc.*, 83 S.W.3d 474 (Sup. Ct. Ky. 2002); *Travelers Casualty and Surety Co. v. United States Filter Corp.*, 870 N.E.2d 529 (Ct. App. Ind. 2007); *Bonewitz v. Parker*, 912 N.E.2d 378 (Ind. App. 2009); *Whitehouse v. Target Corp.*, 279 F.R.D. 285 (D. N.J. 2012); *Robbins v. Physicians for Women's Health, LLC*, 38 S.3d 142 (Conn. App. 2012).

[121] *Curd v. Mosaic Fertilizer, LLC*, 993 So.2d 1078 at 1085 (Fla. App. 2nd Dist., 2008).

5.15 Zoning

The first zoning case to use the term externality shows the common use for the term in similar suits.[122] The owner of an adult bookstore sued the state for violating equal protection. The appeals court upheld the states zoning practice: "The North Carolina law regulates adult establishments different from other bookstores and theaters because of the unique external costs of adult enterprises…. Special regulation of one commercial enterprise with particular externalities but not other enterprises lacking those secondary effects has long been recognized not to violate equal protection."[123]

This view has seen the term externality used in a number of cases that followed, all upholding zoning restrictions for dirty book stores, gravel pits, day care facility location, the number of unrelated adults who can live together in a house, restrictions on a homeless encampment built by a church, billboards, and the kind of structures that can be built in a particular area.[124] It is unlikely that the notion of externality made much difference in the law in this area since strong zoning authority goes back many years.[125] However, zoning rules may not infringe on other protected rights.[126]

The Supreme Court strengthened the use of externality in this regard when it supported restrictions on adult businesses and rejected the First Amendment violation they claimed inherent in restrictions. The court held that there were crime patterns correlated with location patterns of adult businesses, so they could be subject to specific rules.[127] Concurring in the decision, Justice Kennedy stated: "The calculus is a familiar one to city planners, for many enterprises other than adult businesses also cause undesirable externalities. Factories, for example, may cause pollution, so a city may seek to reduce the cost of that externality by restricting factories to areas far from

---

[122]Many zoning cases involve a constitutional law claim but there are enough of these to list them as a separate category.

[123]*Hart Book Stores, Inc. v. Edmisten*, 612 F.2d 821, 831 (4th Cir. 1979).

[124]*Basiardanes v. City of Galveston*, 514 F.Supp. 975 (S.D. Tex. 1981); *George Washington Univ. v. Dist. of Columbia Bd. of Zoning*, 429 A.2d 1342 (Ct. App. D.C. 1981); *City of Los Angeles v. State of California*, 138 Cal.App.3d 526 (Ct. App. Cal. 1982); *American Aggregates Corp. v. Highland Township*, 390 N.W.2d 192 (Ct. App. Mich. 1986); *City of Mandan v. Mi-Jon News, Inc.*, 381 N.W.2d 540 (Sup. Ct. N.D. 1986); *Giger v. City of Omaha*, 442 N.W.2d 182 (Sup. Ct. Neb. 1989); *Howard v. City of Garland*, 917 F.2d 898 (5th Cir. 1990); *France Stone Co. v. Charter Township of Monroe*, 802 F.Supp. 90 (E.D. Mich. 1992); *Dvorak v. City of Bloomington*, 702 N.E.2d 1121 (Ct. App. Ind. 1998); *Louhal Properties, Inc. v. Strada*, 751 N.Y.S.2d 810 (Sup. Ct. N.Y. 2002); *Scenic Arizona v. City of Phoenix Bd. of Adjustment*, 268 P.3d 370 (Ariz. App. 2011); *Green v. Douglas County*, 263 P.3d 355 (Or. App. 2011).

[125]E.g., *Hadacheck v. Sebastian*, 239 U.S. 394, 36 S.Ct. 143 (1915) with respect to brickmaking.

[126]The Washington state high court held that zoning rules could not prevent a homeless encampment supported by a church as that would violate freedom of religion, *City of Woodinville v. Northshore United Church of Christ*, 211 P.3d 406 (Wash. Sup. Ct. 2009); a lower court later followed this decision and recognized the externality argument, but upheld the challenge to a homeless encampment as void on procedural grounds. *Mercer Island Citizens for Fair Process v. Tent City 4*, 232 P.3d 1163 (Wash.App. 2010).

[127]*City of Los Angeles v. Alameda Books, Inc.*, 535 U.S. 425, 122 S.Ct. 1728 (2002).

residential neighborhoods."[128] Alameda Books and the discussion about externalities have been cited in a number of challenges to zoning rules. The rules were upheld in every case that includes any discussion of the problem of externalities.[129] In all zoning cases, externalities is used in the standard economic sense of something that spills costs on to others; the term appears to matter little in the substance of the law as it is not used as the rational for expanding the traditional acceptance of the ability of governments to impose zoning regulations.

5.16 Random Cases

Externality pops up in scattered other areas. In some cases the term externality means outside forces, events or information that the court will or will not take into account.[130] In most cases the usage is in the economic sense of something bad or good that spills over to affect others.[131] It was used in passing, not seeming to play

[128] *Id.* at 446, 1740.

[129] *Ben's Bar, Inc. v. Village of Somerset*, 316 F.3d 702 (7th Cir. 2003); *George Washington Univ. v. Dist. of Columbia*, 318 F.3d 203 (D.C. Cir. 2003); *Greenville County v. Kenwood Enterprises, Inc.*, 577 S.E.2d 428 (Sup.Ct. S.C. 2003); *R.V.S., LLC v. City of Rockford*, 361 F.3d 402 (7th Cir. 2004); *Andy's Restaurant & Lounge, Inc. v. City of Gary*, 466 F.3d 550 (7th Cir. 2006); *Ballen v. City of Redmond*, 466 F.3d 736 (9th Cir. 2006); *City of Chicago v. Pooh Bah Enterprises, Inc.*, 865 N.E.2d 133 (Sup. Ct. Ill. 2006); *State v. Stummer*, 171 P.3d 1229 (Ct. App. Ariz. 2007); *Martin Marietta Materials, Inc. v. Board of Zoning Adjustment of Cass County*, 246 S.W.3d 9 (Ct. App. Mo. 2007); *City of Joliet, Ill. V. New West, L.P.*, 562 F.3d 830 (7th Cir. 2009).

[130] *State of Texas v. Sec. of Interior*, 580 F.Supp. 1197 (D.C. Tex. 1984), referring to outside information; *Peters Township School Dist. v. Hartford Accident and Indemnity Co.*, 643 F.Supp. 518 (W.D. Pa. 1986), regarding events beyond insurance coverage; *In re 523 East Fifth Street Housing Preservation Development Fund Corp.*, 79 B.R. 568 (S.D. N.Y. 1987), concerning outside events; *Wint v. Yeutter*, 902 F.2d 76 (D.C. Cir. 1990), referring to things that damage crops; *Northern California Drywall Contractors Assn. v. Dist. Council of Painters No. 8*, 879 F.Supp. 96 (N.D. Cal. 1995); concerning language outside an arbitration agreement; *Demers v. Snyder*, 659 A.2d 495 (Super. Ct. N.J. 1995), referring to outside information that could taint jury deliberations; *Federal Trade Comm. v. QT. Inc.*, 448 F.Supp.2d 908 (N.D. Ill. 2006), referring to external factors that affect pain relief studies.

[131] *Smith v. City of Riverside*, 34 Cal.App.3d 529 (Ct. App. Cal. 1973), referring to events that spill over from one city to another; *In the Matter of the Valuation Proceedings under Sections 303(c) and 306 of the Regional Rail Reorganization Act of 1973*, 445 F.Supp. 994 (Sp. Ct. R.R.R.A. 1977), noting the need to consider externalities and social values in rail reorganization; *South East Lake View Neighbors v. Dept. Housing and Urban Develop*, 685 F.2d 1027 (7th Cir. 1082), referring to bad effects of investment decisions that are non-actionable by third parties; *Kastenbaum v. Michigan State Univ.*, 327 N.W.2d 783 (Sup. Ct. Mich. 1982), regarding positive benefits from the spread of information; *International Union, UPGWA v. Dept. of State Police*, 373 N.W.2d 713 (Sup. Ct. Mich. 1985), concerning positive benefits from the spread of valuable information; *Martin v. Sullivan*, 912 F.2d 1186 (9th Cir. 1990), referring to the bad consequences of a reduction in SSI benefits; *Vieux Carre Property Owners, Residents and Associates, Inc. v. Brown*, 948 F.2d 1436 (5th Cir. 1991), noting damage to historic properties from bad construction; *Yang v. Reno*, 852 F.Supp. 316 (M.D. Pa. 1994), referring to added costs of further review of deportation orders; *F.D.I.C. v. Perry Bros., Inc.*, 854 F.Supp. 1248 (E.D. Tex. 1994), referring to financial setbacks suffered by debtors from

an analytical role in the decisions. In sum, in the scattered areas where externality pops up, its use is not influential.

5.17 Summary of Externality Usage in Case Law

Approximately two-thirds of the cases that mention externality do so in the context of some sort of cost. In some areas, such as antitrust, the use has mostly been technical, with reference to network externalities. In the zoning area, the term is commonly used to refer to certain negative things regulators intend to control, or, in a few cases, positive things being encouraged. The use of the term is non-analytical; it is essentially a generic term meaning bad things that impose costs. That is the sense in which is it used in other scattered areas too. In some instances, the word is not used properly in a technical economic sense, but the readers of the case will understand that the judge is referring to something generally considered to be a bad. As such, it is a vague descriptive term coming into more common usage.

However, in a minority of the cases, perhaps a dozen, externality appears to be a justification for an expansion of costs that will be considered by the court. In some public utility rate cases there was explicit talk of monetizing the externalities from assorted forms of pollution. The argument did not carry the day, but the economic argument is recognized. Similarly, in a couple air pollution cases there was discussion that went beyond the normal vague use of externality to refer to pollution and instead to use it in a stronger sense of an analytical justification for imposition of liability. In one water pollution case, in dissent, Justice Stevens appeared to use externality as a supposedly scientific justification for his view.

The most aggressive use of externality is in tort cases. In some cases scattered over time we see a discussion of externality in an expansive economic sense, justifying a

(Footnote 131 continued)
setoffs; *Escalera v. New York Housing Authority*, 924 F.Supp. 1323 (S.D. N.Y. 1996), regarding drug dealing in housing projects; *Martens v. Smith Barney, Inc.*, 182 F.R.D. 243 (S.D. N.Y. 1998), noting the social benefits from Title VII; *Gardner v. Allstate Indemnity Co.*, 147 F.Supp.2d 1257 (M.D. Ala. 2001), referring to reducing bad effects from harmful acts; *Brooks v. Pre-Paid Legal Services, Inc.*, 153 F.Supp.2d 1299 (M.D. Ala 2001), regarding bad effects from forum shopping; *Fair Share Housing Center, Inc. v. Township of Cherry Hill*, 802 A.2d 512 (Sup. Ct. N.J. 2002), noting costs imposed on some property developers; *Territory of the United States Virgin Islands v. Goldman, Sachs & Co.*, 937 A.2d 760 (Ch. Del. 2007), concerning costs of financial abuses; *Assurance Co. of America v. Lucas Waterproofing Company, Inc.*, explaining that a party may act strategically in litigation to shift legal costs to another party; *Penn Mont Securities v. Frucher*, 534 F.Supp.2d 538 (E.D. Pa., 2008), making a similar point about improper fee shifting in litigation; 581 F.Supp.2d 1201 (S.D. Fla. 2008); *Department of Children and Family Services v. Chapman*, 9 So.3d 676 (Fla. App., 2nd Dist., 2009), noting that external costs of certain activities may be borne by the taxpayers; *Rock River Communications, Inc. v. Universal Music Group*, 276 F.R.D. 633 (C.D. Cal. 2011), explaining that fee shifting in litigation from one party to another is an externality suffered by the party who is forced to bear an unexpected cost.

non-traditional imposition of liability that would allow nearly anything to be counted as a cost for which a defendant may be held liable.[132]

Given the huge number of cases reported in the federal and state court systems over 40 years, and the pervasive use of externality in economics, the small number of cases in which the term has been employed to provide a justification for a peculiar decision seems remarkable. If judges wish to construct scientific-sounding arguments based on presumed economic analysis, externality provides a handy tool that few judges have stooped to use.[133]

5.18 Externality in Economics

Externality in economics is a concept is so loose as to be useless in an aid to legal analysis. Unlike other areas, such as antitrust, in which economic analysis has played a substantive role in helping courts comprehend market structures, externality may be a useful word that denotes costs or benefits that spill over, but it provides no substantive understanding of particular problems.

Consider some standard definitions of externality: "An externality exists when one (or more) economic actor(s) affect(s) another actor (or group) directly without the intervention of a market transaction." (Russell 2001, p. 45) "Maximum social welfare is only attained…if marginal private costs also equals marginal social cost, for it is only then that marginal social benefits and marginal social cost are equal." (Ferguson 1972, p. 497) "An externality exists when a person does not bear all the costs or receive all the benefits of his or her action. An externality exists when the market price or cost of production excludes its social impact, cost, or benefit. Externalities are everywhere." (Hanley et al. 2001, p. 17) Indeed they are. If you wear a shirt that offends the eyes of some beholders, you have inflicted costs on those persons; they have suffered externalities.

But, as noted before, only some externalities are regarded as worthy of concern in economics. Pecuniary externalities are ignored (Worcester 1969).[134] These are created, for example, when one law firm draws clients away from another law firm, thereby inflicting a financial loss on that firm. No doubt the members of the losing firm feel the pain; they have suffered a loss. But since the transactions are voluntary,

[132]With respect to interesting applications of the notion of externality, one unreported case is worth noting, *Winter v. Office of the President of the United States*, 1997 WL 102513 (N.D. Cal. 1997), in which plaintiff issued a "writ of externality" on behalf of the American people for $50 trillion of laundered drug money, plus $10 million punitive damages. Externalities, used loosely, can cover just about anything imaginable.

[133]This does not imply that the judges who made use of externality to bolster their position were intentionally making clever use of economics they knew to be weak; economists talk so much about externality as if it is a scientific concept that one could easily assume it provides solid theoretical justification for a desired legal outcome.

[134]"[E]verything can be said to affect everything else, so we ignore many things and focus on technological externalities." (Hirshleifer 1980, p. 532).

we presume society as a whole is improved by better allocation of scarce resources. The transitory hits taken by parties in the process of competition are not the kinds of actions that affect others that are of concern. The law is consistent with economics on this point; so long as such transactions are voluntary and no fraud or tort is involved, there is no reason for economic or legal concern. A new contract has been formed by willing parties. Those who did not get cut into the deal cannot make a claim against those who are parties to the deal.

Technical externalities are the kinds of spillover effects of concern in economics. Network externalities, discussed previously, are one example. But the classic eternality is pollution. One makes wheelchairs for disabled persons and, in the process, throws wastes into the air, land and water that inflict harm (costs) on others. How to eliminate the problem?

Guido Calabresi has explained: "Thus if one assumes rationality, no transaction costs, and no legal impediment to bargaining all misallocations of resources would be fully cured in the market by bargains." (Calabresi 1968, p. 78). Those suffering from pollution emitted by the wheelchair maker will bargain with her to reduce emissions. This is often difficult to have happen, explained George Stigler: "The world of zero transaction costs turns out to be as strange as the physical world would be without friction." (Stigler 1972, p. 12). That is, trying to eliminate externalities is like investing in the perpetual motion machine. All "wrongs" cannot be righted and perfection in cost allocation can never occur. The gains to participants in potential exchanges may be very small or deals may be very difficult to make (high transaction costs).

5.19 Relevant Externalities

What externalities can be or should be addressed? One classic article on the subject suggested that Pareto-relevant externalities can be dealt with; Pareto-irrelevant externalities cannot (Buchanan and Stubblebine 1962). Pareto-relevant refers to gains from trade that parties recognize and, to their mutual benefit, bargain into existence. So long as the expected gains outweigh the bargaining costs, the parties can be expected to improve their (and, therefore, societys) welfare. Such externalities are self-correcting, as the parties recognize the potential gain, or they may become solvable if it becomes less costly for the parties to bargain, a point returned to below.

Pareto-irrelevant externalities are cases in which the expected cost of bargaining is greater than the expected rewards, so we let those events pass. The gains may be trivial what is it worth to you for a person not to wear a shirt that offends your sense of fashion? Life is an endless stream of little nicks and cuts that we ignore because the many tiny imperfections, as we see them,[135] are not worth messing with. Add them up across a huge number of people and the supposed losses can be

[135]The person wearing the ugly shirt probably thinks it is just dandy.

immense. Add up little potential gains from the many things that many people enjoy, but do not pay for, and that number can also be immense. The problem is that, in the absence of transactions, reliable measures of value do not exist. Dollar values cannot be placed on such externalities. "In economic terms, people have a difficult time assigning hypothetical dollar values to categories and commodities they virtually never confront in everyday experience." (Sunstein and Pildes 1997, p. 142)

The presumption that because such externalities are not priced, the bad ones will be oversupplied and the good ones will be undersupplied, analytically is not clear (Haddock 2007). Markets function within the law. The purpose of legal rules is to give guidance to market participants as to the boundaries and provide the possibility of restitution in case of improper action. The purpose is not for courts to set proper prices when consideration in contracts is not reflective of "fair market value" nor is it to invent costs, such as in cases of torts or environmental issues, that are not expected by parties to actions.

As the courts in several cases reviewed above noted, if it is desirable to pay for greenhouse gas offsets through the setting of electric rates by a public utility commission, it is for the legislature to give such instructions, not for courts to presume that such costs, traditionally not counted should be counted. Courts have no more business setting such rules than they would declaring there should be compensation for all externality sufferers from electricity. Because the lights are brighter, the offensive shirt is even more glaring, so the offended party should receive some compensation from all users of electricity. Externalities can be an endless regress. Summary Externalities are an empty set. Economists have employed the term for decades but know nothing more today than when the discussion started. Judges have, for the very large part, ignored externality other than as a reference to the fact that many events create bad or good effect. That is a perfectly sensible of the word. To go further and declare that knowledge of the fact of unmeasured costs allows them to be measured by expert economists opens the way to endless mischief of pursuing individual interests in the guise of high science. Decades of fruitless discussions among economists have made little progress in an operational meaning for externalities. Assuming such discussions can be put into practice would terrible consequences for the rule of law in a free society. Bad economics can contribute to the making of bad law. When it does, there is a major externality.

References

Buchanan JM, Stubblebine WC (1962) Externality. Economica 29(116):371–384
Calabresi G (1968) Transaction costs, resource allocation and liability rules-a comment. J Law Econ 11(1):67–73
Coase RH (1960) The problem of social cost. J Law Econ 3(1):1–44
Epstein RA (1987) Foreword: unconstitutional conditions, state power, and the limits of consent. Harv L Rev 102:4
Ferguson CE (1972) Microeconomic theory. Homewood, Chicago
Haddock DD (2007) Irrelevant externality angst. J Interdiscip Econ 19(1):3–18

Hanley N, Shogren J, White B (2001) Introduction to environmental economics. Oxford University Press, Oxford
Hirshleifer J (1980) Price theory and applications. Prentice-Hall, New Jersey
Holmes OW (1899) The theory of legal interpretation. Harv L Rev 12(1):417–420
Liebowitz SJ, Margolis SE (1998) Network externalities (effects). In: Palgrave's N (ed) Dictionary of economics and the law. Macmillan, London
Metcalfe RM (2007) It's all in your head the latest supercomputer is way faster than the human brain. But guess which is smarter? Forbes 179(10):52
Posner RA (1972) A theory of negligence. J Legal Stud 1(1):29–96
Posner RA (1992) Economic analysis of law. Little Brown and Company, Boston
Revesz RL (1992) Rehabilitating interstate competition: rethinking the race-to-the-bottom rationale for federal environmental regulation. NYU L Rev 67:1210–1255
Russell CS (2001) Applying economics to the environment. Oxford University Press, Oxford
Sato S (1972) "Municipal affairs" in California. Cal L Rev 60(4):1055–1115
Stigler GJ (1972) Law and economics of public policy: a plea to the scholars. J Legal Stud 1:1–12
Sunstein CR, Pildes R (1997) Free markets and social justice. Experts, economists, and democrats. Oxford University Press, New York
Worcester DA (1969) Pecuniary and technological externality, factor rents, and social costs. Amer Econ Rev 59(5):873–885

Chapter 6
Why Would Bond Referenda Ever Fail? Do They?

William S. Peirce

Abstract Local bond referenda provide the best available information to test whether agenda setters prefer higher levels of public investment than do the voters. This study examines the entire population of local bond referenda in Ohio from 1963 through 1987. The results do tend to support the hypothesis that agenda setters attempt to raise expenditures above the level preferred by the median voter. Although about half of all referenda fail, most projects eventually pass–as is predicted by the hypothesis of expenditure maximization.

6.1 Introduction

The question of whether spending patterns of local governments reflect the preferences of median voters, or whether politicians and bureaucrats are able to extract more than the preferred amount of resources, has long attracted attention.[1] Politicians and bureaucrats may prefer larger budgets but, especially at the local level, citizens should be able to elect faithful representatives and can exit high tax communities if public spending does not correspond to individual preferences. Still, the power of politicians in setting the agenda may permit exploitation. The superior information of bureaucrats regarding costs and other consequences of budgets and the knowledge by politicians of their next move if a referendum is lost also may give them power to exploit.

Such a list of factors that may be influential suggests that the particular institutional framework within which the expenditure decisions are made will make a difference to the outcome. The case for the decisiveness of the median voter is strongest in Hotelling's (1929) linear model where voters choose between two parties on the basis of a single issue. In a different institutional setting; namely, direct voting on individual

[1] Holcombe (1989) provides a compact and incisive survey of the literature.

W.S. Peirce (✉)
Case Western Reserve University, Cleveland, OH 44106, USA
e-mail: william.peirce@sbcglobal.net

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_6

issues, the case for the median voter is harder to make. Romer and Rosenthal (1978), in a famous study, found that a particular institutional structure enabled agenda setters to extract a school budget that exceeds the preferred spending of the median voter. Briefly, when voters in Oregon and some other states are presented with a referendum on school operating funds, the choice is to accept the level requested by the agenda setter or to reject it, in which case taxes and the operating budget revert to a historically determined level. Romer and Rosenthal (1978) do, indeed, find that spending is inversely related to reversion level.

Local bond referenda can be analyzed using an approach conceptually similar to the Romer and Rosenthal (1978) model. The reversion level is that the capital expenditure is not undertaken. Local bond referenda provide more specific and detailed information about the preferences of voters than can be obtained from any other source. Each referendum provides an exact count of those with strong enough preferences to vote for and against a particular issue, often quite narrowly defined. Votes on particular candidates never provide such specific information on desired expenditures because the elected official will participate in a variety of budgetary and non-budgetary issues - some of which are not even imagined at the time of the election-during his term in office. Even the votes on local tax levies are often not as clear cut because of the uncertainty about what the levy will pay for.

Despite the wealth of information, bond referenda have rarely been studied by specialists in public economics. Mikesell and Blair (1974) use an approach motivated by consumer demand for durable goods. With data from 36 school bond elections in West Virginia, they estimate a demand equation for school buildings and arrive at the favorable vote as a desired adjustment to the stock. DeBartolo and Fortune (1982) use data about the median voter to estimate the probability that a bond issue will pass. They use 205 bond referenda in Ohio between 1968 and 1972 to construct their model. Although the primary focus is on the elasticities of demand for public services, they also conclude that the existing stock of public goods corresponds to the preferences of the median voter. It appears that the main reason for that conclusion is that the probability of passage of a bond referendum is about 0.5.

McEachern (1978) did a cross-sectional analysis of local debt across the 50 states. He found no difference between states that required a referendum and states that allow politicians to contract the debt. This he interpreted as evidence that debt in both states is at the level preferred by the median voter. States that required a supermajority had lower debt levels.

Fort (1988) provides the closest parallel to this paper, but with some key differences. He looked at referenda for construction of rural hospitals and found no support for the hypothesis that the setter extracts the maximum possible expenditure. This study is noteworthy because it makes explicit use of the strength of the vote in favor of a project.[2] Rural hospital referenda generally passed by huge margins. This may be due to the peculiar characteristics of the issue, however. In particular, the rural areas

[2]Romer and Rosenthal (1983) discuss the relationship between the median voter model and the proportion voting for a proposal. It is surprising that this statistic is so widely ignored in the median voter literature.

were threatened with particularly dire consequences if the issues failed, but at the same time, the setters understood that they would have to return to the communities for operating funds to meet the predictable deficits.

By comparing Fort's article with the other attempts to study bond referenda, it is possible to see why such infrequent use has been made of such a rich source of information. In studying annual expenditures, it is easy to relate the annual flow of expenditures to the characteristics of the median voter in the jurisdiction. When dealing with the major capital expenditures financed by bond issues, however, the standard median voter approach is fatally flawed. The sporadic nature of the expenditures combines with the durable nature of the capital stock to complicate the analysis substantially. In particular, while a community with high current expenditure this year can be expected to have high current expenditure next year, the community that spends a lot on capital this year may have a large enough capital stock next year, in which case investment ceases. Alternatively, it may be a community with a strong preference for investing public capital, or high costs, so investment will continue at a high rate. At its best, a median-voter type study estimates a demand function for capital stock. The demand for new investment would depend on the deviation of demanded stock from actual stock. No outside observer, however, can estimate the adequacy of the existing quantity and quality of capital from the perspective of the median voter. In any event, one would not expect to find a close relationship between any one year's capital expenditures and the variables that are typically used to explain annual spending.

Nor should one expect a close relationship between the standard determinants of spending and the approval rate for bond issues. Bond issues are ordinarily proposed by politicians in the relevant community. Approval of a proposal means that the politicians projected correctly the preferences of their constituents. Rejection indicates that the politicians were out of touch with preferences of constituents. There does not seem to be any reason to suppose that politicians in rich or well educated communities are more attuned to the preferences of their constituents than are the politicians in the rural backwaters and slums of society.

Despite the richness of the bond-issue data, therefore, it is necessary to exercise some caution in framing the questions that one tires to answer with the data. For that reason, the next section will spell out a simple voting model in some detail in order to derive testable implications.

6.2 Theory

Figure 6.1 depicts the total cost (TC) and total benefit (TB) curves that a typical voter expects from a potential public project that might be presented as a bond issue. The horizontal axis shows the size of the project measured by the dollar amount of the bond issue. The vertical axis measures in dollars the total benefit and the total cost as perceived by the individual. The TC curve is drawn as a ray from the origin, reflecting the expectations that the bond issue will be financed by an increase in

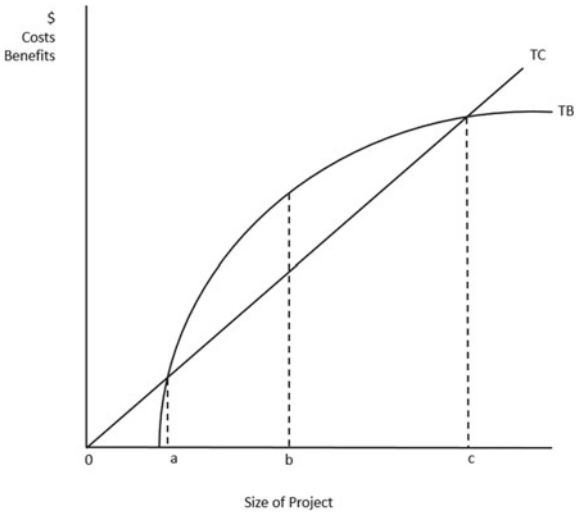

Fig. 6.1 Individual evaluation of a bond proposal

the property tax and the individual voter pays a constant fraction of any increase in property taxes. The steepness of the TC curve reflects the tax share that the particular individual believes he pays. The TB curve is slightly more complex. As drawn, the curve begins at some minimum size, which can reflect indivisibilities in production (a bridge halfway across the river is useless) or the belief by the individual voter that a small project would benefit only some other people (the first street paved will by the one the mayor lives on, so a street-paving bond issue must be big before I get anything). Regardless of the starting point, the total benefit increases at a continually decreasing rate, reflecting decreasing marginal benefit of all but the strangest projects. The TB curve can even turn down, but this will be ignored.

Whether TB ever exceeds TC depends on the tastes and taxes of the individual. If cost always exceeds benefit, the individual will oppose the project, regardless of the size under consideration. In Fig. 6.1, the individual receives a large enough benefit to make him favorably disposed toward projects that are larger than a but smaller than c. This individual derives the net maximum benefit from size b. The bond referendum is such a blunt instrument, however, that the sophisticated behavior of rejecting something between a and c in the hope of moving closer to b in a later election is assumed to be rare.

Although the individual depicted in Fig. 6.1 is favorably disposed toward the project, he may not vote for the bond issue. The curve labeled TC is actually "tax cost" rather than the "total cost." If the cost of voting exceeds the net benefit from the program, the individual will stay home even though he is favorably disposed. This is shown in Fig. 6.2, where the cost of voting is such that the individual never bothers to vote de-spite the range where total benefits exceed the taxes that must be paid. Of course, the figure could have been drawn to leave some range in which the individual would obtain a large enough benefit to more than compensate for the effort of voting,

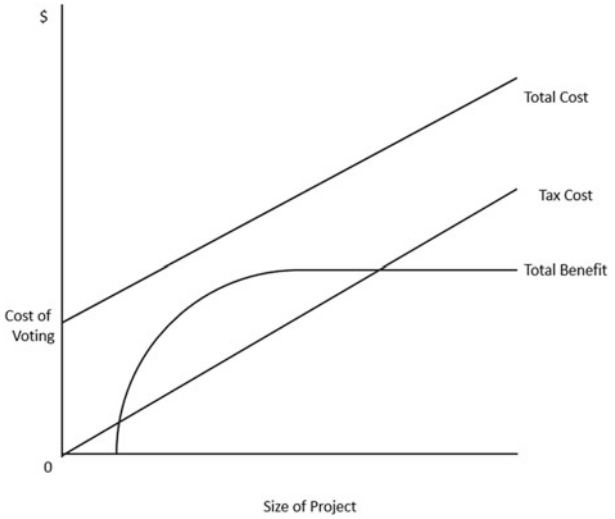

Fig. 6.2 Evaluation of a bond proposal when voting costs are high

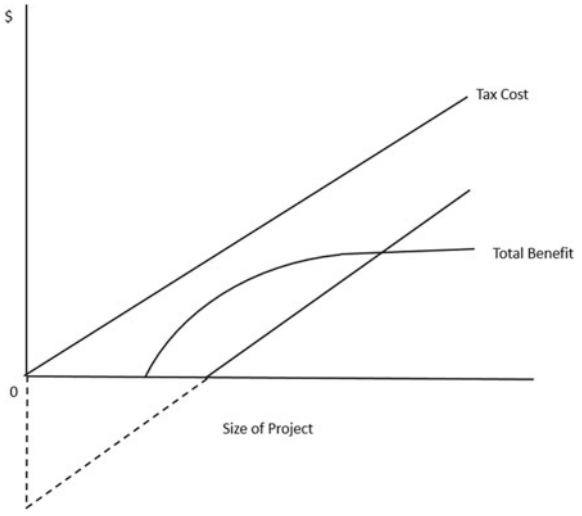

Fig. 6.3 Negatively inclined voter with high voting costs

as well as taxes. Even in that case, however, it is a narrower range of proposals that will attract the individual to the polls.

The implications of the cost of voting for the negatively disposed voter are indicated in Fig. 6.3. Again, depending of the exact shape of the benefit and cost schedules, the individual may not think it worthwhile to register his opposition by voting.

The cost of voting is analyzed separately from the tax cost of the project for several reasons. First the decision to bear the voting cost is made by the individual voter, unlike tax cost, which depends almost entirely of the decisions of other. Second, the

cost to the individual of voting on a specified issue varies from time to time. Not only does it depend on the weather, which is often discussed by commentators on elections, but also it depends on the other issues on the ballot. If the individual is already at the polls because of a presidential election about which he feels strongly, the additional cost of voting on a bond issue is negligible. At the other extreme, if a special election is held for the bond issue, turnout will ordinarily be low because many people will believe that the gain or loss from the passage of the issue will be less than the cost of voting. A third reason for treating voting cost separately, which depends on the variability of costs by election, is that voting costs can be expected to have a greater effect on the negative than on the positive vote.

The reason for this expectation is found in the basic theory of public choice, as well as Figs. 6.1, 6.2, and 6.3. While projects have costs that are spread broadly over the entire population, the benefits tend to be narrowly focused on particular groups. Those with an intense interest in the project will not be deterred by voting cost, but the mildly opposed will stay home.

So far, the discussion has concentrated on the preferences of individuals for proposed projects of various sizes. The natural question, however, concerns the origin of such proposals. Without specifying further, let us state that bond issues are placed on the ballot by politicians in the relevant jurisdiction. Obviously, there is no end to the possible projects that might be presented to the voters. As a working hypothesis, let us assume that politicians prefer to increase the size of government but do not like to be associated with losing causes. As possible capital expenditures are not suggested, politicians look first for sources of financing not subject to referendum. If that attempt fails, the politician tries to estimate whether the voters are likely to accept any version of the project and, if so, the maximum amount that will be approved. This requires great political skill because projects can be designed in different ways to appeal to different people and voters may be persuaded to approve projects that they had never even though about prior to the campaign.

Figure 6.4 depicts a simplified version of the information that the politician tries to estimate. The horizontal axis gives the size of the project. The dollar amount must stand as a proxy for all the subtle variations in quality and quantity that can characterize a capital expenditure. The vertical axis shows the ratio of yes votes to no votes expected by the politician for each size of project.

Assuming that the politician is an expenditure maximizer but does not want to sponsor losing causes, he will want to place on the ballot the largest proposal that will command a yes/no ration greater than 1.0. The curve drawn in Fig. 6.4 could have multiple peaks, but that is irrelevant if the politician is interested only in the largest project the voters will accept. If the project were to be of the size favored by the median voter, a community of reasonably tolerant people with fairly homogeneous preferences would approve it by an overwhelming margin. The expenditure maximizing agenda setter would react to a forecast of overwhelming approval by increasing the size of a project.

Forecasting accurately the point at which the yes/no ration just exceeds 1.0 is very difficult. Some politicians are more successful than others, and the ones that dominate the meetings where bond issues are designed may not be the most accurate

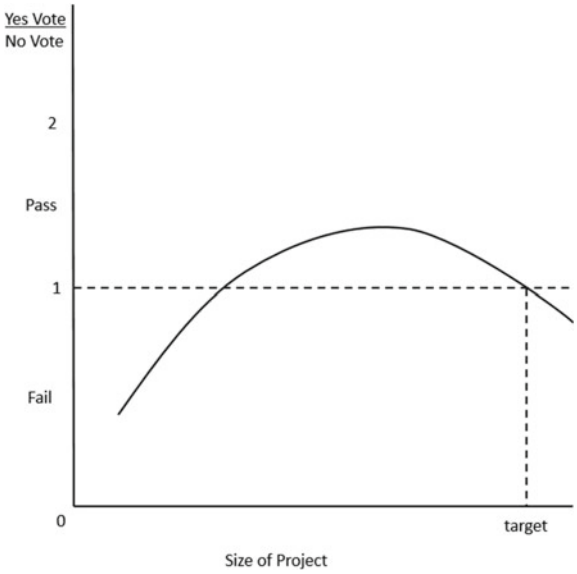

Fig. 6.4 Aggregated preferences of voters

forecasters of voter behavior. Moreover, the yes/no ratio will vary with voter turnout, as well as with changes in the preferences or incomes of individual voters. It is possible that the politician will guess that the yes/no ratio will not exceed 1.0 for any project size. If he cannot suggest a way to modify the proposal so that it will pass, he will not support it at this time. Most ideas for spending fall in this category and most are such obvious losers that they are not even discussed seriously. Presumably, however, those that are placed on the ballot by politicians are expected to pass.

In a world where politicians had perfect knowledge of voter preferences, therefore, one could expect all referenda to pass. Failure means that the politicians guessed wrongly about the preferences of the voters. In the real world we observe that a significant fraction (about one-half) of bond referenda result in failure. This should be a measure of the lack of forecasting success of the politicians and perhaps of the politicians tradeoff between the goals of expenditure maximization and avoiding losing causes. In any event, one would not expect any relationship between the yes/no ration or the passage rate (number of issues passed/number of issues voted on) in a community and the variables that are typically used in expenditure studies such as average income, education, age, proportion of renters, etc. Skilled politicians take all such matters into account before proposing an issue.

Of course, rapid changes in any variables that influence voter behavior could lead to increased errors by politicians. Errors can be of two types: Type I is to propose an issue that fails because it is larger than the voters will accept. Type II is to propose an issue that obtains a yes/no ration much larger than 1.0; for this means that the politician "left money on the table" by not raising the amount requested to the maximum that would pass. Errors can be specific to a particular community that

is changing rapidly or can reflect general changes in the public mood that are not fully anticipated by politicians.

Although about half of all bond referenda result in failure, that does not mean that half of all proposals fail. If each proposal is aimed at the maximum amount that will pass, but the aim is disturbed by a symmetrically distributed random error, then the probability of passage is 0.5. If the same issue that failed in one election is resubmitted (assuming unchanged conditions), the probability of passage is again one half. More generally, in this simple model the probability that an issue would not have passed after n tries is the probability of failure on any one try raised to the n^th power. If the probability of failure is really 0.5, after 3 attempts to pass a proposal only one proposal in eight would still stand rejected. Nevertheless, the aggregate data would show the voters turning down half of the issues at every election. This suggests the importance of following the sequences of votes by particular communities on particular issues.

The task of analyzing sequences of votes is made more difficult by the behavior of politicians in response to failure of a proposal. If the politician takes Fig. 6.4 at face value, he could improve the chances of passage by reducing the size of the proposal. Thus, a change in the dollar amount does not mean that it is a different issue. Indeed, even an increase in the dollar amount may occur, either as a result of price increases between elections or as a result of a reconsideration and redesign of the proposal so that it has special benefits for a larger share of the potential voters in the jurisdiction. Of course, the easiest response by the politician is just to wait until an off-election to submit the same proposal in the expectation that more of the negatively-inclined than of the positively-inclined would fail to vote.

6.3 Data and Empirical Analysis

The date used in this analysis are the "Bond Issue Election Results" collect by the Ohio Municipal Advisory Council (OMAC). The date cover all governments below the state level proposing bonds for voter approval in Ohio during the period from 1963 through 1987. Although some observations have been lost or are incomplete, the date comprise essentially the entire population of bond referenda submitted by cities, villages, school districts, and counties in Ohio.

Each observation includes the name of the issuer, the amount, duration, and purpose of the issue, the date of the election, and the number of votes for and against the issue.

6.3.1 Aggregate Data

Table 6.1 shows a first breakdown of the data. Almost exactly half of the referenda, 1,732 approvals of the 3,515 votes, resulted in passage of the bond issue. School

Table 6.1 Bond approval rates by purpose in Ohio, 1963–1997

Purpose	Passed	Total	% passed
Total	1,732	3,515	49
School	1,179	2,482	48
Non-school	548	1,033	53
Library	11	18	61
Fire, police	135	136	99
Facility for old, young, handicapped	34	37	92
Water, sewer, electricity	135	151	89
Jail	0	3	0
Recreational	24	41	58
Hospitals, etc.	32	33	97
Roads, bridges	36	43	84
Real estate	6	7	86
Government buildings, etc.	38	49	78
Urban development, harbors, pollution control	5	7	71
Airport	1	2	50
No purposes indicated	96	506	19

Source Ohio Municipal Advisory Council, "Bond Issue Election Results"

bonds constituted more than two-thirds of all issues (2,482) and they had an approval rate of 48 %. The total approval rate for the 1,033 non-school proposals was not greatly different at 53 %, but the components require closer analysis. As can be seen in Table 6.1, most of the specific categories had approval rates above 70 %. The exceptions were libraries (61 %), recreational facilities (58 %), airports (1 of 2), and jails (none of 3). The approval proportion for the non-school category was, however, drastically reduced by the low approval rate for bond issues for which no purpose is indicated in the data. Of the 506 such issues, only 96 (19 %) passed.

It is tempting to interpret that as an indication that voters will not approve a bond issue unless it promises some benefits, which an issues with no indicated purpose clearly does not. A bond referendum in Ohio, however, must state some purpose, so "no purpose indicated in the date" reflects a defect in the data, not a characteristic of the bond proposal. If, however, the purpose is stated so vaguely that is does not readily fit a standard category, it may be so vague that the voter rejects the proposal and the reporter fails to record the purpose. A more plausible explanation is that, in the haste to report results, OMAC is most concerned with the issues that have passed and hence is likely to skip over the "purpose" of the failed issues if higher priority activities are pressing.

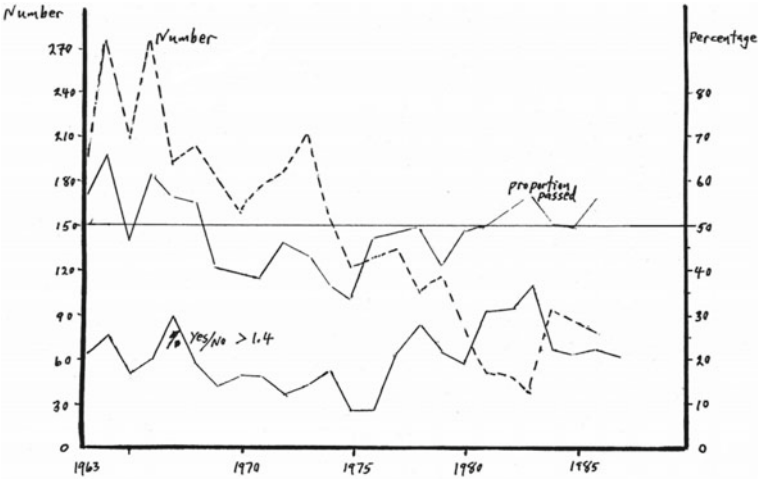

Fig. 6.5 Number of referenda and proportion passed, 1963–1986

An examination of the approval rates by county reveals a range of 0 to 83 % for Ohio's 88 counties. The number of referenda in individual counties for the entire 25 years ranged from 1 to 353. Restricting the analysis to the seven counties in which at least 100 issues were presented, 607 of 1,101 issues were approved. This is a rate of 55 %, which is slightly higher than the overall rate. The difference is not remarkable enough to inspire any disparaging remarks about the forecasting skill of back-country politicians. The approval rates in the 7 most active counties ranged from 41 % to 73 %.

Figure 6.5 shows three annual time series: (1) the number of referenda in Ohio; (2) the percentage approved; and (3) the percentage in which the Yes/No ration exceed 1.4. The number of issues presented to the voters exceeded 250 in some years during the 1960 s and then decreased erratically to less than 50 in the early 1980 s before rising to 85 in the final period. Despite the decrease in the requests of the politicians, the voters rejected a higher proportion of the issues in the early 1970s. After 1975 the approval rate increased and the submission rate followed with a lag. Although one could explain the data in a variety of ways, it does not seem unreasonable to interpret them as evidence of a "taxpayers' revolt" that surprised the politicians, who subsequently reduced their requests for bond issues.

The lower percentage the proportion of issues in which the Yes/No ratio exceed 1.4 fluctuated wildly. The lowest observation was 8.42 % in 1975, while the highest was 36.84 % in 1983. I would interpret this percentage as an indication of the frequency with which setters left money on the table. Whether this reflects an attempt to satisfy the median voter, or whether the expenditure maximizer made a mistake, cannot be distinguished from the data.

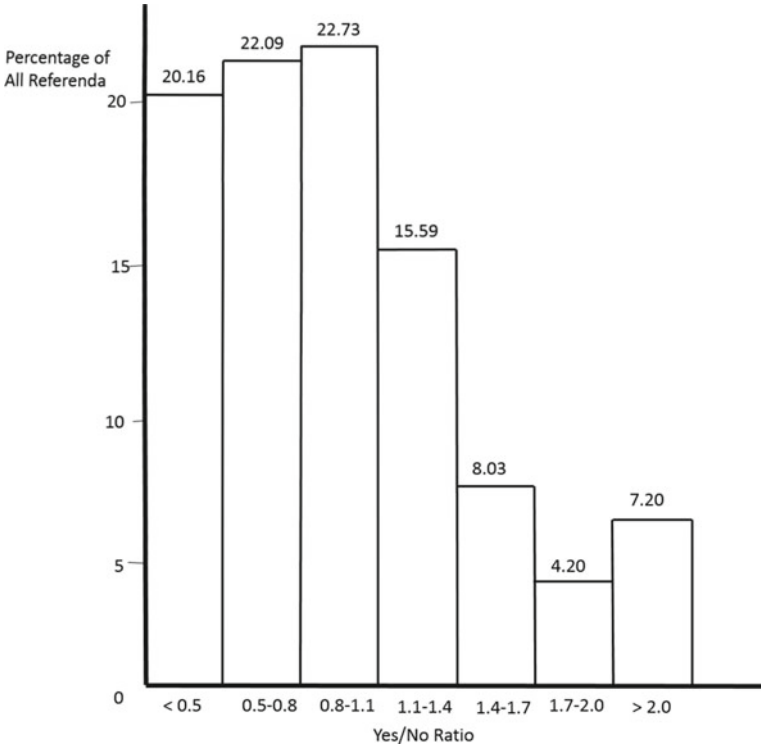

Fig. 6.6 Distribution of votes

Figure 6.6 shows the frequency distribution of the yes/no ratio in the entire data set. In less than 20 % of the referenda did the ratio exceed 1.4. It appears that politicians generally do not leave money on the table, but rather ask for at least as much as they can get. Indeed, one would expect that a change in the proportion of very high or very low ratios would indicate an unexpected shift in voter preferences. In subsequent periods, politicians would react to their error in forecasting by adjusting either the number of size of bond proposals.

At the aggregate level, the data support the hypothesis that those who are negatively inclined have less intense preferences than those who favor a proposal. This is indicated in Table 6.2, which shows the approval rates by time of election. November elections, when turnout is greatest, have an approval rate of 47 %. Referenda presented at the May or June election, which often has interesting primary contests, are approved 49 % of the time. At the special elections held at unusual times, however, 66 % of the bond issues pass.

Table 6.2 Passing proportions by date of election

Date	Number of referenda	Number passed	Proportion passed
Total	3,515	1,732	0.49
November	2,321	1,084	0.47
May or June	833	411	0.49
Other	361	237	0.66

6.3.2 Analysis of Sequences

The most fundamental grouping of referenda for analysis is the sequence. A sequence is defined as one or more referenda on the same issue. A complete sequence ends with the passage of the issue, but a sequence may be abandoned after losing one or more times. The data also include sequences that are incomplete because they are still going on. The analysis of early sequences may be distorted because some were ongoing when the data began. If 0 is used to signify failure and 1 signifies passage, complete sequences will be (1), (0, 1), (0, 0, 1), and so on. Regardless of the number of failures, each complete sequence ends in passage. The jurisdiction may proceed to vote on a similar issue after passing a bond issue, but that is treat as a new sequence.

Empirical analysis of sequences is difficult because the amount and duration of the bond proposal may change with time. Indeed, reducing the amount or extending the duration of the proposed bonds may be ways to increase the probability of passage. Another empirical problem is the large number of failed referenda for which no purpose is indicated, which may be the failures that belong in some other sequences.

If passage is essentially random with some fixed probability of occurrence, then failure could lead to resubmission of the same proposal at a later election, with the same unchanged probability of passage. Alternatively, issues presented the first time may constitute two separate populations: good ones have a high probability of passage, while bad ones are likely to fail and then join a population of losers that fail repeatedly. Those speculations ignore the possible responses of politicians. In addition to presenting the same issue at every regular election until is passes, politicians might (1) modify a proposal to make it more appealing or cheaper, (2) try to slip it though at a special election, or (3) abandon the issue.

As a first step in the empirical analysis, Table 6.3 lists the expected frequencies of various sequences under the assumption that passage is totally random with a probability if 0.5. The table also lists the observed frequencies of various sequences in the entire set of bond issues. Proposals that pass on the first try are somewhat more common than expected (54.7 % vs. 50 %). More notable is the rapid drop in the observed frequency relative to the expected for longer sequences. The real surprise, however, is the large proportion of sequences that end in abandonment of the proposal, rather than passage. It seems appropriate to describe the set of issues presented for the first time as consisting of two types-those that will pass quickly and those that will never pass-rather than assuming that the probability is the same for all issues.

Table 6.3 Expected and observed frequencies of sequences

	Sequence	Expected frequency	Observed frequency		
			Total	School	Non-school
1	1	0.5	0.547	50.8	61.6
2	01	0.25	0.116	16.4	3.3
3	001	0.125	0.049	7.4	0.7
4	0001	0.0625	0.019	3.0	0.1
5	00001	0.03125	0.011	1.7	
6	000001			1.0	0.1
7				0.3	
8				0.4	0.1
9				0.1	
...					
17				0.1	
	Abandon	0	0.244	18.9	34.0

This finding also suggest that politicians do not respond actively to the failure of a bond issue by searching for the modifications that will enable it to pass.

Further insights are available from a comparison of the frequency of various sequences for school and non-school issues. The data are presented in the last two columns of Table 6.3. Among non-school issues, 62 % pass on the first try and 34 % are abandoned. That does not leave much room for complete sequences of more than one presentation! School issues present some contrast because the initial passage rate is lower (51 %), the abandonment rate is lower (19 %), and issues are presented again and again until they pass. For example, of the 51 issues that passed after 4 or more tries, 49 were school is-sues. This difference probably reflects the fact that the conditions that lead a school district to place a bond issue on the ballot do not disappear with repeated rejection of the issue. In the non-school issues, a different type of capital expenditure, or even a tax decrease, may be useful to the politician in obtaining his goal of reelection. The politician is more careful to put what he thinks will be a winner on the ballot-and will abandon it if it fails. Boards of education, however, cannot ordinarily substitute a popular bridge for an unpopular high school.

Although the preceding results cast considerable doubt on the strength of the processes discussed in the theoretical section, it seemed useful to test directly some of the expected behavior. Specifically, the last two elections in complete sequences having two or more elections were examined. That means that the issue lost in the first election and won in the second. It was hypothesized that politicians would increase the probability of success of an issue by decreasing the amount, increasing the duration, and scheduling the election sometime other than November. Furthermore, it was assumed that the specially interested people voting yes would vote more consistently than those opposed.

In about 30 % of the sequences the proposal that passed was for a lower amount. A similar proportion showed an increase in duration. This it appears that some effort was made to adjust proposals in a way that would increase support. About 62 % of the sequences showed a decrease in the no vote, while about 64 % showed an increase in the yes vote. This is perhaps to be expected in moving from a lost election to a won election. A better test of the stability of the yes vote relative to the no vote is to compare the standard errors directly. Contrary to expectation, the standard error of the yes vote exceeded the standard error of the no vote in 465 of the 679 sequences examined.

6.4 Conclusion

Local bond referenda do offer a wealth of information, but to make the most of that information requires a more sophisticated modeling of the political process that results in referenda being placed on the ballot. The importance of this is revealed by the differences between the school and non-school issues, which have different origins. Although bond referenda do have their unpredictable aspects, it is probably not useful to apply the simple model of a random process. Politicians (other than, perhaps, boards of education) do not keep rolling the dice, waiting for a victory. Instead, they respond actively by either modifying the proposal or withdrawing from it.

As for the questions posed in the title, it is clear from the theory that in reasonably homogeneous communities of reasonably flexible voters, bond referenda would rarely fail if skilled politicians were trying to satisfy the median voter. The fact that about half of all referenda fail is strongly suggestive of an attempt to maximize project size. It is quite possible, however, that certain types of referenda are set by a political process that represents voters, rather than budget maximizers. Perhaps rural hospitals and the affairs of small communities, generally, fall into this category. Nevertheless, it is noteworthy that the school data, which are most numerous and most complete, fall more neatly into the expenditure maximizing model than do the non-school issues.

Do issues ever fail? The model of repeated presentation and modification until passage is overly simplistic, but it comes closer to describing the process than the casual leap from a 50 % passage rate of referenda to the unwarranted assumption that half of all issues fail.

References

DeBartolo G, Fortune P (1982) The demand for public services: inferences from municipal bond referenda. Natl Tax J 35:55–67

Fort RD (1988) The median voter, setters, and non-repeated construction bond issues. Public Choice 56(3):213–231

Holcombe RG (1989) The median voter model in public choice theory. Public Choice 61(2):115–125

Hotelling H (1929) Stability in competition. Econ J 39(153):41–57

McEachern WA (1978) Collective decision rules and local debt choice: a test of the median-voter hypothesis. Natl Tax J 31:129–136

Mikesell JL, Blair JP (1974) An economic theory of referendum voting: school construction and stock adjustment. Public Financ Rev 2(4):395–410

Romer T, Rosenthal H (1978) Political resource allocation, controlled agendas, and the status quo. Public Choice 33(4):27–43

Romer T, Rosenthal H (1983) Voting and spending: some empirical relationships in the political economy of local public finance. Local provision of public services: the tiebout model after twenty-five years. Academic Press, New York, pp 165–182

Chapter 7
The Effect of Early Media Projections on Presidential Voting in the Florida Panhandle

Russell S. Sobel and Robert A. Lawson

Abstract The media incorrectly called Al Gore the winner of Florida at 7:48 p.m. eastern time with 12 min still remaining to vote in the 10 Central time zone counties in Florida. In addition, the media reported that polls "close in Florida at 7:00 p.m. eastern time," which may have misled some panhandle voters into thinking their polls closed at 6:00 p.m. central time. Given the closeness of the popular vote in Florida, and the degree to which the outcome in the state was contested, these media miscues could have been decisive in the election. When Bush was behind in the recount, his supporters adamantly claimed their candidate suffered a loss of votes because of these media miscues. We test this hypothesis and reject it. Our regression results find no significant impact on the Gore/Bush vote differential, nor do we find any impact on voter turnout or third party voting, in these counties.

7.1 Introduction

Many strange things may have happened in the state of Florida during the presidential election of 2000. As never before, the Internet allowed a rapid dissemination of research on these events. Within one week after the election, at least 20 different scholarly analysis of the butterfly ballot in Palm Beach County issue were posted on the Internet. While this paper is on a slightly different issue in this election, a very early draft of it was part of this historic research revolution and was circulated and discussed on the Internet.[1]

[1] Several of these studies were featured on the www.yahoo.com site and were summarized on Greg Adam's web page at Carnegie Mellon University (madison.hss.cmu.edu) and Jonathan Okeefe's webpage (bestbookmarks.com), which are archived at the Internet Archive.

R.S. Sobel (✉)
The Citadel, Charleston, SC 29409, USA
e-mail: russell.sobel@citadel.edu

R.A. Lawson
Southern Methodist University, Dallas, TX 75275, USA
e-mail: rlawson@smu.edu

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_7

In this paper we analyze the possibility of voting anomalies in Floridas western panhandle-or what became more commonly known in the aftermath of the election as "the Central time zone hypothesis." The polls in Florida close at 7:00 p.m., however there are ten counties in the western panhandle of Florida that are in the Central time zone (rather than the Eastern time zone like the remainder of the state), so polls in those western counties do not close until 7:00 p.m. Central time, which is 8:00 p.m. Eastern time, or an hour later than the polls in the rest of the state. The major media networks, relying on exit polls, pronounced Gore the winner of the state about 7:48 p.m. Eastern time–12 min before the polls closed in the ten western counties in the Central time zone.

Perhaps just as problematic, the media repeatedly reported throughout the evening that the polls in Florida "close at 7:00 p.m. Eastern time which is believed to have deterred some western panhandle voters from going to the polls because they mistakenly inferred that this meant polls in their county closed at 6:00 p.m. Central time. Supporters of George W. Bush maintained throughout the recount that these media influences hurt their candidate more than it did Al Gore because the panhandle is disproportionately Republican. A poll conducted by McLaughlin and Associates (2000), a Republican polling outfit, "found that 15 % of registered non-voters did not vote because the networks declared Gore the winner," and that the effect was larger for Republican voters than for Democrat voters. A group calling itself the Committee for Honest Politics has even filed suit against the media networks for their erroneous and untimely predictions (Perrin 2001). So this question is not entirely academic, but is certainly worth academic inquiry.[2]

7.2 Voter Turnout

The literature on voter turnout is extensive. Most studies examine the impact of various demographic or institutional variables on the likelihood of voting (Matsusaka and Palda 1999; Merrifield 1993). This literature has uncovered a number of empirical regularities, but lacks a comprehensive theory of why people vote (Matsusaka 1995). In the literature, there are two main theories about what motivates individuals to vote. Shachar and Nalebuff (1999) summarize these as the (1) "pivotal-voter model" in which people vote in order to affect the outcome of the election, and (2) the "consumption model" in which voters choose to vote because the act of voting itself provides them with satisfaction.

The pivotal-voter model is founded in the rational voter hypothesis developed by Downs (1957) and extended by Tullock (1967) and Riker and Ordeshook (1968). These models explain how the voting calculus of individuals is affected by the costs

[2] We know of only two other attempts to study this particular issue. Romely (2000) finds that George W. Bush's relative performance was better in the central time zone counties in 2000 than was Bob Doles performance in the election of 1996. Using a larger panel of Florida presidential elections since 1976, Lott Jr (2005) finds an "unusual and large drop off in Republican voting rates" in Florida's western panhandle.

and benefits of voting. In these models, a key factor influencing voter turnout is the probability that the voters single vote will change the outcome of the election, (i.e., the probability that he or she is the decisive voter). While this probability is very small, even in close elections, and despite some rather heavy-handed criticisms of this rational-choice approach to voting (see, for example, Green et al. 1994), empirical research does tend to support the idea that changes in the probability of being decisive do indeed have the predicted influence on voter turnout (Matsusaka 1993; Shachar and Nalebuff 1999).

According to the pivotal-voter model, the medias early call of the election in Florida for Al Gore before the polls were closed has the potential to lower turnout because it lowers voters' expectations about their votes being decisive. The effect would be very similar to what might happen in pacific coast states in presidential elections where the popular media has already pronounced an overall winner before the polls have closed. Voters who have yet to vote in these places may logically conclude that their vote can no longer have an impact on the outcome of the election. Despite the apparently clear link with theory, the limited empirical research that has been done on this issue is not very supportive of this hypothesis. Carter JR et al. (1984), for example, finds that the 1980 early projection of victory for Ronald Reagan did not affect voter turnout in those affected Pacific coast states. In addition, in the case of the 2000 election, the first media call of the Florida election for Gore happened at 7:48 p.m. eastern time with only 12 min still remaining to vote. With such a short time remaining until the polls closed its unlikely that many voters were affected by this announcement.

We have mixed priors before looking at the issue empirically. The theory seems pretty clear but given the limited time remaining until the polls closed, and given the previous research we remain skeptical that there will be much of an effect on voter turnout from the media's call of the Florida election for Gore. If there is any impact it should be that some voters were misled into thinking the polls closed one hour earlier than they really did, rather than from any changes in their personal assessment of the probability of casting the decisive vote. The consumption model of voting most closely fits with the literature on expressive voting first proposed by Buchanan (1954), and further developed by Tullock (1971) and Brennan and Lomasky (1993). The theory of expressive voting holds that even when there is a relatively low probability of casting the decisive vote, individuals will chose to vote solely as an act of expressive behavior, often voting along ideological or moral lines for what might be considered "public minded" policies. In these models, the act of voting itself is a consumption benefit for the voter. Of course, the pivotal-voter model and the consumption model are not mutually exclusive. People may vote both to influence the outcome of the election *and* for consumption purposes. However, the presence of a consumption component to the act of voting has the potential to change significantly the possible impacts of the medias actions in this election.

First, if voters are not voting to influence the outcome but rather they are voting to enjoy consumption benefits, it makes it less likely that the medias early call of the election would discourage as many voters. However, it may imply that it could have a significant impact on the decisions voters made when casting their ballots.

With a voter thinking the election had already been decided they would be much more likely to vote for a third party candidate, potentially as an act of expression. In the 2000 election many registered Libertarian Party voters, for example, chose to vote for George W. Bush in the election because they perceived the vote between frontrunners Al Gore and George W. Bush was going to be close. In principle, if a voter prefers a third party candidate who has no real chance of winning it can make sense for them to vote for their second best choice (or more properly their first choice among the pool of viable winners) if the election is going to be close. By calling the election for Al Gore, even if voters were not dissuaded from voting (and thus it did not impact voter turnout), it still could have potentially changed the votes of those individuals who still went to the polls and voted.

Thus we are left with three empirical questions regarding these Florida counties in the central time zone: (1) was the share of the votes between Al Gore and George W. Bush altered; (2) was voter turnout lowered; and (3) was there an increase in the share of votes going for third party candidates? The next section of this paper presents our empirical estimates that address each of these questions.

7.3 Empirical Evidence

In an attempt to get directly at these issues we look at the 2000 election results in Florida using county- level data. In the regressions we employ a conventional range of demographic control variables for each county: per capita income, the percentage of the population that is black, Hispanic, age 65 or over, and college educated. We also include a dummy indicator variable that equals one if the county is one of the 10 Florida counties in the central time zone. The results of our regressions are shown in Tables 7.1 and 7.2.

7.3.1 Bush/Gore Vote Differential

Regression (1) shows results using the difference between the Gore vote share and the Bush vote share as the dependent variable.[3] In addition to the standard control variables we also include the similar index from the previous presidential election, the 1996 Clinton/Dole spread. The coefficient on the Central time zone variable is negative but statistically insignificant at standard levels of confidence. The negative coefficient indicates that, if anything, the media effect hurt Gore more than Bush, contrary to the claims made by Bush supporters and the results reported by the Republican polling outfit, however the result is statistically insignificant. We decided to explore the sensitivity of this result by adding and subtracting several variables from the regression. In only one out of the eight attempts did this variable become

[3] (Gore %-Bush %) × 100.

Table 7.1 Presidential voting in Florida's central time zone

Independent variables	Gore/Bush vote spread	Voter turnout (relative to 1996)
	(1)	(2)
Constant	−50.286 *** (6.450)	71.290 * (1.847)
Central time zone	−3.779 (1.530)	−7.819 (0.666)
Pct. black (× 100)	0.376 *** (4.054)	−0.055 (0.125)
Pct. hispanic (× 100)	0.073 (0.720)	−0.369 (0.742)
Pct. age 65 and older (× 100)	0.793 *** (5.768)	1.000 (1.463)
Pct. college educated (× 100)	0.577 *** (2.981)	0.822 (0.897)
Per capita income ($1000 s)	0.456 (1.547)	0.741 (0.548)
1996 Clinton/Dole spread	0.898) *** (13.368)	
1996 Third party vote share		
R-squared	0.896	0.137
Observations	67	67

Notes: * indicates statistical significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level. Absolute t-statistics in parentheses

significant (at the 10 % level), and in no attempt did the sign switch to being positive as would support the previous survey result. If anything, by selectively reporting only the significant result it would be possible for someone to try to show that the media call actually hurt Gore relative to Bush. Our analysis certainly seems to contradict the McLaughlin and Associates (2000) results. Based on our results we conclude that the relative share of votes for Bush and Gore in the Central time zone counties was not significantly different than in the rest of the state after controlling for our other explanatory variables. Thus, we find no evidence that the media miscues significantly altered the relative vote shares in the central time zone counties.

7.3.2 Voter Turnout

Regression (2) from Table 7.1 and (3) from 7.2 show the regression results that look directly at the issue of voter turnout. The dependent variable in (2) is the total votes cast in the county in 2000 relative to 1996.[4] We do find that turnout in 2000 relative to 1996 was lower in the central time zone counties, but the coefficient is again

[4](Total votes 2000/Total votes 1996).

Table 7.2 Presidential voting in Florida's central time zone

Independent variables	Voter turnout (% of Registered voters)	Third party vote share
	(3)	(4)
Constant	25.552 (1.235)	1.357 (1.440)
Central time zone	−1.900 (0.302)	−0.036 (0.255)
Pct. black (× 100)	0.025 (0.106)	−0.009 (1.294)
Pct. hispanic (× 100)	−0.178 (0.668)	−0.014 ** (2.231)
Pct. age 65 and Older (× 100)	0.817 ** (2.231)	0.015 (1.656)
Pct. college educated (× 100)	0.698 (1.422)	0.065 *** (6.022)
Per capita income ($1000s)	0.756 (1.043)	−0.047 ** (2.214)
1996 Clinton/Dole spread		
1996 Third party vote share		0.083 *** (3.385)
R-squared	0.257	0.598
Observations	67	67

Notes: * indicates statistical significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level. Absolute t-statistics in parentheses

statistically insignificant. The dependent variable in (3) is the total votes cast in the county in 2000 as a percent of the registered voters in the county.[5] Again we find similar results, the coefficient is negative (implying lower turnout) but the effect is far from being statistically significant at standard levels of confidence. Based upon these results we conclude that voter turnout in the central time zone counties was not significantly different than in the rest of the state.

7.3.3 Third Party Voting

Even if the media announcement did not affect voter turnout, it is still possible that it could change the outcome of the election if it caused some voters to change their votes to third party candidates when they believed the election was already decided. If so, then the issue becomes whether a larger number of Gore or Bush supporters switched to alternate third party candidates. After all, nobody is claiming that this announcement caused voters to switch their votes from Bush to Gore or vice versa.

[5] ((Total votes 2000/Registered voters) × 100).

Regression (4) in Table 7.2 explores this issue using the percentage of presidential votes in the 2000 election cast for third party presidential candidates as the dependent variable. In this regression we add another control variable, the percentage of the county voting for third party candidates in the previous 1996 presidential election. The coefficient is negative (which would imply less third party voting than expected) but again the result is not even close to being statistically significant at standard levels of confidence. Based upon this result we conclude that third party voting in the central time zone counties was not significantly different than in the rest of the state.

7.4 Conclusions

Our conclusion is that the popular media's erroneous and premature pronouncing of Gore as the winner of Florida before the polls closed in the 10 western Florida counties had no significant impact on the vote totals overall, or for third party candidates. Thus, our results contradict the widely publicized McLaughlin and Associates (2000) survey result that concluded Bush lost votes relative to Gore in these counties. If anything, it appears that Gore, not Bush, did unusually poorly in the affected panhandle counties (although the result is not statistically significant at standard levels).

Perhaps our strongest findings are that there was no impact on either voter turnout nor on third party voting in these counties. Any theory of either why George W. Bush or Al Gore was affected must be rooted in a theory of how the announcement affected one of these two variables. Even if our results of the relative Bush/Gore vote share regression had been significant, it would have been hard to explain how it was possible for this effect to be present without it being supported by the turnout or third party voting findings.

We conclude with a simple caveat. Many commentators have suggested that the statewide margin of victory for Bush was within the bounds of the margin of error in the election process. Along those lines, we cannot rule out, given the extreme closeness of the overall election, that the media influence could not have affected the outcome. An influence small enough to be statistically insignificant may still be large enough to matter in an election that was for all intents and purposes a statistical tie.

References

Brennan G, Lomasky L (1993) Democracy and decision. Cambridge University Press, Cambridge

Buchanan JM (1954) Individual choice in voting and the market. J Polit Econ 62(4):334–343

Carter JR et al (1984) Early projections and voter turnout in the 1980 presidential election. Public Choice 43(2):195–202

Downs A (1957) An economic theory of democracy. Harper and Row, New York

Green DP, Shapiro I, Shapiro I (1994) Pathologies of rational choice theory: a critique of applications in political science. Yale University Press, New Haven

Lott JR Jr (2005) The impact of early media election calls on republican voting rates in floridas western panhandle counties in 2000. Public Choice 123(3–4):349–361
Matsusaka JG (1993) Election closeness and voter turnout: evidence from California ballot propositions. Public Choice 76(4):313–334
Matsusaka JG (1995) Explaining voter turnout patterns: an information theory. Public Choice 84 (1–2):91–117
Matsusaka JG, Palda F (1999) Voter turnout: how much can we explain? Public Choice 98(3–4): 431–446
McLaughlin and Associates (2000) Panhandle poll summary: Networks' wrong Florida call for gore cost bush votes
Merrifield J (1993) The institutional and political factors that influence voter turnout. Public Choice 77(3):657–667
Perrin DB (2001) Committee for honest politics: testimony before the United States senate governmental affairs committee
Riker WH, Ordeshook PC (1968) A theory of the calculus of voting. Am Polit Sci Rev 62(01):25–42
Romely J (2000) Statistical analysis of Florida election, working paper
Shachar R, Nalebuff B (1999) Follow the leader: theory and evidence on political participation. Am Econ Rev 89(3):525–547
Tullock G (1967) Toward a mathematics of politics. University of Michigan Press, Ann Arbor
Tullock G (1971) The charity of the uncharitable. West Econ J 9(4):379–392

Chapter 8
Ballots, Bribes, and Brand-Name Political Capital

R. Morris Coats, Thomas R. Dalton and Arthur Denzau

Abstract Campaign effort is allocated towards resources that increase voter support the most for the marginal effort. A model is developed for the derived demand for investment in brand name and voter persuasion and for votes bought outright. We see that a secret ballot increases the cost of monitoring paid voters and changes the relative prices for obtaining voter support through bribery and investment in brand name capital, increasing the demand for campaign or brand-name-induced capital. Alternately, restrictions on new spending on brand-name political capital increase the demand for votes gained through direct bribery, treating and conveyance of voters to the polls, while at the same time increase the relative value of incumbency, of existing brand-name political capital. The model we develop is used to examine restrictions on spending and on brand-name investment in candidates and the effects of such spending limits (and limits on campaign contributions) on the demand for additional votes gained by vote buying, treating and conveyance of voters to the polls.

8.1 Introduction

> No man has ever placed his money corruptly without satisfying himself that the vote was cast during the last campaign without proof that "the goods were delivered": and when there is to be no proof but the word of the bribe-taker (who may have received thrice the sum to

R.M. Coats
Nicholls State University, Thibodaux, LA 70310, USA
e-mail: morris.coats@nicholls.edu

T.R. Dalton (✉)
Department of Economics, Eller College of Management, University of Arizona, Tuscon, AZ 85721, USA
e-mail: tr.dalton@gmail.com

A. Denzau
Virginia Polytechnic and State University, Blacksburg, VA 24061, USA
e-mail: artd@vt.edu

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_8

> vote for the briber's opponent), it is idle to place any trust in such a use of money. In other words, take away all interest in committing an offense, and the offense will soon disappear. (Wigmore 1889, p. 31) (quoted in Heckelman (1998))

The influence of money in elections has long been a worry of those concerned with the legitimacy of democracies. If politicians sell their decisions to the highest bidders, important "donors," then it is feared that wealthy interests and the unscrupulous would be able to use the force of the state to further enrich, empower and exploit the rest of society. Two Supreme Court cases in recent years, the Citizens Unitedand McCutcheon cases have overturned many restrictions on campaign donations established under the McCain-Feingold "Bipartisan Campaign Reform Act of 2002", leading to fears that candidates would be more beholding to wealthy donor interests.

Further, if a candidate wins elections with unfair means, such as bribing voters to vote for him, then challengers end up adopting similar means and competitive bribery escalates. Having to compete in order to win, all candidates have an incentive to pursue wealthy donors and may agree to donor wishes. High levels of political corruption bring about distrust of laws and the social order and encourages calls to overthrow the system, as we recently witnessed with the so-called "Arab Spring".

We examine the supply and demand for votes, considering that there are at least two routes to "produce" increased electoral support. The first route to additional support is through buying or inducing individual votes which is likely to have little lasting impact on voter decisions, which we will call "bribes". The second route to additional support is through persuading voters of the advantage one candidate has over opponents, by providing information, biased as it may be, which we will call campaigning or "investment in brand-name capital". The demand for by either bribed votes or campaign effort by candidates or electoral elite are both seen in terms of derived demands: demands derived from the demand for winning office, which comes with some prize. We suggest that increases in the marginal cost of gaining support for one route to votes and office increase the demand for an alternative route to votes.

8.2 Bribery and Campaign Reforms

While it seems unnecessary to discuss collective incentives to vote provided through information and other forms of mass-media voter persuasion, individual or selective incentives should be addressed. Gherghina (2013) discusses the various forms of vote buying commonly seen throughout the ages, from direct payment of money for votes, to "treating" or in-kind payment through meals, drink and tobacco. Another form of selective incentive is provided through conveyance to the polls. Here, the provided transportation reduces the voter's costs of voting. In addition, a candidate or party may pay a driver above his opportunity cost for his vote and the votes of relatives.

Though the payment and acceptance of a selective incentive, whether through money, treats or rides seems to taint both candidate and electorate, some selective incentives may prove more insidious but not so voluntary on the part of the voter. A candidate or his supporters may use their positions as employers or landlords to exploit their monopoly or monopsony positions with many voters to gain votes. Threats to discontinue contracts, where the voters have little choice, have been used in the past to gain votes. Similarly, threats of physical harm have also been used to pressure voters. Such intimidation of voters was a well-known tactic of voter suppression seen in the Jim Crow South (Wang 2012).

8.2.1 Ancient Rome

Linderski (1985), Lintott (1990), (Lintott 1999, p. 205) and Yakobson (1999) discuss electoral bribery and intimidation in the elections of the Roman republic. A key feature to elections before the Roman secret ballot was bribery. The established families and the politicians they supported gave money to "their voters" and expected "their" voters' support in return.

Candidates influenced voters with bribes, or what may be called "selective incentives" to vote, as well as with campaigning with promises of shared or collective incentives for those in a particular tribe or group. The vote contract was enforced by the law of continuous dealings and by threats of violence and financial harm, as many of the voters were often dependent upon staying in the good graces of their benefactor. On the other hand, votes of groups could also be bought somewhat collectively through the games, lavish banquets, or money given out to show the candidate's generosity, but which was not tied directly to individual voters.

Politicians borrowed heavily and ran for office as long as the returns from running for office exceeded the interest rate. As Cicero wrote, "Follow me to the Campus ... Bribery is flaring up ... this shall be a sign unto you ... Interest has gone up from four to eight per cent" (Lintott 1990, p. 8).

8.2.2 Victorian England

The United States has certainly seen its share of both electoral bribery and intimidation, but some of the most noted electoral bribery took place in the United Kingdom in the 19th century. The classical economist, John Stuart Mill, won a seat in parliament in 1865 representing Westminster. Mill is known to have refused to bribe voters, cutting his tenure in Parliament to only one term. Another major classical economist, David Ricardo, is well known to have won his seat in parliament the old-fashioned way-he bought it. Ricardo's one-term, was not from his refusal to pay voters, but rather his unfortunate death at the age of 51.

Bribery was commonplace in nineteenth century British politics (Gwyn 1962; Lloyd 1968; Gash 1977; Pinto-Duschinsky 1981). Pinto-Duschinsky (1981) remarks that the main features of British elections from 1830 to 1883 were bribery and the high costs of elections. Buying a seat was a familiar route to Parliament (Gash 1977). The practice was so common that this route to office was not even the cause of a decline in respect among Members of Parliament (MPs). As the *Westminster Review* summed it up:

> It is a painful truth that a wealthy man, known to have bribed may actually be convicted of bribery, is not the whit less respected by the majority of the House .... That a candidate spent 10,000 in the corruption of a borough will no more exclude him from the general society of the House of Commons, than a man of fashion would have been tabooed in the age of Congreve, because he had laid out a similar sum to corrupt a friend's wife. (Gwyn 1962, p. 72)

Although bribery was seldom the only consideration in an election, it was, nonetheless, a familiar one. Illegal inducements were offered to and accepted by voters regularly in most boroughs. Bribery was not the only means candidates, their agents and others had to influence votes. Employers would sometimes direct the votes of their workers. Landlords had influence over their tenants, customers over shopkeepers, clients over solicitors, and clergy over congregations. This sort of influence was universal. Less widespread, yet a still common means of influencing the outcome was violence. The banners and processions during electoral contests often led to riots (Lloyd 1968; Gash 1977).

Not all of the corruption was in the form of direct cash payments for votes. Treating, the funding by candidates of innkeepers to provide liquor, cigars, meals, and rooms to voters, was an even more common practice (Gash 1977; Pinto-Duschinsky 1981). Candidates were also expected to transport voters to the polls even in small borough districts (Gash 1977). To have a chance of winning, candidates also had to employ electors as cab drivers, messengers, canvassers, clerks, agents, and poll watchers (Pinto-Duschinsky 1981).

It was not merely their desire to serve the public, nor their thirst for power that led candidates to spend great sums for the chance of gaining a seat in Parliament. They were able to secure jobs and sinecures for friends and relatives, a great source of political payoff. Decisions about transportation infrastructure were made in small, private bills committees, putting an MP in a position to receive substantial rent (Pinto-Duschinsky 1981).

The MP's rents, especially from the private bills committees, were largely dissipated by bribery and treating. MPs attempted to increase their net rents by limiting rent dissipation by making those acts illegal, passing a number of measures to control their expenses (Gash 1977). Among the objectives of the First Reform Act (1832) was the reduction of corruption and election expenses. An act passed in 1841 made it less likely that seats gained by bribery could be retained, by allowing evidence about instances of bribery to be taken before proof of agency was established. The penalties for bribery were stiffened in 1854 (Gwyn 1962). This law discouraged bribery by overwhelming the candidates with electors who they might have to bribe. In 1868, the Corrupt Practices Act was passed in an attempt to increase the severity

of punishments and to reduce the expense of trying election petitions. The year 1872 produced two more laws to combat bribery. The first, recognizing the difficulty of having pure national elections when the local elections (in particular, elections for the returning officers) were fought with corrupt means, made fraud in municipal elections just as illegal as fraud in the Parliamentary elections (Pinto-Duschinsky 1981). The Secret Ballot Act of 1872 was passed to raise the cost of monitoring the votes of paid voters, making the practice of buying votes less certain in determining election outcomes (Gwyn 1962).

After the corrupt and monetarily expensive general election of 1880, the British Parliament passed the Corrupt and Illegal Practices (Prevention) Act in 1883. According to (Pinto-Duschinsky 1981, p. 26):

> This measure proved to be a landmark in British electoral history and has provided the basis for all subsequent legislative control of political spending in Britain. The 1883 act introduced strict limits on permitted campaign expenditures by candidates. Certain types of expenditure were banned altogether (for instance, the provision of refreshments or payments to transport voters to the polls). Moreover, it was forbidden for anyone to incur an election expense without written permission from the candidate or from his legally appointed election agent. This was intended to eliminate expenditure on a candidate's behalf but allegedly without his knowledge. Finally, tight disclosure rules set exact procedures for the presentation and the public inspection of campaign accounts.

The monetary cost per vote fell dramatically after the act (Pinto-Duschinsky 1981). But such efforts at increasing net rents were, of course, futile, because there was still a payoff for which to compete. The true costs of informing voters did not decrease. Information costs increased because either information was not transmitted at all or it was transmitted by more costly means (volunteers pounding pavement, etc.).

8.3 The Market for Voter Support

We formally model the supply and demand for votes to make our perspective explicit. We first examine the supply of voter support and then discuss the demand for votes by politicians. We then note that the demands for voter support produced through selective incentives and produced through ordinary campaign investment in brand-name capital are derived demands, derived from the demand for office.

In examining the market for voter support, we look at two different ways in which candidates or electoral elite gain voter support. One is with outright exchanges and the other is with investment in brand-name political capital through advertising and other "selling" methods, such as door-to-door selling and public speeches.

8.3.1 *Instrumental and Expressive Voting and the Supply of Votes*

The questions of why voters vote and what information voters obtain before they vote have held the attention of public choice scholars since Downs (1957). Downs (1957) shows that one component of this decision is the value of changing the group decision from a less preferred to a more preferred choice of that individual. The Downsian voting participation choice model can be stated with the following equation:

$$V = pB - C \tag{8.1}$$

where V is the citizen's decision to vote if $V > 0$ or abstain if $V < 0$; p is the probability that the voter's action at the polls changes the election's outcome; B is the benefit to the voter of changing the election outcome in his favor; and C is the cost of voting.

The only case in which a voter alters the outcome is if his action is pivotal: when one's action causes a tie or breaks a tie. Expected closeness of the race, then, affects p, the probability that a voter's choice is pivotal. Voting that is motivated by the expectation of being pivotal and changing the outcome has come to be called "instrumental voting," where the voter's action does something to change the outcome. Instrumental votes are cast expecting to make a difference.

Downs noted that the probability of a tie without one's vote or that one's vote will create a tie is so small, that in most mass elections, the probability is effectively zero, so that the pB term disappears altogether, leaving just the cost of voting in the choice equation and a negative value for V. The prediction of universal abstention, or at least exceptionally small turnout (p may become "visible" if almost no one else votes), while we note that people make it to the polls in substantially large numbers, leading to a notion known as the "paradox of voting," people vote in spite of the lack of apparent rational basis to do so.

Since $p = 0$, closeness should have no effect on voter behavior. As (Schwartz 1987, p. 118) cleverly notes, "saying that closeness increases the probability of being pivotal ... is like saying that tall men are more likely to bump their heads on the moon". Even so, closeness is sometimes found to be positively related to voter turnout (Barzel and Silberberg 1973; Silberman and Durden 1975; Settle and Abrams 1976). Coats (1984) similarly finds higher turnout in closer Victorian era British Parliamentary elections. Rallings and Thrasher (2007) note a relationship between closeness and voting in their study of modern British elections using individual-level data. Fauvelle-Aymar and François (2006) provide evidence that expected closeness increases voter participation using aggregate results from French legislative elections. Matsusaka and Palda (1993), however, find that voters in California do not seem to respond to closeness when voting on propositions, while turnout is often related to closeness in legislative elections. In a review of the empirical literature on turnout, Geys (2006) concludes that turnout is higher when campaign spending is higher, when districts are smaller and when elections are expected to be closer. Instead of

examining mass elections, Boudreaux et al. (2011) examined votes of the 110th U.S. Senate, with the body split 51–49, finding support for instrumental voting. The overall empirical results seem mixed in support of instrumental voting, but as we shall see, finding a relationship between closeness and turnout does not really imply support for instrumental voting.

In an attempt to rescue the Downsian model to resolve the paradox, Riker and Ordeshook (1968) proposed adding a *D* term to Downs' equation, a term to represent the costs of non-voting that a citizen may face, such as social pressure and a civic duty to vote, so that Eq. (8.1) reduces to:

$$V = D - C \tag{8.2}$$

Adding the *D* term to the model changes the rationale for voting from being instrumental to what has been termed "expressive". Expressive voting, then, is that votes are motivated by the desire to make a statement rather than to secure some preferred outcome. Voting becomes much like flying the controversial Confederate flag a form of protest, cheering for one's team. Copeland and Laband (2002) find that voter participation is positively related to other forms of expressiveness, such as flag flying and showing visible support for the local college team. Cebula (2008) notes the presence of emotionally charged ballot issues, such as abortion, are associated with higher turnout rates. Karahan and Shughart II (2004) show that elections with expressive elements, such as the Mississippi vote on continuing to show the Confederate flag as part of the Mississippi state flag, explain voter turnout better than those without such characteristics.

Aldrich (1993) explains that closeness affects turnout, not because of instrumental voting, but because candidates and their political organizations have a greater incentive to entice voters through influencing the voter's cost of voting or cost of abstaining–the so-called "elite mobilization theory (Key 1949). Controlling for campaign spending, Cox and Munger (1989) show evidence of closeness increasing turnout for House of Representative elections, suggesting that tighter races elicit greater campaigning effort by candidates.

The elite mobilization theory foreshadows Olson (1965), where candidates or electoral elite provide selective incentives to individuals to provide them with selective incentives to act collectively. Although Morton (1987) and Uhlaner (1989) discuss the roles of groups and group leaders in providing individual incentives to voters to entice them to vote, the reason elite group leaders provide this "group" public good (Olson 1965, 1982) remains unanswered in their work. Anderson et al. (1988) advocate an interest-group perspective of politics, with interest-group organizations creating the demand for votes in vote markets. Consistent with their view, we claim that candidates and their supporting interest-group and party organizations provide information and individual incentives as well as group benefits to voters because there is a payoff to holding office.

The supply of an individual vote is based on the reduced Riker and Ordeshook (1968) formulation of Downs' rational voter participation function we see in Eq. (8.2). Aldrich (1993) notes that for most citizens, the *D* and *C* terms are likely both small

so that only a small incentive is needed to change the sign. If both D and C are not exactly correlated and one is sometimes larger than the other, then we can rank $V's$ of individual voters, in increasing values from negative to positive so that we may aggregate across suppliers to get the market supply. We can think about this ranked V both before election competition begins, unaffected by incentives and actions a candidate might take as well as after a candidate provides competitive motivation to voters. The supply of votes, then, will be upward sloping with a quantity-of-votes intercept where the V is zero without the incentives provided by candidates. Negative "*Vs*" must be overcome with some selective incentive, d, to get the voter to the polls.

8.3.2 The Demand for Votes

The theoretical basis for understanding voter participation lies in the theory of rent seeking (Crain et al. 1988). Just as the quantity exchanged in a market is due to the interaction of supply and demand, with neither being completely explanatory on its own, voter participation is the result of the interaction between candidates (vote demanders) and voters (vote suppliers). Karahan et al. (2006) develop a model of rent-seeking demand that is dependent on the rents of office (the size of the prize) and the change in the probability of winning that is produced by additional voter support, along with the rational, expressive supply of votes by constituents. Closeness affects turnout more because of its effect on the demand for extra support than on the voters' responses to the perceived probability of affecting the outcome. The marginal benefit of an additional vote, or additional support, is based on the "prize" of the office, its value, which we assume remains constant. As the election gets closer from an expected losing position, the probability of winning increases at an increasing rate up to where the candidates are expected to be even. After the expected even position, the probability of winning continues to grow with additional support, but now at a decreasing rate. The marginal benefit of an additional vote leads to an eventual decline in the marginal expected benefit as the probability grows at a decreasing rate.

Peltzman (1976) argues that the candidate's job as political entrepreneur is to devise transfer plans to the beneficiary group which, subject to costs, garner as much support as possible (also see Lott Jr (1986, 1987)). Since straightforward tax/subsidy transfers are very transparent, a high proportion of those who are taxed (on net) will oppose the transfer, reducing the size of the total possible transfer. To reduce the "visibility" of the transfer plan, politicians devise indirect transfers, such as farm programs, defense contracts, etc., so that one dollar of net revenues (tax revenues after the politician's cut) translates into less than one dollar of benefits to the beneficiary group. Less transparent transfers entail benefits to the gainers that are less than the losses to the taxed group.

The politician receives a return for arranging these transfers, which may be monetary, though this is not necessary. Monetary payments need not be direct as a politician may to receive her fee in the form of jobs for friends and relatives. The candidate may

also take the fee in the non-monetary form of support for programs that she prefers that do not directly benefit her in her home district in terms of political support (Lott Jr 1987). Of course, being in a position to receive corrupt payments from those who may wish to do business with an elected official's government or to profit from what one might call "political insider trading," trading on the special knowledge that an elected official may possess. Political transfer fees, corrupt payments and profits from political insider trading are all sources of the prize of an elected official's office, which we will call, K. We should point out that (Lintott 1990, pp. 1–2), in discussing Roman bribery mentions that the Latin word for electoral bribery, *ambitus*, shares its root with Latin words "to canvass support" and the pursuit of political fame (what we call brand-name political capital) and political office. Rational candidates pursue an office not just because of the fame and power that winning may afford them, but because the offices provide other gains for the holders, and as is suggested, the value of the office, what one will pay or give up for the office, are connected with various gains of office: fame, power and wealth.

In competing for office, then, the candidate not only hopes to receive the prizes of the office that we lump together as K, but face the certain costs of competing for office, which we call EC. The candidate's net expected benefits from winning the election can then be stated as

$$Z = gK - EC \tag{8.3}$$

where Z is the net benefit of elected office; g is the probability of electoral victory and EC is the cost of competing for office. We should note that g is a function of the expected vote difference $v_1 v_2$, where v_1 refers to the vote for the candidate of interest, while v_2 represents the votes for the opposition, and ${}^{\delta \mathbf{q}}/_{\delta v_1} > \mathbf{0}$; ${}^{\delta^2 \mathbf{q}}/_{\delta {v_1}^2} > \mathbf{0}$ if $v_1 > v_2$; ${}^{\delta^2 \mathbf{q}}/_{\delta {v_1}^2} > \mathbf{0}$ if $v_1 < v_2$; so that ${}^{\delta \mathbf{q}}/_{\delta v_1}$ is maximized where net support, $v_1 - v_2 = 0$.

Also, EC is a function of the votes, v_1 and v_2, such that ${}^{EC}/_{\delta v_1} \geq \mathbf{0}$ and ${}^{\delta^2 EC}/_{\delta v_1 \delta v_2} \geq \mathbf{0}$, with non-decreasing marginal costs of another vote and increased support for the opposition tends to drive up the cost of acquiring another supporter for the candidate. From basic microeconomic theory, we know that the candidate (or other electoral elite) maximizes net gains, Z, where the marginal benefits of adding net support, ${}^{\delta g K}/_{\delta v_1 - v_2}$, equal the marginal costs of adding net support, ${}^{\delta EC}/_{\delta v_1 - v_2}$.

Modeling the instrumental demand of candidates, Karahan et al. (2006) extend the model proposed by Peltzman (1976) and refined by Lott Jr (1986, 1987) to show that closeness of the election can increase turnout by increasing the value of marginal support (vote). As Karahan et al. (2006) show, the closer the election becomes, the greater is the change in the probability of winning brought about by additional voter support. Since winning the election is rewarded by a valued prize, the closer the election is expected to be, the higher is the expected value of an additional vote. Candidates and their parties compete more intensively the closer the election is expected to be, because an extra vote is more valuable in a close race than in a landslide. In this model, candidates for office or their parties are the sources of the demand for votes, a demand that is derived from the demand for the rents from

political office. So, increasing the size of a majority is valuable, but with decreasing marginal benefits, much as was suggested by Stigler (1972).

When candidates can monitor votes, as is the case with no secret ballot, a candidate can purchase a vote from those who the candidate thinks are likely to vote for the opposition. If a vote is switched from the opposition will increase the candidates net support by 2. With a secret ballot, however, votes of opposition supporters are unreliable, but whether or not a supporter votes can be monitored. In these secret ballot elections, some supporters will have a negative V without some selective incentive. Such payments will increase the candidate's net support by 1 vote instead of 2. The value of votes of supporters will be lower than of opposition votes, if the vote is performed according to the contract, but not when the voter cannot be monitored. Nichter (2008) tells us that many election activities are less about paying off swing voters than getting core voters to make it to the polls-what he refers to as "turnout buying". Instead of having to monitor how someone votes, which is not possible if the ballot is truly secret, the elite need merely monitor whether voters show up to the polls. Canton and Jorrat (2003) discuss the role of party workers in Argentinian elections in monitoring voters by making sure that the poll official marks the word for "cast" next to a voter's name. Casas et al. (2014) show that this activity by party officials is associated with raising the vote count of the party worker's preferred party by anywhere from 1.7 to 7 %. Nichter (2008) discusses the role of districts and groups (e.g., unions) in the process, noting that districts and groups vary in the probability of support for one side or the other in an elections. The members of these districts and groups can be easily identified, so that candidates and other electoral elite are able to target groups and districts that have a high probability of support for their party or position. If voters from a particular group or voting precinct cast their votes for party X 85 % of the time, than paying 100 voters from this group or precinct to vote in the election will be expected to result in 85 votes for your candidate and 15 for the opposition. Similarly, Cox and Kousser (1981) note that with the introduction of the secret ballot in rural New York, there were many instances of parties paying known opposition supporters to abstain.

In the next section, we extend this basic model of supply and demand for support to one with two ways of increasing support, much like a model of derived demand for two factors of production, or alternatively, with two plants with different marginal costs. We use the model to examine changes in election institutions, such as the introduction of the secret ballot and restrictions on campaigning, such as restrictions on campaign spending. Our perspective is one of rent-seeking politicians who appreciate the costs and benefits facing potential voters and expend their resources in an attempt to influence these values to increase the probability of being elected. Our theory of supply and demand for votes and the production of additional votes with both individual incentives and mass appeals suggest that placing restrictions on one path to additional votes increases the demand for its substitute.

8.4 Bribes and Mass Campaigning: Twin Routes to Voter Support

Consider two different ways of attracting voters. One way is to give individual incentives to voters, which we refer to as bribes, while the other is to expend resources through campaign activities, such as providing information about the candidate's qualifications and positions on issues. The latter expenditures represent investments in brand-name capital, while the former buys current but not future support because bribed support must be repeatedly purchased at each election.

Brand-name capital, invested through advertising, campaign stumps, hand-shaking at public events, and going door-to-door to talk to voters in their homes, requires effort and time, as well as money. Additional increments of brand-name capital produce decreasing marginal expected support for the candidate because it is subject to normal diminishing returns. The marginal cost of increasing expected support by increasing brand-name capital, which we denote as $_j$ in our model, increases with support.[1] The increasing marginal cost assumption, however, is not crucial.

The supply of bought votes is positively sloped because the difference between C and D varies from voter to voter, and the direct payment must at least cover the difference, CD, though it may have to be even larger to overcome the voter's moral reticence to selling his or her vote. The market for bribed votes may be somewhat oligopsonistic when a limited number of candidates compete for office. The marginal cost of a bribed vote can exceed the price paid when there is monopsony power and the supply of bribed votes is positively sloped.[2] Our concern here is not so much with the price paid for bribed votes, but with the marginal cost to the candidate for a bribed vote. We denote the marginal cost of a bribed vote as β_j, which may exceed the price for bribed votes.

The marginal cost to purchase an additional vote and the marginal cost of an expected additional vote may not be the same, because a purchased vote from a voter may or may not be added to the candidate's total, as the voter may not honor the vote-bribe contract. The probability that a voter who has sold his vote will vote according to that contract depends largely upon the costs of monitoring contract performance. Increased costs of monitoring contracted behavior decrease the probability, φ, of contract performance. We assume that the probability of contract performance does not vary with the level of support, though this is merely for convenience. This implies that the marginal cost of increasing expected support by bribing voters is $^1/_\varphi * \beta_{\mathbf{j}}$, where β_j is the cost to candidate j of to bribe a voter. Of course, if candidates face an upward-sloping marginal cost of bribed votes, β_j, and $^1/_\varphi * \beta_{\mathbf{j}}$ is upward sloping as more voters are bribed.

---

[1] For consistency of notation the following should be noted: (a) subscript j refers to a candidate and (b) subscript i refers to a voter.

[2] Here we refer to simple monopsony power, not pure monopsony. For simple monopsony power to exist, all that is necessary is that as a single buyer attempts to purchase more of some input, the price of the input will be pushed up.

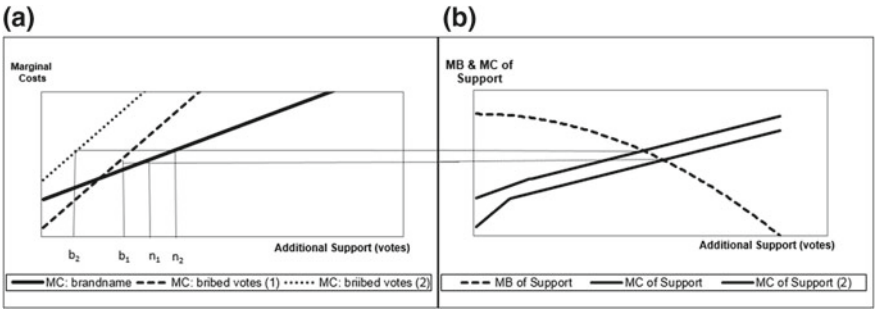

Fig. 8.1 MCs and MBs of additional support before and after the secret ballot

As mentioned, there are two basic methods of increasing support, bribery and brand-name investment, with little or no complementary relationship between the two. The two methods are combined in a way to minimize costs, analogous to cost minimization in a two-plant firm. The rule is to set the marginal costs of the two methods equal, but sum quantities of votes across the two methods. The production is allocated between the two methods such that they have the same marginal costs which are equal to the marginal benefits at the net benefit maximizing level of output. Candidates use bribery (as well as treating, conveyance to the polls, and intimidation), up to a point, to compete for support under open or non secret voting, because it is initially cheaper than alternative methods of gaining support, such as investing in additional brand name capital. The marginal cost of an additional vote, c_j, follows either $^1/_{\varphi} * \beta_{\mathbf{j}}$, or $_j$, whichever is lower.

The problem facing the candidate before the Secret Ballot was introduced is described in Fig. 8.1. The marginal cost of gaining support follows MC: Bribed votes (1) = $^{\beta j}/_{\varphi}$ until MC: bribed votes (1) is equal to the y-axis intercept of MC: brand-name (1), and then follows the horizontal sum of the two marginal cost curves. The candidate stops trying to improve his vote margin when the marginal cost of support is equal to the marginal benefits of additional support, as in Fig. 8.1b. The candidate produces additional support with the combination of bribery and brand-name inputs shown where their marginal costs are equal in panel (a) of Fig. 8.1. Of course, the alternative case, where MC: brand-name is first lower and then finally intersects with MC: bribed votes may also be the case. It matters very little for our analysis. For illustrative purposes, we assume that $^1/_{\varphi} * \beta_{\mathbf{j}}$ and $_j$ are independent resources in production. In this case, all that we need to assume is that for some candidates, the solution to the cost minimization problem is an interior one, that some candidates use a combination of bribery and brand-name capital in the open balloting or non-secret voting case.

The secret ballot, by reducing the probability of contract performance by the bribed voters, increases the marginal costs of bribery, pushing up the marginal costs of additional support. Since the marginal costs of support from bribery went up from "MC: bribed votes (1)" to "MC: bribed votes (2)," and so the quantity of bribed votes

demanded falls from b_1 to b_2 in Fig. 8.1, while the optimal amount of brand-name capital increases from n_1 to n_2.

8.5 The Importance of Voting Institutions: Ballots and Campaign Reform

Voter participation is often seen as essential for democratic choice to be considered legitimate, because otherwise, only a small group determines the political outcome-oligarchy replaces democracy. Burnham (1965, 1974) laments the decline in US voter turnout from its pinnacle in the late nineteenth century, attributing it to an "industrial-capitalist political hegemony" that alienated poor voters by giving them little choice between the two major parties. Converse (1972, 1974) and Rusk (1974) argue that institutional change causes voter behavior change. They point out that registration and the secret ballot reduced turnout by raising the cost of monitoring the vote contract for bribed votes and, as a consequence, reducing the payment of this incentive to potential voters. Cox and Kousser (1981) extend the discussion by showing that corruption continues with the advent of the secret ballot, but changes forms from turnout-inflating corruption to turnout-deflating corruption as corrupt candidates begin paying opposition voters to abstain, since abstention can be monitored. Anderson et al. (1988), however, view the secret ballot's introduction into the market for votes as a "market failure" or perhaps as interferences that reduce gains from trade arguing that purchasing votes is more efficient than trying to influence voters. While the purchase of votes might be more efficient, widespread vote buying raises questions concerning the legitimacy of the election process, just as does a lack of voter participation in the modern era.

Heckelman (1995) shows that the secret ballot did indeed lead to a substantial decline in turnout. Following Cox and Kousser (1981), Heckelman (1998) elsewhere shows that if a candidate can identify her opponents' supporters and pay them to stay away from the polls, the opponent can identify her own supporters and pay them to go to the polls, where they cast ballots for their preferred candidate once behind the voting booth's curtain, leading to higher turnout, rather than lower turnout. We should also point out that if a candidate can identify his opponents' supporters, the candidate can both pay them to stay away from the polls, but worse, can also intimidate them to stay away. Heckelman and Yates (2002) suggest incumbents of either party should favor the secret ballot, because an active bribed vote market erodes the ordinary incumbency advantage. As Pecquet et al. (2003) suggest, however, incumbents only favor secret ballots if they expect penalties against bribery to be enforced equally, unlike the case in Rome where laws against vote buying were only enforced against challengers.

As Cox and Kousser (1981) and Heckelman (1998) point out, not all bribed-vote contracts require monitoring. Similarly, Lott Jr (1986) and Lott (1987) discuss why there may be no need to be concerned with a need to punish politician shirking by

voting against their district interests in the last period as the politicians' preferences become known by their constituency and are selected largely due to their known ideology. Bender and Lott Jr (1996) show that there are no differences in voting patterns of political incumbents in their planned last terms and before their last terms. Where there is a good chance that a voter, once behind the election booth curtain, will vote against the person for whom that voter was bribed, it makes little sense to pay such a voter. However, if the voter is a member of a group known to support a particular candidate or party at exceptionally high rates, then such a voter can be counted on to vote as contracted. Such voters are all the more vulnerable to corrupt candidates if they are rather poor and more valuable to corrupt if the group to which they belong often experiences low rates of participation (Heckelman 1998, 2000).

The secret ballot, however, reduces the marginal productivity of a bribed vote in two ways. First, votes must now be purchased from supporters rather than opponents, meaning such a bribed voted increases net support only by one instead of by two, as we point out above. Secondly, each vote purchased only increases the expected vote by the probability that a voter chosen at random from the supporter group or district, rather than something much closer to one under the low costs of monitoring votes instead of turnout.

8.5.1 The Secret Ballot, the Value of Brand-Name Capital, and District Switching in Britain

Coats (1984) notes that a critical difference between British parliamentary elections and U.S. congressional elections is the lack of residency requirements for office for parliament so that candidates can represent any district without establishing residency. Some candidates switch districts after unsuccessful campaigns. Occasionally, however, an incumbent leaves his seat to run in another district, as can be seen by examining the list of candidates in the appendices of Craig (1977). Switching districts entails a loss of some brand-name capital, as the stock of information in a district about a particular candidate is a local stock and is of little value to other districts as that capital mostly resides in the district. It makes sense to give up one/s brand-name capital when there are possible differences in the costs of winning and in the probabilities of retaining a seat across different constituencies Coats and Dalton (1992b). The incumbent is less likely to switch districts as sunk brand-name capital becomes more valuable. In other contexts, this is referred to as the "sunk cost effect" and explains the reluctance of firms who have already committed to a particular technology through sunk investment to adopt a lower-cost technology (e.g. see Besanko et al. (1996)). As brand-name capital becomes more valuable in home districts, where the candidate is an incumbent, relative to new districts, the less likely the incumbent is to changing to a new district, other things being equal.

In Fig. 8.1a, we see that after the Secret Ballot is introduced the chance that a bribed voter will vote as contracted, φ, decreases. At the same time, β_j/φ, the marginal cost of gaining support by bribing another voter, increases to "MC: bribed votes (2)". We can see in Fig. 8.1b that the higher costs of gaining support through bribery increases the marginal costs of support from "MC of Support (1)" to "MC of Support (2)," which decreases the support sought through bribery and increases the support sought through increased investment in brand-name capital. The marginal costs of additional support through bribery and the marginal costs of support before the Secret Ballot are marked "MC: bribed votes (1)" and "MC of Support (1)," respectively. Note that the marginal value of brand-name capital is higher with the increase in the marginal cost of gaining extra votes through bribery.

8.5.2 Fighting the High Costs of Campaigning Through Campaign Reform and the Value of Incumbency

Incumbents have forever complained about the cost of campaigning. Competition for the rents of office largely consume the value of these rents. The candidates who anticipate higher gains from holding office bid up the cost of campaigning, through both investing in brand-name capital and through purchases of votes, a point reminiscent of Hayek's argument about why the worst get on top (Hayek 1944).[3]

Since voting is costly, Downs (1957) suggests that many voters have little incentive to vote or register to vote without subsidization (transportation, food, drink, or money). For instance, Democrats in Monroe County, New York, offer young adults beer to register to vote under the guise of a taste test (Spector and Arbelo 2004). Nichter (2008) mentions that during the 2004 election, Democratic Party workers in East St. Louis were convicted for offering various cash and in-kind bribes to increase voter participation by poor voters. Nichter (2008) goes on to mention that campaign workers and politicians have handed out coupons for food, cash payments and hiring members of large families to increase turnout, even under secret ballot rules.

Not only is voting costly, but information is as well, so Downs further suggests that voters have little incentive to become informed about elections. Just as a limit on health-care prices does not reduce the costs of health care, a limit on campaign expenditures does not reduce the cost of elections. The costs are merely passed on to voters through higher information costs. Though voters may have little incentive to become informed, candidates do have an incentive to lower information costs faced by voters through investment in brand-name capital (information provision).

A strict limit on campaign expenditures can be seen as making the marginal cost of getting another vote through brand-name investment turn vertical at the point of

[3]Coats and Dalton (1992a, b) and Yen et al. (1992) have discussed brand-name investment by candidates and how brand-name investment by candidates reduces the entry of rivals. Coats et al. (2014) discuss corruption in democracies relative to Hayek's chapter.

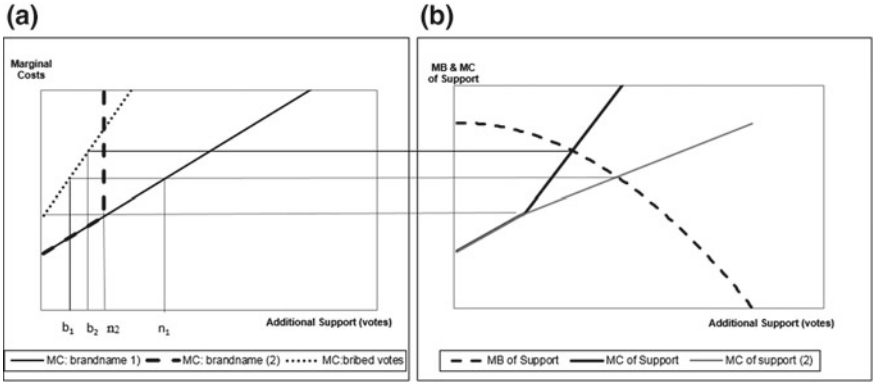

Fig. 8.2 MCs and MBs of additional support after the secret ballot and before and after campaign spending reform

the limit, as we see with the heavy dashed line in Fig. 8.2a.[4] While that marginal cost is infinitely high, there is still another route to gaining voters, bribery. In Fig. 8.2, we show the impact of introducing limits on campaign expenditures on a candidate's choice of election strategy. The dotted line marked "MC: bribed votes" in Fig. 8.2a, and the darker solid line in Fig. 8.2b marked "MC of Support (2)" show their positions after the Secret Ballot is introduced. We can clearly see from the figure that expenditure limitations increase the demand for bribed votes, with the quantity demanded of bribed votes increasing from b_1 to b_2, while the quantity demanded of brand-name capital decreasing from n_1 to n_2.

An objection that could be raised to the result shown in Fig. 8.2 is that limits on campaign spending will limit all spending equally. However, modern candidates mostly run using money raised by contributors, and there is a natural limit to that source of funds. Or another way of thinking about this is that as a candidate spends more time and effort raising funds he has less time to attend to his current office and less time to spend campaigning – there is a rising marginal cost to increased campaign spending. With limits on campaign spending the candidate could ask that the money be given to a separate get-out-the-vote organization instead of having contributors give money to the candidate directly. This organization, like candidates in nineteenth century Britain, could provide rides to polls, food, beer, smokes or even money to potential voters. The form of the get-out-the-vote effort depends on which means is more effective at increasing votes. While bribing voters is illegal, spending on voter bribery and often questionable get-out-the-vote efforts are much more difficult to monitor than spending on advertising or campaign rallies.

In recent years in the United States, limits on contributions from individuals and from organizations have come to the forefront. McCain-Feingold established some

[4]The kink in he marginal cost of support through brand name investment and the intercept of the marginal cost of support through bribery are at the same marginal cost for convenience, to avoid extra kinks and complicated graphics.

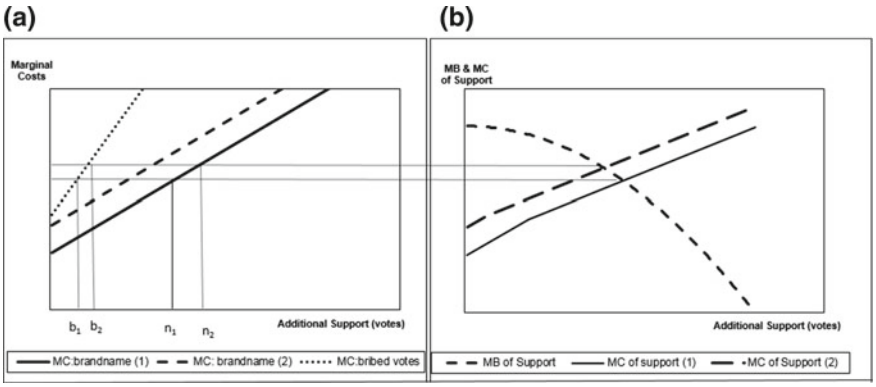

Fig. 8.3 Limits on campaign donations raise the MCs of campaigning and the derived demand for votes

limits. In both *Citizens United v. Federal Election Commission* and *McCutcheon v. Federal Election Commission*, the United States Supreme Court overturned many contribution limits. One of the reasons given for such limits is that campaign contributions are seen as corrupting.

We suggest that such limits on contributions as well as limits on spending can increase the demand for bribed votes or what might be called "get-out-the-vote" efforts of political parties and their proxies. Contribution limits lead to candidates having to go after more individual contributors, raising the politician's costs of brand-name capital. The rise in the costs of campaigning or in brand-name capital, by reducing what might be contributed by any donor, raises the demand for bribed votes and can be seen as the result of raising the marginal costs of campaigning. As we see, this rise in the demand for bribes leads to an increase in the quantity demanded of bribed votes in Fig. 8.3a from b_1 to b_2. We should also note that often, many organizations are involved in turnout buying with only lose ties to a candidate or a party, making it very difficult to monitor and very difficult to tie to a party or candidate to get-out-the-vote effort and such expenditures would be very difficult to limit.

Of course, not only does restricting new investment in brand-name political capital increase the demand for what we have called bribed votes, or turnout buying, but it should be immediately clear that restrictions on investment in brand-name capital increases the value of existing brand-name capital, increasing the value of incumbency. Campaign spending and giving restrictions amount to a form of protectionism not unlike that fostered under mercantilism. The American humorist and pundit, P.J. O'Rourke (2005) labeled the McCain-Feingold campaign-finance reform as "Incumbent-Protection Acts," and with good cause.

8.6 Conclusions

This chapter develops a rational exchange model to answer the question: "what effect can we expect in candidate competition from changes in election institutions?" e began by exploring expressive versus instrumental rationales for voting, finding that instrumental motivations alone were unlikely to account for voting and that an expressive rationale may account for voting in some cases and for some, it may be insufficient to get a voter to the polls, giving us an upward sloping supply of votes. We saw that the demand for votes was a reflection of the value of the office to candidates. The demand for votes can be explained as largely instrumental, based on the declining change in the probability of winning.

Then we noted that the demand for any one routed to office, through vote buying or investing in brand-name capital, can be modeled as production in across two plants, so that increasing the cost of in one plant provides incentives to increase utilization of the other plant. We find that the secret ballot made it more expensive to compete for a seat, instead of reducing campaign costs as some had hoped. British MPs found the election following their introduction of the secret ballot, the general election of 1880, to be their most expensive ever, leading to campaign expenditure limitations in 1883 (Pinto-Duschinsky 1981). The Secret Ballot Act, by raising the cost of securing additional support through bribery or other forms of selective incentives, increased the demand for political brand-name capital.

In cases where monitoring of voter behavior is indirect, the use of selective incentives (Olson 1965) can have a significant effect on candidate competition. Last-period problems in contract performance by individuals are less likely to occur when the performer's preferences are discovered to be consistent with the requested performance. That is, a candidate or his organization can be fairly certain that certain voters will support that candidate and feel safe that the voters who get rides to the polls will not support the opposing candidate when they get behind the curtain. This is especially true when it is known that voters with certain characteristics, such as class, gender or ethnicity, have a very strong tendency to vote a particular way.

Downs (1957) argues that no voter has an incentive to gather information about candidates and election issues, and the burden of this cost is largely internalized by the candidates. However, the information obviously has some bias, and surely voters take those biases into account to some degree.

Legislation aimed at controlling campaign expenses seems to do the opposite of what is intended. The burden of this increase in costs falls somewhat on the voters, either as information that is not communicated to the voters or as income denied to voters by certain selective incentives. Limiting campaign spending, whether directly or indirectly by limiting campaign contributions, raises campaign costs, increasing the demand for selective incentives, as selective incentives can be hidden from public scrutiny by having a third party act on a candidate's behalf. While we may attempt to reduce campaign spending with limits, as long as office-holding is valuable, campaign costs are bid up to that value. Reducing campaign costs and the influence of money in elections requires reducing the value or rents of office holding.

References

Aldrich JH (1993) Rational choice and turnout. Am J Polit Sci 37(1):246–278
Anderson GM, Tollison RD et al (1988) Democracy, interest groups, and the price of votes. Cato J 8(1):53–70
Barzel Y, Silberberg E (1973) Is the act of voting rational? Public Choice 16(1):51–58
Bender B, Lott JR Jr (1996) Legislator voting and shirking: a critical review of the literature. Public Choice 87(1–2):67–100
Besanko D, Dranove D, Shanley M (1996) Economics of strategy. Wiley, New Jersey
Boudreaux CJ, Coats RM, Walia B (2011) Voting and abstaining in the US Senate: Mr. Downs goes to Washington. Southern Bus Econ J 34(1-2):55–72
Burnham WD (1965) The changing shape of the american political universe. Am Polit Sci Rev 59(1):7–28
Burnham WD (1974) Theory and voting research: some reflections on converse's "change in the American electorate". Am Polit Sci Rev 68(3):1002–1023
Canton D, Jorrat JR (2003) Abstention in Argentine presidential elections, 1983–99. Lat Am Res Rev 38(1):187–201
Casas A, Díaz G, Trindade A et al (2014) Who monitors the monitor? Effect of party observers on electoral outcomes. Technical report
Cebula RJ (2008) Influence of the number of statewide referenda involving emotionally-charged issues on voter turnout, 2006. Atl Econ J 36(4):383–393
Coats RM (1984) Voter participation in nineteenth century British parlimentary elections. PhD thesis, Virginia Polytechnic Institute and State University
Coats RM, Dalton TR (1992a) A note on the cost of standing for the British Parliament: 1852–1880. Legis Stud Quart pp 585–593
Coats RM, Dalton TR (1992b) Entry barriers in politics and uncontested elections. J Public Econ 49(1):75–90
Coats RM, Karahan G, Shughart W (2014) Why the worst get on top: corruption in democracies
Converse PE (1972) Change in the american electorate. In: Campbell A, Converse PE (eds) The human meaning of scoial change. Russell Sage Foundation, New York, pp 263–338
Converse PE (1974) Comment on Burnham's "theory and voting research". Am Polit Sci Rev 68(3):1024–1027
Copeland C, Laband DN (2002) Expressiveness and voting. Public Choice 110(3–4):351–363
Cox GW, Kousser JM (1981) Turnout and rural corruption: New York as a test case. Am J Polit Sci 25(4):646–663
Cox GW, Munger MC (1989) Closeness, expenditures, and turnout in the 1982 US House elections. Am Polit Sci Rev 83(1):217–231
Craig F (1977) British parliamentary election results 1832–85
Crain WM, Shughart II WF, Tollison RD (1988) Voters as investors: a rent-seeking resolution of the paradox of voting. In: The political economy of rent-seeking. Springer, Heidelberg, pp 241–249
Downs A (1957) An economic theory of democracy. New York
Fauvelle-Aymar C, François A (2006) The impact of closeness on turnout: An empirical relation based on a study of a two-round ballot. Public Choice 127(3–4):461–483
Gash N (1977) Politics in the age of peel: a study in the technique of parliamentary representation, 1830–1850. Harvester Press, Sussex
Geys B (2006) Explaining voter turnout: a review of aggregate-level research. Electr Stud 25(4):637–663
Gherghina S (2013) Going for a safe vote: electoral bribes in post-communist romania. Debatte: J Contemp Cent and East Europe 21(2–3):143–164
Gwyn WB (1962) Democracy and the cost of politics in Britain. Althalone Press, London
Hayek FA (1944) Why the worst get on top. In: The road to Serfdom, University of Chicago Press, Chicago, Chap 10, pp 134–152

Heckelman JC (1995) The effect of the secret ballot on voter turnout rates. Public Choice 82(1–2):107–124
Heckelman JC (1998) Bribing voters without verification. Soc Sci J 35(3):435–443
Heckelman JC (2000) Revisiting the relationship between secret ballots and turnout a new test of two legal-institutional theories. Am Polit Quart 28(2):194–215
Heckelman JC, Yates AJ (2002) Incumbency preservation through electoral legislation: the case of the secret ballot. Econ Gov 3(1):47–57
Karahan GR, Shughart WF II (2004) Under two flags: symbolic voting in the state of Mississippi. Public Choice 118(1–2):105–124
Karahan GR, Coats RM, Shughart WF II (2006) Corrupt political jurisdictions and voter participation. Public Choice 126(1–2):87–106
Key V (1949) Southern politics in state and nation. University of Tennessee Press
Linderski J (1985) Buying the vote: electoral corruption in the late republic. Anc. World 11(3–4):87–94
Lintott A (1990) Electoral bribery in the Roman republic. J Roman Stud 80:1–16
Lintott AW (1999) Violence in republican Rome. Oxford University Press, Cambridge
Lloyd TO (1968) The general election of 1880. Oxford University Press, Cambridge
Lott JR (1987) The effect of nontransferable property rights on the efficiency of political markets: some evidence. J Public Econ 32(2):231–246
Lott JR Jr (1986) Brand names and barriers to entry in political markets. Public Choice 51(1):87–92
Lott JR Jr (1987) Political cheating. Public Choice 52(2):169–186
Matsusaka JG, Palda F (1993) The downsian voter meets the ecological fallacy. Public Choice 77(4):855–878
Morton RB (1987) A group majority voting model of public good provision. Soc Choice Welf 4(2):117–131
Nichter S (2008) Vote buying or turnout buying? machine politics and the secret ballot. Am Polit Sci Rev 102(1):19–31
Olson M (1965) The logic of collective action. Harvard University Press, Cambridge
Olson M (1982) The logic in the rise and decline of nations. In: Baker SH, Elliott CS (eds) Readings in public sector economics, DC Heath & Company, Lexington, pp 193–206
O'Rourke P (2005) Incombent-protection acts. Atl 295(4):36
Pecquet GM, Coats RM, Dalton T (2003) Roman elections, vote buying, and campaign reform legislation
Peltzman S (1976) Toward a more general theory of regulation. J Law Econ 19(2):211–240
Pinto-Duschinsky M (1981) British political finance, 1830–1980. American Enterprise Institute Press, Washington DC
Rallings C, Thrasher M (2007) The turnout gapand the costs of voting-a comparison of participation at the 2001 general and 2002 local elections in England. Public Choice 131(3–4):333–344
Riker WH, Ordeshook PC (1968) A theory of the calculus of voting. Am Polit Sci Rev 62(1):25–42
Rusk JG (1974) Comment: the American electoral universe: speculation and evidence. Am Polit Sci Rev 68(3):1028–1049
Schwartz T (1987) Your vote counts on account of the way it is counted: an institutional solution to the paradox of not voting. Public Choice 54(2):101–121
Settle RF, Abrams BA (1976) The determinants of voter participation: a more general model. Public Choice 27(1):81–89
Silberman J, Durden G (1975) The rational behavior theory of voter participation. Public Choice 23(1):101–108
Spector J, Arbelo E (2004) Free beer if you register to vote. Democrat and Chronicle 18
Stigler GJ (1972) Economic competition and political competition. Public Choice 13(1):91–106
Uhlaner CJ (1989) Rational turnout: the neglected role of groups. Am J Polit Sci 33(2):390–422
Wang TA (2012) The politics of voter suppression: defending and expanding Americans' right to vote. Cornell University Press, Ithica

Wigmore JH (1889) The Australian ballot system as embodied in the legislation of various countries. Charles C Soule, Boston
Yakobson A (1999) Elections and electioneering in Rome: a study in the political system of the late Republic. Franz Steiner Verlag, Stuttgart
Yen ST, Coats RM, Dalton TR (1992) Brand-name investment of candidates and district homogeneity: An ordinal response model. South Econ J 58(4):988–1001

Chapter 9
The Effect of Inter-School District Competition on Student Achievement: The Role of Long-Standing State Policies Prohibiting the Formation of New School Districts

Katie Sherron and Lawrence W. Kenny

Abstract Efforts to estimate the effect of having more school districts (i.e., having more competition among school districts) have been hampered by the difficulty of finding a good instrument for the number of school districts. We identify 9 states in which the state requires that the school districts be county-wide or state-wide; these laws have been in place for almost 7 decades. In states with no restrictions on the formation of school districts, larger metropolitan areas have more school districts, and thus more inter-district competition. As expected, student test scores are higher in larger metro areas that do not require county-wide or state-wide districts. On the other hand, test scores are no higher in large metro areas than in small metro areas in states that prohibit any rise in the number of districts as the metro area grows.

9.1 Introduction

Most economists believe that greater competition provides each of the players with a stronger incentive to be more efficient. Accordingly, greater competition in primary and secondary education should make that sector perform more efficiently. For this reason, many advocates of reform of the educational system support programs that would bring about more competition. School vouchers increase enrollment in private schools, which creates greater competition between public schools and private schools. Charter schools operate under fewer regulations than traditional public schools and compete with nearby traditional public schools. We study the impact of a third form of competition - having more competition between public school

K. Sherron
Florida State University, Tallahassee, FL 32306, USA
e-mail: ksherron@fsu.edu

L.W. Kenny (✉)
University of Florida, Gainesville, FL 32611, USA
e-mail: kenny@ufl.edu

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_9

districts - on the efficiency of public school districts.[1] Having more school districts forces school districts to compete with one another for students, giving the districts a greater incentive to use their limited resources efficiently. Thus, school districts in school markets with more school districts should have higher test scores, holding parental inputs and school inputs constant. The evidence is inconclusive in the small empirical literature testing this hypothesis.[2]

Borland and Howsen (1992) and Zanzig (1997) found that student test scores were higher in areas with more inter-district competition, but these papers do not address issues raised by the endogeneity of the number of school districts. That is, the number of school districts (and therefore the amount of competition) and academic performance are determined simultaneously. For example, imagine two neighboring districts, one with strong performance in math the other with strong English skills. If these districts decide to merge, it is conceivable that average academic achievement would increase. This is an example of student learning causing the number of districts. On the other hand, the number of districts can affect learning through inter-district competition and scale efficiencies.

Papers that specifically deal with the endogenous nature of the number of public school districts and consider the impact on student learning are described below. Hoxby (2000) uses the number of rivers in a metropolitan area to instrument for the number of districts and finds that in markets with greater choice of school districts, students in public schools perform better academically. Rothstein (2007) reports that Hoxbys instrument is very difficult to code consistently and finds that in only a few regressions is there any evidence that more inter-district competition raises test scores.

Another paper by Rothstein (2007) uses the number of districts in 1942 to instrument for the number of districts in 1997. The author utilizes SAT scores and data from the National Educational Longitudinal Survey 1998 and concludes that increasing school district choice does not improve academic performance of students.

We contribute to the literature by developing two empirical specifications that do not require using potentially flawed instrumental variable procedures to deal with the endogeneity of the number of school districts in many states. We identify 9 states in which the state requires that school districts be county-wide or state-wide, or have some other long-lasting configuration. With the exception of Nevada, the law mandating no more than one district in the county or state or a similar restriction has been in place since at least 1947, covering 67 years. Nevada's current configuration of 17 school districts for its independent city and 16 counties was established in 1954. Since these laws have been in place for more than six decades, we consider them to be exogenous to student achievement in 1997.

---

[1] Note also that having more school districts makes it possible to have a more complete sorting of families according to their desired level of school quality. See Tiebout (1956).

[2] In related research, Hoxby (2000) finds that per-pupil spending is lower in metropolitan areas with greater choice. Kenny and Schmidt (1994) find that state-wide and county-wide districts are less efficient than districts where competition is present. The laws that prevent within-county competition leads to a higher cost of education. As a result, these monopoly-like school districts spend an additional $866.5 million each year (1992 dollars).

It also is important to recognize that metro areas with more people tend to have many more unrestricted school districts than smaller metro areas.[3] Regression estimates from Fisher and Wassmer (1998) indicate that unrestricted metropolitan statistical areas (MSAs) with populations of a quarter million, half a million, and one million people have (respectively) three, five and ten times the number of districts as an MSA with a population of only one hundred thousand residents, all else constant. A larger MSA typically has both bigger school districts (allowing each district to take advantage of scale economies but strengthening the teachers union) and more districts (allowing for more competition between school districts and for better Tiebout-sorting of preferred school qualities into provided school qualities).

Small counties like St. James, Louisiana have few students and would most likely have only one school district even if the law permitted them to have many more.[4] Since the law mandating a county-wide district is not binding in a small county, the law is expected to have no impact on student learning in small counties. As the size of the educational market increases, eventually the restriction that there be no more than one district in each county becomes binding. Further increases in the size of the market (1) in unfettered states typically lead to more school districts, and thus more competition between school districts, and (2) in limited-district states are accompanied by no change in inter-district competition. So student test scores in states that require county-wide or state-wide districts should be lower than student test scores in states that do not restrict competition between school districts.

Furthermore, the disparity between test scores in limited-district states and test scores in unfettered states should be greater in larger educational markets than in smaller markets. This is because there is more competition among unfettered districts in larger markets than in smaller markets. These predictions are tested with exogenous variables that measure the size of the market and whether the state requires county-wide or state-wide school districts. No instrumental variables (IV) procedures are used in these specifications. Both predictions are strongly supported in the empirical analysis that follows. We find that students learn more in unrestricted large metro areas in which there are typically many districts than in large metro areas that have stringent limits on the number of school districts. Also, the gap between test scores in unrestricted states and test scores in states with a restriction that districts be county-wide is greater in larger school markets.

Of course, a metro area with 500,000 families can be broken into 100 school districts with 5,000 families in each district, 50 school districts with 10,000 families in each district, etc. A closely related literature (see Rose and Sonstelie (2010); Brunner and Squires (2013); Lott and Kenny (2013)) examines the effects of having larger (but fewer) school districts. Parents have less influence in large school districts than in small school districts, presumably because the free riding of some parents on the efforts of other parents is more of a problem in large school districts. Free

[3] See Ross et al. (2014) for an interesting study of the alignment of municipal boundaries and school district boundaries on class size.

[4] The county of St. James, Louisiana has a population of nearly 21,000 land area of 246 mi^2 and population density of 84 people/mi^2.

riding by teachers does not appear to rise as rapidly as free riding by parents, as school districts get larger. If there are economies of scale, students learn more in larger school districts. But the greater influence of teacher unions in larger school districts makes it more difficult to fire incompetent teachers and to make numerous other decisions that could make schools more efficient. The evidence in this small literature suggests that students learn less in larger school districts.

9.2 States Limiting the Formation of School Districts

In Table 9.1 nine states are listed that require that school districts be county-wide or state-wide, or have some other long-lasting configuration; the number of (state-limited) districts in these states and the District of Columbia for 1947, 1950, 1960, 1970, 1980, and 1997 also are reported.[5] Florida, Maryland, Nevada, Virginia, and West Virginia all have laws in place that mandate that each county in the state have one county-wide school district. The District of Columbia has had a single federal district-wide public school district for decades. Hawaii has had a single state-wide public school district since it became a state in 1959. The number of school districts has also remained relatively constant for the past half century in Georgia, Louisiana, and Utah.

Between 1947 and the late 1970s, the number of public school districts in Georgia's 159 counties has been as great as 198 and as few as 186; since the late 1970s the number of school districts in Georgia has remained at 187. Louisiana has 64 counties and had 67 public school districts in the 3 decades after World War II; since then it has had 66 school districts. Finally, Utah has had 40 public school districts in 29 counties since before 1947. None of the aforementioned states have replaced their county-wide or state-wide school districts with systems that allow multiple districts in a county. With the exception of Nevada, the law mandating no more than one district in the county or state or a similar restriction has been in place since at least 1947, covering 67 years.[6] Nevada's configuration of 17 school districts for its independent city and 16 counties was established in 1954. Throughout the rest of the paper these 9 states and the District of Columbia will be collectively called *LimitedDistrictStates*.[7]

Note that the county-wide school district states had only one third the number of school districts that would have been predicted for them (Kenny and Schmidt 1994). The typical state with county-wide limits had 164 fewer districts than unfettered states.

[5]The earliest data available are from 1947 and obtained from the Council of State Governments (2002).

[6]Nevada had 196 school districts in 1952 but due to a funding crisis, consolidated to 17 districts after 1954 (Strang 1987).

[7]Admittedly, this is a slight misnomer, since DC is a federal district and not a state.

Table 9.1 Characteristics of states with an exogenous number of school districts

State	Districts in 1947	Districts in 1950	Districts in 1960	Districts in 1970	Districts in 1980	Districts in 1997	Mandate
DC	1	1	1	1	1	1	Yes
Florida	67	67	67	67	67	67	Yes
Georgia	189	186	198	190	187	187	No
Hawaii	NA	NA	1	1	1	1	Yes
Louisiana	67	67	67	66	66	66	Yes
Maryland	24	24	24	24	24	24	Yes
Nevada	222	196	17	17	17	17	Yes
Utah	40	40	40	40	40	40	Yes
Virginia	125	127	131	134	141	141	Yes
West Virginia	55	55	55	55	55	55	Yes

Notes Data from 1947 from Council of State Governments (2002). Data from 1950–1980 is from Kenny and Schmidt (1994). Data for 1997 is from Snyder (1998). All states but Georgia have a law in place preventing the number of districts from changing. Florida consolidated from 720 to 67 school districts in 1947. Nevada consolidated from 196 to 17 districts in 1952. In Virginia, the number of school districts equals the number of counties plus the number of independent cities; changes in the number of school districts are due to changes in the number of independent cities

9.3 Empirical Model

Two empirical specifications are used to assess the impact of limiting competition on student learning. The first specification is relatively simple.

$$\begin{aligned} Learning_{icsv} = \alpha_0 + \alpha_1 LimitedDistrictState_s + \alpha_2 MarketSize_c \\ + \alpha_3 LimitedDistrictState_s X MarketSize_c + StudentInputs_i \chi \\ + ParentalInputs_i \lambda + SchoolInputs_v \theta \end{aligned} \tag{9.1}$$

The subscripts depict the learning by student i attending school v in county c of state s. The variable *LimitedDistrictState* is a dummy that equals one in the states where the number of public school districts is exogenous. In the remaining states, where residents' preference for education, income heterogeneity, population, population density, and land area determine the number of school districts, the variable *LimitedDistrictState* equals zero. We define the size of a public school market as the population living in the area.

In states with no restriction on the formation on the number of school districts, the number of school districts in the MSA is greater in large metro areas than in small metro areas. But in limited-district states there is no increase in the number of school districts as the market gets larger. The difference between the number of districts in states with a county-wide mandate and the number of districts in states with little to

no restrictions on the formation of school is greater in larger school markets. Thus, the disparity in the level of competition, and consequently in test scores, between unfettered states and limited-district states is expected to be greater in larger markets. The coefficient on the interaction term (α_3) is predicted to be negative.

$$\frac{\partial Learning}{\partial LimitedDistrictState} = \alpha_1 + \alpha_3 MarketSize \tag{9.2}$$

In the simple model described above, the effect of having a larger market on student achievement (i.e., the slope) does not depend on the size of the market. The second empirical specification allows the slope to differ between small markets and larger markets. For example, economies of scale may be more important in small markets. This more complex functional form is a spline function.[8]

$$\begin{aligned} Learning_{icsv} = \beta_0 + \beta_1 LimitedDistrictState_s + \beta_2 MarketSizeSpline1_c \\ + \beta_3 MarketSizeSpline2_c + \beta 4 LimitedDistrictState_s \times MarketSizeSpline2_c \\ + StudentInputs_{i\chi} + ParentalInputs_{i\lambda} + School\ Inputs_{v\theta} \end{aligned} \tag{9.3}$$

The spline is a continuous piecewise linear function with a discrete change in the marginal effect (slope) at a breakpoint. We search for the appropriate breakpoint by finding the market size breakpoint, M*, which gives the best fit. The spline function enables us to better describe the relationship between test scores and the size of the school market.

As before, the slope in the larger markets is allowed to differ between unfettered states and limited-district states. This effect from Eq. 9.4 corresponds to Eq. 9.2 in the simpler specification. This is captured by the interaction term *LimitedDistrictState* × *MarketSizeSpline*2. The variable takes a value of zero when the state or federal district has no restriction on the number of school districts. This variable also equals zero when *LimitedDistState* = 1 and *MarketSize* is "small." In small markets, the restriction that there are only a few school districts is not binding. The mandate for a small number of districts will have little or no impact on learning. However, when *LimitedDistState* = 1 and *MarketSize* is "large," the restriction is binding.

$$\frac{\partial Learning}{\partial LimitedDistrictState} = \beta_1 + \beta_4 MarketSizeSpline2 \tag{9.4}$$

[8]The spline functions are defined as follows:

$$MarketSizeSpline1 = \begin{pmatrix} MarketSize\ if\ MarketSize \leq M^* \\ M^*\ if\ MarketSize > M^* \end{pmatrix}$$

$$MarketSizeSpline2 = \begin{pmatrix} 0\ if\ \ MarketSize \leq M^* \\ MarketSize\text{-}M*\ \ if\ Markesize > M^* \end{pmatrix}$$

.

In states without laws limiting competition, the number of school districts within that market also tends to increase as the school market gets larger. Residents are able to better take advantage of economies of scale and to better sort themselves according to their preference for school quality. The number of public school districts increases as the size of the market increases because people are sorting themselves more completely according to their preference for education.

On the other hand, in limited-district states there is no increase in the number of school districts as the market gets larger. The disparity in competition between unrestricted states and limited-district states is greater in larger markets. Thus, the difference in test scores between unfettered states and limited-district states is greater in larger markets. Equivalently, the coefficient β_4 is predicted to be negative.

If Eq. 9.4 is evaluated at the break point ($MarketSizeSpline2 = 0$), where county-wide restrictions are just becoming binding, then the effect of having limited-district choices equals β_1. We expect that the coefficient on *LimitedDistrictState*, β_1, will not be statistically distinguishable from zero.

9.4 Data

What is the "area" that best describes the municipal marketplace? In a typical metropolitan statistical area (MSA) workers commute to a central business district, and households sort in a Tiebout marketplace for government services, which are provided by local governments. The MSA is a better measure of the extent of the school market than the county, since families rarely live in a different MSA from that in which the parents work but often work in a different county from the county in which the parents send their kids to school.[9] The size of the market for education seems to be ideally measured by the number of people living in the MSA.

The only problem with using the MSA is that some people do not live in a metro area; in our sample 1,148 (22 %) of the students do not live in an MSA. An alternative measure is the number of people living in the county. But most MSAs are comprised of several counties. The county population measure cannot distinguish between a metro area consisting of one county with 100,000 people and a second metro area consisting of three counties, each with 100,000 inhabitants. Thus the extent of the municipal marketplace is not as well described by the number of people living in a county in an MSA.

The variable $MarketSize$ is taken from the 2000 census and is measured in two different ways: (1) number of people living in the MSA, and (2) the number of people living in the county. Summary statistics for these variables are presented in Table 9.2. We match each $MarketSize$ variable with information from the 1997 National Longitudinal Survey of Youth (NLSY97).

All individual and school level data that we use come from the NLSY97 survey data. The NLSY97 follows a sample of 8,984 students in their transition from school

[9] See Hoxby (2000), Zanzig (1997), and Rothstein (2007).

Table 9.2 $MarketSize$ variables summary statistics

Variable	Mean	Std. dev.	Min	Max
MSA population	818,545	1,987,796	57,813	21,199,870
County population	89,596	292,462	67	9,519,338

to work beginning in 1997. These data are weighted to be representative of the U.S. population. During the first round of interviews, 6,039 students were given the Peabody Individual Achievement Test (PIAT) Math Assessment. The variable Learning is the students score on the PIAT. This test is a computer administered exam that has been found to reliably measure mathematic achievement (Markwardt 1998).

The PIAT begins with a few questions of varying difficulty, based on the students grade level. The computer then presents more complicated questions if a students prior answers were correct and easier questions if his prior answers were wrong. Raw exam scores are reported on a scale of zero to one hundred. As expected, students in higher grades receive higher scores. To better track a student's relative progress over time and to be able to compare students across grades, we normalize the raw score with respect to grade.[10] That is, we take each students raw test score, subtract the mean score for that students grade level and then divide by the standard deviation for that students grade level. Therefore, within grade level, the learning measure that we use has a mean of zero and standard deviation of one.

The vector $StudentInputs$ is measured using three dummy variables (Black, Hispanic, and female), and the student's age. The vector $ParentalInputs$ consists of six different variables: mother's years of education, father's years of education, the number of children under eighteen living at home, an indicator equal to one if the biological father lives at home, an indicator equal to one if a family member volunteered at school, and an indicator equal to one if a family member attended a PTA meeting in the last month. In 1997, one parent from each NLSY household was interviewed. About ten percent of students surveyed had no parent interview recorded and were excluded from my analysis. The students biological mother was the respondent for 80 % of completed parent interviews. The parent was asked questions about her education level and about the students biological fathers education level and if he still lived at home. If the education level of a biological parent is missing, we use the resident parents education instead. The interviewer also asked if the biological mother or another family member attended PTA meetings or volunteered at the students school (correlation coefficient of 0.29).

The vector $SchoolInputs$ is made up of three different pieces of school and individual level information. To gather information on the students' daily life, NLSY Interviewers asked, "Have you ever been threatened at school?" This question may

[10]We also used raw scores as the dependent variable in all regressions, including grade dummies. We do not report these results here, as they were very similar.

Table 9.3 Variable summary statistics

Variable	Mean	Std. dev.	Min	Max
Dependent variable				
Normalized test score	0.000	1.000	−4.299	2.721
Student inputs				
Female	0.472	0.499	0	1
Black	0.282	0.450	0	1
Hispanic	0.200	0.400	0	1
Age	13.63	1.17	12	17
Parent inputs				
Mother's years of education	12.47	2.89	1	20
Father's years of education	12.48	3.19	1	20
Live with biological father	0.531	0.499	0	1
Volunteer at student's school	0.641	0.722	0	2
Attend PTA meetings	0.933	0.745	0	2
Number of children under 18 living in household	2.52	1.27	1	12
School inputs				
Threatened	0.224	0.42	0	1
Student teacher ratio	17.19	5.31	10	26
School size	819	359	50	1250

also be viewed as indicator of student effort. The other gauges of school quality are the student teacher ratio and school size, which NLSY interviewers obtained directly from the school. The final sample that we use consists of 710 students who live in areas where state law restricts the number of school districts, and 4,587 students from states without a potentially restrictive mandate. Summary statistics are presented in Table 9.3.

9.5 Results

The MSA $MarketSize$ and *LimitedDistrictState* dummy results for the simpler interaction specification (Eq. 9.1) and for the spline specification (Eq. 9.3) are reported in Table 9.4. In the spline we find that mandating only one school district per county does not negatively impact test scores until the MSA population reaches approximately 91,000. When the market is defined as the county, results indicate that test scores are no different in unfettered states and countywide districts until the county population reaches 250,000 residents.

The strong prediction that the gap in test scores between students in states that restrict the formation of school districts and students in unfettered states is greater

Table 9.4 Regression results on the effect of limiting competition on student learning

	Simple interaction term		Spline specification	
	MSA pop	County pop	MSA pop	County pop
LimitedDistrictState	0.080 (0.125)	−0.0081 (0.106)	0.083 (0.125)	−0.020 (0.095)
MarketSize	2.93×10^{-9} (4.29×10^{-9})	-2.10×10^{-8}** (7.87×10^{-9})		
MarketSize ×*LimitedDistrictState*	-2.46×10^{-8}* (1.31×10^{-8})	-8.51×10^{-8} (9.22×10^{-9})		
*MarketSizeSpline*1			1.11×10^{-5}*** (2.14×10^{-6})	$-1.12 \times e^{8}$ $(2.20\text{x}10^{-7})$
*MarketSizeSpline*2			3.36×10^{-9} (4.33×10^{-9})	-2.11×10^{-8}** $(1.33 \times e^{-8})$ (8.63×10^{-9})
*MarketSizeSpline*2 ×*LimitedDistrictState*			$-2.51 \times e^{-8}$*	-9.77×10^{-8} (1.02×10^{-7})
Market size break point			91,000	245,254
Number of observations	4,149	5,297	4,149	5,297
R-squared	0.206	0.196	0.207	0.196

Note * indicates significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level respectively

Table 9.5 Sample of MSAs negatively impacted by mandates restricting competition among public schools

MSA name	Population	Negative impact on test scores
Washington DC/Baltimore, MD	7,608,070	0.186
Atlanta, GA	4,112,198	0.100
Miami/Fort Lauderdale, FL	3,876,380	0.094
Tampa/St. Petersburg/Clearwater, FL	2,395,977	0.058
Norfolk/Virginia Beach, VA	1,569,541	0.037
Salt Lake City, UT	1,333,914	0.31
West Palm Beach/Boca Raton, FL	1,131,184	0.026
Honolulu, HI	876,156	0.020

Notes Impact on test scores represents the estimated number of standard deviations lower the average test score is as a result of the restrictive mandate. The impact is calculated based on the first regression in Table 9.4: $2.46e^{-8}\times$ (MSA population minus 57813)

in larger markets is supported when the market is defined as the number of people living in the metro area. The coefficients on the interaction terms in the first and third regressions are negative and are significant at the 0.07 level with a two-tailed test and at the 0.04 level using a one-tailed test.

We present a list of the eight largest MSAs whose students are negatively impacted by restrictive mandates in Table 9.5. These markets serve between 60,000 and 400,000 students each. Using the coefficient for the interaction term in the first regression in Table 9.4, we estimate that these markets have test scores that are 0.02–0.19 standard deviations lower than they would be in the absence of restrictions mandating few school districts.

Table 9.6 Regression results for first regression in Table 9.4

Variable	Coefficient	Standard error
Student inputs		
Female	−0.0339	0.0227
Black	−0.5547	***0.0496
Hispanic	−0.1922	***0.0538
Age	−0.0611	***0.0201
Parent inputs		
Mother's years of education	0.0571	***0.0080
Father's years of education	0.0480	***0.0086
Live with biological father	0.1127	***0.0359
Volunteer at student's school	0.0766	***0.0247
Attend PTA meetings	−0.0172	0.0194
Number of children under 18 living in household	−0.0169	0.0142
School inputs		
Threatened	−0.0824	***0.0273
Student teacher ratio	−0.0045	0.0030
School size	$1.96 \times e^{-5}$	$5.54 \times e^{-5}$

Note Dependent variable is grade normalized PIAT math test score. The market size variable is MSA population. Standard errors are clustered at the state level. *** indicates significant at the 1 % level

On the other hand, the coefficient on the interaction variable is insignificant when the school market is measured by the number of people in the county. We argue that the county measure is inferior to the MSA measure because the county measure does not take into account whether the metro area has 2 counties, 3 counties, and so on. That is, the county measure does not take into account the full extent of the school market. Note that the R^2 is noticeably higher when the metro population is used than when county population is used.

All of the coefficients on the dummy variable *LimitedDistrictState* are not statistically significant. This is expected because there should be no difference in student test scores when a very small unfettered school market provides no more districts than would be allowed in a limited-districted state. This result also rules out the possibility that states with restrictive mandates are somehow intrinsically different and less productive than states without mandates. It is only when the mandates are binding and effectively limit the number of school districts that test scores are negatively impacted.

The results for the other independent variables, which describe the effects of student, parent, and school inputs, are reported in Table 9.6. We find that males and females perform equally well on the Math PIAT. However, Black and Hispanic students earn grades approximately 0.6 and 0.2 standard deviations lower than white students, respectively. Since the test scores are normalized by grade, it is not sur-

prising that the coefficient on Age is negative. This indicates that if grade is held constant, older students (perhaps students that have been held back a grade) do not perform as well on the exam. Both mother and fathers education positively impact test scores. Results indicate that students whose mother or father has a four-year college degree will likely score 0.2 standard deviations better on the exam than a student whose parents have not finished at least a year of college. Students who live with their biological father perform, on average, 0.1 standard deviations better than students who do not live with their biological father. Finally, we find that the student teacher ratio is inversely related to a students score on the exam. Specifically, an average student at a school with a teacher for every ten students will score 0.07 standard deviations better on the exam that an average student at a school with a teacher for every twenty-five students.

9.6 Discussion

There has been a sharp decline in the number of school districts in the U.S.: from 127,422 in 1931–32 to 83,642 in 1949–50 to 15,987 in 1980–81 and to 13,051 in 2007.[11] There were some exceptions to the pattern of decline; in half a dozen states the number of school districts increased between 1950 and 1980. Kenny and Schmidt (1994) found that a fall in the transportation cost of adding one more student to a school district led to larger and fewer school districts.[12] Over time the fall in the farm population and in the school age population and the increase in population density have been associated with the drop in the number of school districts. Since there is a fixed cost to organizing a school district, teacher unions have pressed for the consolidation of school districts. Perhaps in response to the growing strength of teacher unions, many states have offered financial incentives for school districts to consolidate. Kenny and Schmidt (1994) found that a rise in membership in the main teachers' union (National Education Association) led to a fall in the number of school districts. The growth in the share of revenue coming from state governments in recent years has been used to reduce the inequality of spending across school districts, which in turn has reduced the benefit from having more school districts. Kenny and Schmidt found that a rise in the state share in school revenue indeed resulted in fewer school districts. The fall in the number of school districts due to all these effects has resulted in less competition between school districts, which we have found makes school districts less efficient.

[11] Kenny and Schmidt (1994) and Bureau (2012).

[12] See Fischel (2009) for a fascinating account of the determination of the number of school districts in the early 1900s.

9.7 Conclusion

There is no consensus in the literature on whether competition between school districts makes schools more efficient. We identify states that have had a restriction that school districts be county-wide or state-wide, or have some other configuration that has been in place for more than six decades. These restrictions on district formation are treated as exogenous. Our empirical analysis also takes advantage of the fact that large educational markets tend to have many more school districts than small markets. The empirical analysis does not utilize any data on the number of school districts. Instead, the classification of states into limited-district states and unfettered states is used in conjunction with the size of the market to predict that restricted-district students fare worse than students in unfettered states. There should be more inter-district competition in large markets than in small markets in unfettered states, and there should be no more competition in large school markets county-wide restrictions than in small school markets with county-wide restrictions. Thus the disparity in math scores between limited-district states and unfettered states should be greater in larger markets than in smaller markets.

We find strong evidence that restricting competition among public school districts has an adverse impact on student learning. In smaller markets where having more than one school district is inefficient, restricting competition has no impact on student learning. In larger markets, math scores are lower in limited-district states than in other states. In the school districts with the most students, these restrictive mandates are especially harmful. We find that binding laws that mandate county-wide or state-wide school districts cause student test scores to drop by 0.02–0.19 standard deviations.

References

Borland MV, Howsen RM (1992) Student academic achievement and the degree of market concentration in education. Econ Educ Rev 11(1):31–39

Brunner EJ, Squires T (2013) The bargaining power of teachers unions and the allocation of school resources. J Urban Econ 76:15–27

Bureau UC (2012) Statistical abstract of the United States. US Census Bureau, Washington DC

(2002) The book of the states. Council of State Governments, Chicago

Fischel WA (2009) Making the grade: the economic evolution of American school districts. University of Chicago Press, Chicago

Fisher RC, Wassmer RW (1998) Economic influences on the structure of local government in US metropolitan areas. J Urban Econ 43(3):444–471

Hoxby CM (2000) Does competition among public schools benefit students and taxpayers? Am Econ Rev 90(5):1209–1238

Kenny LW, Schmidt AB (1994) The decline in the number of school districts in the US: 1950–1980. Public Choice 79(1–2):1–18

Lott J, Kenny LW (2013) State teacher union strength and student achievement. Econ Educ Rev 35:93–103

Markwardt FCJ (1998) Peabody individual achievement test-revised. American Guidance Service, Circle Pines
Rose H, Sonstelie J (2010) School board politics, school district size, and the bargaining power of teachers unions. J Urban Econ 67(3):438–450
Ross JM, Hall JC, Resh WG (2014) Frictions in polycentric administration with noncongruent borders: evidence from Ohio school district class sizes. J Public Admin Res Theory 24(3):623–649
Rothstein J (2007) Does competition among public schools benefit students and taxpayers? Comment. Am Econ Rev 97(5):2026–2037
Snyder TD (1998) Digest of Education Statistics, 1998. National Center for Education Statistics
Strang D (1987) The administrative transformation of american education: school district consolidation, 1938–1980. Admin Sci Quart 32(3):352–366
Tiebout CM (1956) A pure theory of local expenditures. J Polit Econ 64(5):416–424
Zanzig BR (1997) Measuring the impact of competition in local government education markets on the cognitive achievement of students. Econ Educ Rev 16(4):431–441

Chapter 10
The Endowment Effect in a Public Goods Experiment

Edward J. Lopez and William Robert Nelson Jr.

Abstract The endowment effect suggests that consumer preferences are reference-dependent; i.e., that the shapes of indifference curves depend on an agent's initial endowment and the direction of exchange offers. Hence, a person may value a good more highly once ownership is established, causing disparity between willingness to accept and willingness to pay value measures. In this paper we test for the endowment effect in a manner that does not rely on observing value disparities. We employ a one shot voluntary contribution mechanism (VCM) with treatments for account framing, duration effects, and the physical handling of the initial endowment. The treatments are designed to vary subjects' perceived ownership over their experiment endowments. Results generally fail to support reference-dependence in manners suggested by the endowment effect. Contribution rates are higher when initial endowments begin in subjects' private accounts compared to when originating in the shared public account. Contributions are no different when subjects hold their endowments for up to one week. And contributions are higher among subjects who physically handle cash compared to those indicating their decisions in writing.

10.1 Introduction

The endowment effect suggests that preference formation is reference-dependent; i.e., that loss aversion, status quo bias, or inertia can create some manipulation (shift, kink, rotation, etc.) of indifference curves about the point of initial endowment (Knetsch 1989; Tversky and Kahneman 1991; Kahneman et al. 1991; Morrison 1997; List 2004). If preferences depend on the reference state, one consequence may be that individuals exhibit disparities between willingness to accept (WTA) and willingness to pay (WTP) measures of value. Such value disparities have been the key observational medium for testing the endowment effect. Ongoing theoretical work on

E.J. Lopez (✉)
Western Carolina University, Cullowhee, NC 28723, USA
e-mail: edwardjlopez@gmail.com

W.R. Nelson Jr.
Eqis Capital Management, San Rafael, CA 94901, USA

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_10

reference-dependence continues to unfold in terms of value disparities (Kőszegi and Rabin 2006). In surveys, lab experiments, and field tests, value disparities have been so widely reported that it now seems naive to argue that such apparent anomalies do not exist under certain conditions.[1]

Rather, the developing literature has centered on whether value disparities are anomalous or instead are borne of substitution effects consistent with conventional preferences. Hanemann (1991, 2003) predicted that value disparities should be smaller for goods that are more substitutable. In experiments, competitive market forces caused observed value disparities to diminish (Brookshire and Coursey 1987; Shogren et al. 1994; List 2003). In addition, experience and information also tended to mitigate value disparities (Coursey et al. 1987; Knetsch and Sinden 1987; Kahneman et al. 1990).[2] Beyond the impact of market discipline and experience, the question remains fairly open whether value disparities are due to endowment effects, low substitutability, or other.

Furthermore, there is a tendency to conflate "endowment effect" with value disparities.[3] This unfortunately obscures the broader nature of the endowment effect; it is a statement that preferences change when the reference state changes, of which only one potential consequence would be WTA-WTP disparities. The endowment effect and WTA-WTP disparity should not be treated synonymously. These reasons and others suggest the usefulness of testing for the endowment effect in a manner that does not rely on observing WTA-WTP disparities.

This study presents a one-shot public good (voluntary contribution mechanism) design that allowed the experiment simultaneously to: (1) eliminate market discipline; (2) eliminate market experience; (3) hold substitution effects constant; and (4) observe treatments that one would expect to elicit the endowment effect.[4] If preferences depend on the initial endowment in manners suggested by the value disparity literature, then the endowment effect is likely to emerge under the conditions of the one-shot voluntary contribution mechanism (VCM), because it eliminates both market discipline and experience. To control for substitution effects, we used a public good that is perfectly substitutable: cash. Therefore, our experimental environment created favorable circumstances for observing an endowment effect, while holding constant the leading alternative explanation for value disparities.

---

[1] See Horowitz and McConnell (2002) for a survey of WTA-WTP studies, but note that results vary. Most of this type of evidence have come from laboratory experiments, but increasing amounts of field evidence continues to emerge. Macmillan et al. (1999), for example, compared donations to an actual charity under alternative contingent valuation procedures, and List (2003, 2004) observed bid and ask prices for sports memorabilia in actual markets.

[2] Plott and Zeiler (2005) found no evidence of a WTP-WTA gap after extensive subject education and practice with a modified Becker–DeGroot–Marschak mechanism.

[3] One anonymous reader characterized an early version of this paper as considering "whether the endowment effect–people seem to dislike giving something up more than they like getting it–exists in a VCM..."

[4] Some of the WTA-WTP auction experiments have controlled for substitution effects (magnitude of the *MRS*) as well as income effects (movement among alternative indifference curves). See, in particular, List (2004) and Shogren et al. (1994). The present design holds substitution constant and implicitly assumes negligible income effects.

Results from this experimental design do not support the thesis that preferences depend on initial endowments. In one set of treatments, the duration for which participants held a cash endowment before making their public good decisions failed to influence participants' allocations. In another set of treatments, participants contributed more to the public account when they were told their initial endowment originated in their private accounts. As we will argue, the direction of this difference was the opposite of what the endowment effect should impart. Furthermore, participants who physically held cash contributed more than those who did not. This effect is also in the opposite direction than the endowment effect would suggest, but the differences were not statistically significant.

In the next section, we discuss previous tests of the endowment effect and explain how the VCM can be applied. In Sect. 10.3 the experimental design and hypotheses are explained. Section 10.4 contains results and their discussion, and Sect. 10.5 concludes with extensions for future research.

10.2 The VCM and the Endowment Effect

The endowment effect has usually been studied in market auctions by comparing the extent to which agents' WTA exceeds their WTP. The early literature on this topic explained observed value disparities in terms of Thaler's endowment effect (Thaler 1980), which suggests that agents may value a good more highly when their property right is already established. Knetsch and Sinden (1984), for example, showed that WTA > WTP for lottery tickets and attributed the disparity to the endowment effect and loss aversion. Subsequent experiments-e.g., Coursey et al. (1987), Knetsch and Sinden (1987), Kahneman et al. (1990)-allowed for subject experience, yet also explained observed value disparities as endowment effects.[5] Knetsch (1989) presented similar experimental evidence and concluded that the endowment effect implies anomalous preference formation-the shapes of indifference curves depend on the agent's initial endowment and the direction of exchange offers.[6]

These researchers invoked the endowment effect explanation because received theory (Willig 1976; Randall and Stoll 1980) indicated that value disparities for private goods would depend on the magnitude of the income elasticity, which is negligible for magnitudes of typical experiment earnings. In contrast, Hanemann (1991) showed that the value disparity for quantity changes of public goods depends on both the income and the substitution effects. His solutions demonstrated that as the substitution effect becomes smaller (greater) the value disparity becomes greater (smaller), holding the income elasticity constant. With negligible income effects or

[5]The subsequent cited studies also achieved greater control by eliminating the need for subjects to calculate expected winnings, and differing attitudes toward risk, associated with lottery tickets.

[6]Going further, Kahneman et al. (1991) argued that the endowment effect can result in intersecting indifference curves.

perfect substitutability, there should be no value disparity (Hanemann's Proposition 3).[7] This implies that the substitution and endowment effects are alternative, though not necessarily mutually exclusive, explanations for value disparities (cf. Morrison (1997)).

Shogren et al. (1994) tested endowment versus substitution explanations for value disparities using multiple-trial, second-price, sealed-bid Vickrey (1961) auctions for two goods: one with close substitutes (candy bars) and one with few substitutes (sandwiches with decreased health risk). For the high-substitutable good they found that the value disparity diminished to negligible amounts, and converged to the approximate market price after approximately four trials. However, for the low-substitutable good, the value disparity persisted, even after many trials. Shogren et al. (1994) isolated the effects of different auction mechanisms (i.e., institutions) on measured value disparities by recreating the coffee mug experiments using a Vickrey auction instead of a random bid auction (Becker et al. 1964), which Kahneman et al. (1990) used. Shogren et al. (1994) found, contrary to Kahneman et al. (1990), that the value disparity diminished after the first of ten trials. The results were more consistent with the substitution effect than the endowment effect, which Shogren et al. explained by the Vickrey mechanism being more market-like than the Becker–DeGroot–Marschak.[8] List (2003) further demonstrated the market discipline and experience effects using an innovative field experiment.

Morrison (1997) suggested that the experimental design of Shogren et al. (1994) was insufficient for rejecting the endowment effect because the design required the endowment and substitution effects to work mutually exclusively. Following the logic of irreversible indifference curves (Knetsch 1989), Morrison graphically demonstrated how the value disparity can be larger for goods with fewer substitutes if the endowment effect is allowed to reinforce the substitution effect, such that the indifference curves pivot in a particular manner. In response, Shogren and Hayes (1997) noted that Morrison's pivots were seemingly arbitrary. By using different pivots, they showed that the value disparity can be of equal size for linear and convex indifference curves. Thus, *if observed value disparities change in magnitude while holding the*

[7] Hanemann (1991) reformulated the bounds on the neoclassical compensating and equivalent variations determined earlier by Willig (1976) and Randall and Stoll (1980). He reduced the difference between WTA and WTP to the ratio of the income elasticity of the public good to the elasticity of substitution between public and private goods. As we will argue, our experiment assumes negligible income effects and holds the substitution effect constant in treatments designed to elicit an endowment effect.

[8] See also Brookshire and Coursey (1987), Coursey et al. (1987), and List (2003, 2004). The effect of market discipline/experience appears to be sensitive to institutional design. There are many institution-specific explanations for observed value disparities. First the perceived illegitimacy of a transaction might cause the required (narrowly interpreted) surplus from the transaction to exceed epsilon, thereby driving a wedge between WTP and WTA (e.g., Rowe et al. (1980)). Second, buyers are often able to negotiate a lower price if they understate their WTA; if the associated rules of thumb are adopted, then equilibrium WTA exceeds WTP (Knez et al. 1985). In surveys and one-shot acutions, reported preferences might be misrepresentations/mistakes, but in repeated market interactions such mistakes tend to diminish in magnitude and frequency. Third, WTP and WTA might vary according to which elicitation mechanism is used (Shogren et al. 2001).

substitution effect constant, it would be due to the endowment effect even for goods that are perfect substitutes.

In short, the literature on value disparity currently offers the following stylized facts. (1) Value disparity is observed under certain elicitation conditions. (2) The endowment and substitution effects are alternative but not mutually exclusive explanations. (3) The disparity diminishes as agents gain market experience and as the experimental environment imposes more market discipline. (4) The endowment effect can be described as some manipulation (e.g., pivot, kink, rotation) of agents' indifference curves.

The VCM simultaneously addresses several aspects of testing for the endowment effect. First, a parameter in the VCM is the agent's marginal rate of substitution between proceeds from the private and public accounts. We introduce treatments that are designed to elicit an endowment effect-i.e., to change the *MRS* based on the subjects' perceived control over the initial endowment. With experimental control and a constant substitution effect, any difference in contribution levels between treatment groups would be due to the endowment effect. Thus, it can be inferred whether subjects' indifference curves pivot/kink/rotate sufficiently so as to alter their observed decisions. Second, proceeds from either the private or public account are cash-denominated. This feature allows substitution and endowment to work mutually inclusively, since the goods are perfect substitutes. Third, the design creates favorable conditions for the endowment effect to emerge because we eliminate both market discipline (by using a public good) and market experience (by allowing only one trial). Fourth, a public good experiment may afford a closer test of value disparity theory. Hanemann's advance was the result of considering the exchange offer as a change in the quantity of a public good. Yet, to our knowledge, no public good (VCM) experiments have been used to test for value disparities. One study (Brookshire and Coursey 1987) used a public good (trees in a neighborhood park, which have "a large degree of substitutability" [p. 555]), but its purpose was to compare the results of contingent valuation versus auction mechanisms. Furthermore, Cherry et al. (2005) provide recent evidence on VCM experiments that indicates contribution rates are sensitive to certain treatment effects on subjects' endowments. Thus, the VCM public good experiment presents the opportunity to test for endowment effect-style preference formation without the need to observe WTA-WTP disparities.

10.3 Experimental Design

10.3.1 The VCM

In the two-player VCM, each player $i = 1, 2, i \neq j$, is given an initial endowment of ω dollars to be invested in two accounts-one shared, one private. Define x^i and y^i as $i's$ dollar proceeds from the public and private accounts, respectively. Total dollar

payoffs to each player i equal the sum of x^i and y^i. The rational agent's objective in this environment is to maximize

$$u^i = u^i(\omega - c^i, g(\Sigma c^i)) \tag{10.1}$$

where $i's$ choice variable is c^i dollars contributed to the public account. In this experiment, private account payoffs are unweighted such that $y^i = \omega c^i$. To define payoffs from the public account and to characterize contribution incentives, differentiate equation 10.1 with respect to x^i and normalize by u^i_x to obtain

$$du^i = -1 + \frac{u^i_y}{u^i_x} g'. \tag{10.2}$$

Note that Eq. 10.1 contains the agent's marginal rate of substitution between the private and public goods. In a seminal study on VCM experiments, Isaac et al. (1984) defined the second term in Eq. 10.1 as the marginal per capita return (*MPCR*) from the public account. It is the product of the agent's $MRS_{x,y}$ (under a given payoff structure) and the marginal rate of transformation (as specified by experiment parameters). Proceeds from the public account depend on the technology of the experiment, g, which characterizes the *MRT*. The general form of the VCM public good production function is

$$g = \frac{a\Sigma c^i}{N}, \tag{10.3}$$

which, for this two-player experiment is

$$g = \frac{1.5(c^1 + c^2)}{2}, \tag{10.4}$$

such that the $MPCR = 0.75$. Proceeds from either account are denominated in dollars such that

$$\frac{u^i_y}{u^i_x} = MRS_{x,y} = 1. \tag{10.5}$$

The socially optimal contribution is $c^i = \omega$, but the Nash equilibrium is the strong free rider prediction $c^i = 0$. Previous experimental results under these types of conditions revealed contribution rates approximately 40 % of the optimal (Dawes and Thaler 1988; Ledyard 1995). Reference dependence in general, and the endowment effect in particular, suggests that $MRS_{x,y}$ will vary as subjects' perceived control over ω is varied under experimental control. If treatments successfully elicit the endowment effect, this will increase the disutility of c^i, the marginal dollar contributed to the public account. Hence, this will increase the $MRS_{x,y}$ such that the indifference curve

is rotated in the manner discussed above, which would decrease average contribution levels.

10.3.2 Treatment Designs

Within the above VCM environment, this study features two primary treatment effects-account framing and duration framing-which, according to the surveyed literature, are expected to elicit an endowment effect. A third effect is also presented, by which the endowment effect may arise if participants' physical handling of cash imparts an endowment effect. We discuss these three effects in turn.

First, in the account framing (AF) treatments, the originating account is varied. In one treatment, participants were told ω (the initial endowment) began in the shared public account; in the other treatment, participants were told ω began in their private account. The AF treatments may elicit an endowment effect if subjects perceive the originating account as an initial property right. When ω originates in the private account, participants may initially feel a greater sense of ownership than when the endowment begins in the public account. Accordingly, if the AF elicits an endowment effect, there is reason to expect average contributions to be lower in treatments where ω starts in the private account.[9]

Second, in the duration framing (DF) treatments, the length of time that participants held the endowment prior to making their allocation decision is varied (by up to one week). Research on the endowment effect provides several reasons to expect a duration effect. First, some scholars have speculated that the endowment effect may have a temporal component that it may take time to bind in some sense (e.g., Knetsch and Sinden (1984); Kahneman et al. (1991); Strahilevitz and Loewenstein (1998).[10] Second, individuals may be more readily willing to part with windfall gains than earned wealth (Thaler and Johnson 1990; Cherry et al. 2005). Third, current spending may increase by less following a temporary increase in income compared to a longer duration increase (Friedman 1957). Participants who make their experiment decision immediately after receiving the endowment may perceive the endowment as a windfall gain and play as if they are using the "house's money." Subjects who are

[9]In value disparity experiments, WTA assigns the agent rights to the good while WTP offers the opportunity to acquire the good. By varying the originating account, the VCM experiment may mimic this difference regarding the direction of the exchange offer.

[10]To our knowledge, Strahilevitz and Loewenstein (1998) is the closest to this study in testing for duration effects. The derive duration-effect hypotheses by combining a prospect theory value function with *adaptation*, a concept in psychology, which "in the context of object ownership, is the tendency for people to become psychologically accustomed to changes in their material situation" (Strahilevitz and Loewenstein 1998, p. 277). Duration treatments up to one hour have been introduced in a variety of experiments employing WTP and WTA elicitation questionnaires. Results indicated that subjects generally express greater WTP and WTA as the endowment is held for a longer duration. We are unaware of other experimental results on duration effects.

able to savor the increase in wealth for enough time may play as though the money is their own.

Third, in all the DF treatments, participants physically held cash as their endowment. In all AF treatments, participants submitted their decisions in writing without handling cash. Cash in hand may induce feelings of ownership more so than money in an account. Accordingly, if cash-in-hand elicits an endowment effect, there is reason to expect average contributions to be lower in treatments where participants handled cash.

10.3.3 Treatment Groups and Hypotheses

We conducted six treatment groups, two for AF and four for DF. Each treatment group consisted of two sessions, meeting simultaneously in separate rooms, for a total of 12 sessions. All AF sessions were run in a laboratory setting. For reasons that will become apparent, the DF sessions were run in both laboratory and classroom settings. Table 10.1 summarizes the six treatment groups. Details regarding logistics and the protocol are available from the authors.

As is apparent from Table 10.1, the six groups were organized as three pairs of treatments. Comparing average contribution levels between the treatment pairs provides the basis for the hypothesis tests. The null hypothesis is no identifiable treatment effect. The alternative hypothesis is provided by the direction of the anticipated endowment effect.

All AF sessions were conducted in a laboratory setting. No AF participants handled cash until after all decisions were made. Rather, in one AF treatment subjects were told the initial endowment ω originated in their own private account. In the other AF treatment, subjects were told ω began in the public account. As Table 10.1 shows, these groups are named *ALR* and *ALU*, respectively. Subjects then wrote

Table 10.1 Summary of treatment groups and hypotheses

	Account framing (Did not handle cash)			Duration framing (Handled cash)			
	Laboratory			Laboratory		Classroom	
	ω begins in			ω held for		ω held for	
	Private account	Public account		25 min	1 min	1 week	1 min
Treatment group name	*ALR*	*ALU*		*DLL*	*DLS*	*DCL*	*DCS*
Endowment effect hypotheses							
Account and duration framing	$\bar{c}_{ALR} < \bar{c}_{ALU}$			$\bar{c}_{DLL} < \bar{c}_{DLS}$		$\bar{c}_{DCL} < \bar{c}_{DCS}$	
Handling versus not handling cash			$\bar{c}_{AL} > \bar{c}_{DLS}$ or $\bar{c}_{AL} > \bar{c}_{DCS}$				

Notes The initial endowment is ω, and $\bar{c}$ is the mean or median contribution within a treatment group

down their allocation decisions and were paid accordingly at the end of the session. This enables the test of the following, where is the mean (or median) public account contribution within a group.

H1: Account framing imparts an endowment effect; *ALR* participants will contribute less to the public account than *ALU* participants. That is $\bar{c}_{ALR} < \bar{c}_{ALU}$.

In the DF treatments, this experiment had "Short" and "Long" groups, which were defined by the length of time that subjects held the endowment prior to making the allocation decision. Duration framing is easy to accomplish in the laboratory, but the length of the duration treatment is limited by how long participants can be asked to stay. We were cautious not to make our sessions too long so as to hinder subject recruitment. More importantly, because Short and Long participants were recruited simultaneously, having one session last longer could introduce a loss of experimental control. Therefore, in the laboratory treatments we varied the duration by only 25 min, the length of time required to complete the instructions. As shown in Table 10.1, we assigned the Short and Long laboratory groups the treatment names *DLS* and *DLL*, respectively.

The design problem was more challenging for observing the longer duration treatments. We considered scheduling participants for two laboratory sessions. In Session 1 we would explain the potential winnings and take care of paperwork, such as consent forms. We would give the cash to the Long group in Session 1 but not to the Short group. In Session 2, perhaps a week later, all participants would reconvene and play the VCM. According to the endowment effect, we would expect the Long participants to contribute less on average to the public account than the Short group. The obvious difficulty with this approach is that participants who attended Session 1 might fail to show up for Session 2. If those who do show up for Session 2 are more trustworthy than those who do not, this design would likely select cooperators. To minimize this risk and obtain results as free from selection bias as practicable, we decided to run the longer duration treatments under the structure of regularly meeting university classes. With instructor permission, we were allowed to visit four different classes during two consecutive weeks. In Week 1, we explained to students that they would have the opportunity to participate in an experiment that would take place in the same class one week later, and we took care of paperwork. For the Long treatment, we also distributed cash in Week 1 and asked students to bring an equal amount of cash with them to Week 2. For the Short treatment, we simply told students they could participate in an experiment, for monetary earnings, during the following week's class. Using this approach, students had the added incentive to show up for Week 2, reducing the likely extent of selection effects. As shown in Table 10.1, the endowment effect suggests the following.

H2: The length of time one possesses an item increases the strength of the endowment effect; Long group participants will contribute less than Short group participants. That is $\bar{c}_{DLL} < \bar{c}_{DLS}$ or $\bar{c}_{DCL} < \bar{c}_{DCS}$.

Finally, all DF participants handled cash when making their experiment decisions, while all AF participants wrote their decisions without handling cash. Comparing

only Short group DF participants with AF participants, it is possible to test the following.

H3: Cash-in-hand increases the strength of the endowment effect; Short group participants in the laboratory or classroom setting, who handled cash, will contribute less than AF participants, who did not handle cash. That is, $\bar{c}_{DLS} < \bar{c}_{AL}$ or $\bar{c}_{DCS} < \bar{c}_{AL}$.

10.4 Results

10.4.1 Overview

We present all results in Fig. 10.1, Tables 10.2, and 10.3. Through 12 experiment sessions a total of 284 undergraduate participants, from a wide range of majors at two large public universities, one in New York (NY) and the other in Texas (TX), each made one allocation decision. In the first set of sessions, 75 students from NY were divided among four laboratory treatments. The second set consisted of 80 students from TX, divided among the four classroom sessions. In the third set were 129 students, also from TX, divided among the remaining four laboratory sessions. All duration framing (DF) treatments took place in TX. In the combined account framing (AF) treatments, NY students contributed less than TX students (mean 4.12 versus 5.17, p = 0.07 according to a two tailed t test assuming unequal variances). In each of the pairs of AF treatments at each school, the directions were the same and the sizes

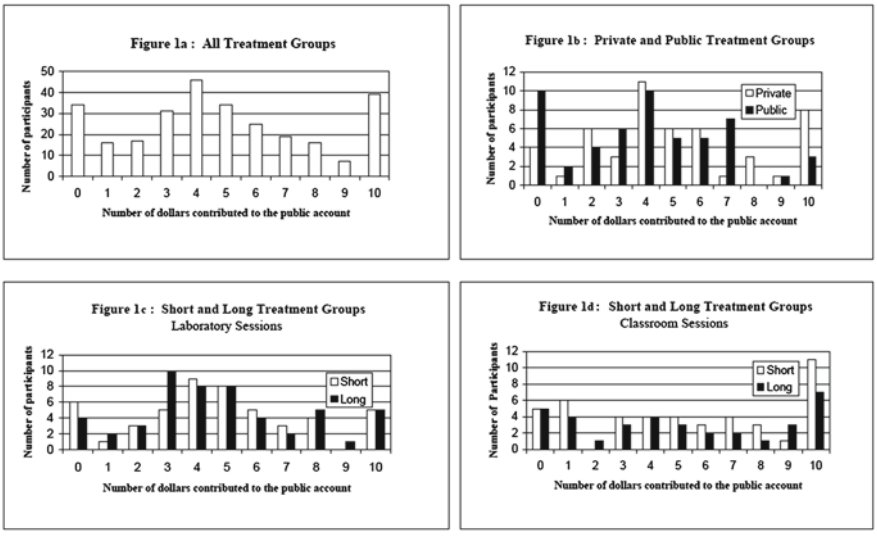

Fig. 10.1 Contribution frequencies by treatment groups

Table 10.2 Contribution descriptive statistics by treatment group

	Treatment groups divided by location								Six treatment groups						Combined groups		
	A_{LRY}	A_{LUY}	D_{LSY}	D_{LLY}	A_{LRT}	A_{LUT}	D_{LST}	D_{LLT}	A_{LR}	A_{LU}	D_{LS}	D_{LL}	D_{CS}	D_{CL}	AL	D_{short}	D_{long}
Mean	5.4	3.94	4.1	4.14	4.89	4	5.14	5.2	5.04	3.98	4.71	4.75	5.4	5.06	4.5	5.04	4.87
Standard error	0.7	0.69	0.68	0.59	0.54	0.48	0.5	0.51	0.43	0.39	0.41	0.39	0.54	0.62	0.29	0.33	0.34
Median	5	4	4	3.5	4	4	5	5	4.5	4	5	4	5	5	4	5	4
Mode	4	5	4	3	4	4	5	4	4	0	4	3	10	10	4	10	3
Standard deviation	2.69	2.94	3.06	2.78	3.17	2.84	2.71	2.78	3.02	2.85	2.87	2.81	3.6	3.66	2.96	3.24	3.16
Sample variance	7.26	8.64	9.36	7.74	10.05	8.06	7.34	7.75	9.1	8.1	8.25	7.88	12.97	13.41	8.78	10.51	9.99
Range	8	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
Minimum	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maximum	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
No. observations	15	18	20	22	35	35	29	30	50	53	49	52	45	35	103	94	87

Total number of subjects is 284. Group name key: First letter denotes kind of treatment: A = account framing and D = duration framing. Second letter denotes settings: C = classroom and L = lab. Third letter denotes specific treatment: L = long, S = short, R = private account and U = public account. Subscript denotes location: Y = New York and T = Texas

Table 10.3 t-Tests for account framing effect

	ALR		ALU	DLS		DLL	DCS		DCL	D		AL
Mean amount deposited into public account	5.04		3.98	4.71		4.75	5.4		5.06	4.96		4.5
Variance	9.1		8.1	8.25		7.88	12.97		13.41	10.22		8.78
Observations	50		53	49		52	45		35	181		103
Hypothesized mean difference		0			0			0			0	
df		100			98			73			226	
t Stat		1.83			−0.06			0.42			1.24	
P(T <= t) one-tail		0.04			0.47			0.34			0.11	
t-critical one-tail		1.66			1.66			1.67			1.65	
P(T <= t) two-tail		0.07			0.95			0.68			0.22	
t-critical two-tail		1.98			1.98			1.99			1.97	
Mann–Whitney two-tail P,(test of medians)		0.10			0.90			0.64			0.26	
Number of observations required for one-tailed t-test (p = 0.1 and a power of 0.80)	N-1: 71 N-2: 67			N-1: 46,040 N-2: 45,077			N-1: 1,022 N-2: 1,039			N-1: 419 N-2: 388		

Notes Two-Sample assuming unequal variances. Group Name Key. First letter denotes kind of treatment: *A* = account framing and *D* = duration framing. Second letter denotes setting: *C* = classroom and *L* = lab. Third letter denotes specific treatment: *L* = long, *S* = short, *R* = private account and *U* = public account

of the differences in means were similar. The mean contribution within treatment groups ranged from a low of 3.94 in the *ALU* treatment at NY to a high of 5.40 in the *DCS* treatment. The contributions overall appear normally distributed around four, but with spikes at 0 and 10 (Fig. 10.1a). All the laboratory sessions had similarly shaped trimodal distributions, but the distribution of the classroom appeared more uniform with a single mode at 10 (Fig. 10.1d) while not significantly increasing the mean in these groups. The mean contribution in the classroom treatments equaled 5.25, and the mean in the laboratory treatments equaled 4.61.

10.4.2 Tests of Hypotheses

H1 is weakly rejected. Participants contributed more when the money began in their private account than when the money began in their public account. This variation is the opposite of what the endowment effect would impart. As evident from Table 10.2 and Fig. 10.1d, over $1 more was contributed when ω began in the private account ($ALR = 5.04$) than when the money began in the public account ($ALU = 3.98$). However, the difference is significant at slightly less than conventional levels (two-tailed $\rho = 0.07$) according to a two-sample t-test assuming unequal variances.

H2 receives no support. The results give no indication that the duration for which one holds cash has an impact on an individual's contributions to the public good. In the laboratory, the long group contributed more, although the difference was negligible ($DLL = 4.75$ and $DLS = 4.71$) and not close to significant (two-tailed $\rho = 0.95$). In the classroom the difference was larger in magnitude, but also in the opposite direction than expected ($DCL = 5.06$ and $DCS = 5.40$, with two-tailed $\rho = 0.68$). According to a power test, over 45,000 laboratory observations would be required for the t-test to identify 0.10 significance with 0.80 probability. Similar power would be accomplished with just over 1,000 classroom observations.

H3 is rejected. Cash-in-hand participants contributed more ($DCS = 5.4$ and $DLS = 4.71$) than participants who indicated their contributions in writing ($AL = 4.5$).[11] This is the opposite of what the endowment effect would impart, though not statistically significant. The DLS mean was 0.21 higher (two-tailed $\rho = 0.38$), and DCS was 0.90 higher (two-tailed $\rho = 0.14$). Combining the *DLS* and *DCS* data, mean contributions were 0.46 higher (4.96, two-tailed $\rho = 0.22$).[12]

[11] Note that we compare only the Short groups from the duration treatments.

[12] In the AF treatments, the participants were told that the initial endowment originated in a particular account, but the DF instructions included no reference to the originating account.

10.4.3 Discussion

This experimental design created circumstances that value disparity experiments have shown to be favorable for eliciting an endowment effect. In particular, as discussed in Sect. 10.2 above, recent work has shown that value disparities diminish with market experience and discipline. By using a one-shot public good game, one would expect evidence of endowment effects to emerge in participants' contributions to the public account. On the contrary, the data from this experiment generally point to a negative result. The endowment effect is elusive in a cash-based, one-shot VCM, which fails to support the thesis that preferences are sufficiently reference-dependent so as to alter observed public good contributions. More specifically, the *DF* and cash-in-hand treatments fail to support earlier speculations regarding the "immediacy of the transaction" (Knetsch and Sinden 1984) or "gambling with the house's money" (Thaler and Johnson 1990).

Inherent features of the VCM or the use of cash in this experiment may be confounding matters. For example, the results from AF treatments-which were designed to frame the contribution decision as giving versus taking are consistent with the result from Andreoni (1995) that the warm glow effect in a positive frame is stronger than the cold prickle in a negative frame. Andreoni's experiment, and several that have followed, were linear public good games like the experiment in this study.[13] Thus, the AF results are consistent with the results obtained in the broader literature, regardless of the attempts to elicit reference-dependence through framing. Second, while the DF and cash-in-hand treatments do not support the "house money" effect, this may derive from there being little expectation of an endowment effect with money. Cash is more divisible than the goods used in earlier value disparity tests (coffee mugs, candy bars, sports cards, etc.). Moreover, as the numeraire good cash also differs in that subjects hold cash for future purchases, not for consumption per se (Kahneman et al. 1990, p. 1328). However, participants in a linear public good experiment are not sensitive to earned income either (Cherry et al. 2005). The VCM has well documented deviations from the Nash prediction, which can create control problems when attempting to elicit a deviation from consumer theory such as reference-dependence. We cannot rule out, for example, that the endowment effect in a VCM is of some statistical magnitude regardless of account/duration framing or handling cash.

Thus, while these results cannot point to the existence (or non-existence) of the endowment effect, the design of this experiment offers potentially fruitful new directions for testing reference-dependence. Traditional value disparity approaches have grappled recently with the relative merits of alternative elicitation procedures and their institutional attributes (e.g. Shogren and Hayes (1997); List (2004)), rather than

[13] Similarly, Andreoni (1995) also found that the warm glow is stronger than the cold prickle. In his experiment, contributions to the public good were greater when the game was explained in terms of a positive rather than a negative externality. In both his treatments, all money began in each individual's "Investment Account," and participants chose between depositing tokens in a "Private Exchange" and a "Public Exchange."

whether preferences are reference-dependent the central question of the endowment effect. Furthermore, too few value disparity experiments have investigated substitution between private and public goods, which would be a closer test of received consumer theory on WTA-WTP (Hanemann 1991, 2003). The VCM approach disentangles the endowment effect from value disparities and institutional differences within alternative auction mechanisms. In principle, reference-dependence generally, or the endowment effect more specifically, can be tested in a variety of experimental environments that offer subjects an initial endowment with which to play. Our results invite even stronger account, duration, and endowment-in-hand treatments with games other than the VCM and using a less divisible, more consumable initial endowment than cash.

10.5 Conclusion

According to neoclassical theory, when the public good is perfectly substitutable with at least one private good and income effects are negligible, there will be no disparity between willingness to accept (WTA) and willingness to pay (WTP) measures of value (Hanemann (1991), Proposition 3). In the presence of an endowment effect, however, individuals may consider a good in possession as less substitutable due to loss aversion or status quo bias (Kahneman et al. 1991). The voluntary contribution mechanism (VCM) provides a tool for inferring whether the (unobserved) marginal rate of substitution between a public and private good is sufficiently sensitive to subjects' perceived control over the initial endowment so as to alter their (observed) contributions to the public account. The one-shot, cash-denominated VCM creates favorable circumstances for the endowment effect to emerge because it eliminates market discipline and experience while holding the substitution effect constant. Within this environment, we designed treatments that framed the initial endowment in several different ways. According to the endowment effect, treatments in which subjects had greater perceived control over the initial endowment should have contributed less to the public account. The results of 284 subjects in 12 different treatment sessions are not consistent with this expected effect.

Finally, our approach touches on critiques of standard experimental methods. Suppose there is some temporal component to how subjects respond when given an initial endowment with which to participate in an experiment. Skeptics could argue, as we hinted earlier in the paper, that subjects' decisions are unreliable if the experiment decisions are made immediately or soon after receiving the endowment. We liken this to the criticism of using student subject pools to represent the behaviors of actual economic agents in relevant markets (Davis and Holt 1993, p. 17). A preponderance of experimental evidence comparing students with professionals indicates that this "subject surrogacy" critique does not seem to detract from standard methodology. Similarly, our results on duration framing do not suggest evidence of problems

associated with allowing subjects to make their experiment decisions soon after receiving the initial endowment. Alternative explanations for these results are possible, thus calling for further investigations.

Acknowledgements We thank the Russell Sage Foundation Behavioral Roundtable for financial support. We thank Daniel Houser, John List, Vernon Smith, and Bart Wilson for helpful comments. Kari Battaglia, Mark Strazicich, and David Molina graciously donated class time for generating some of the data. We also thank Todd Jewell, Ife Isiekwe, Sangkyoo Kang, Jae Hoon Kim, Kaunyoung Lee, Nathan Roseberry, Diego Segatore, Shanhong Wu, and Oksana Zhuk for assistance monitoring the experiments.

References

Andreoni J (1995) Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. Quart J Econ 110(1):1–21

Becker G, DeGroot M, Marschak J (1964) Measuring utility by a single-response sequential method. Behav Sci 9(3):226–232

Brookshire D, Coursey DL (1987) Measuring the value of a public good: an empirical comparison of elicitation procedures. Am Econ Rev 77(4):554–566

Cherry TL, Kroll S, Shogren JF (2005) The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab. J Econ Behav Org 57(3):357–365

Coursey DL, Hovis JL, Schulze WD (1987) The disparity between willingness to accept and willingness to pay measures of value. Quart J Econ 102(3):679–690

Davis DD, Holt CA (1993) Experimental economics. Princeton University Press, Princeton

Dawes RM, Thaler RH (1988) Anomalies: cooperation. J Econ Perspect 2(3):187–197

Friedman M (1957) A theory of the consumption function. Princeton University Press, Princeton

Hanemann WM (1991) Willingness to pay and willingness to accept: How much can they differ? Am Econ Rev 81(3):635–647

Hanemann WM (2003) Willingness to pay and willingness to accept: How much can they differ? Reply. Am Econ Rev 93(1):464–464

Horowitz JK, McConnell KE (2002) A review of wta/wtp studies. J Env Econ Manage 44(3):426–447

Isaac RM, Walker JM, Thomas SH (1984) Divergent evidence on free riding: an experimental examination of possible explanations. Public Choice 43(2):113–149

Kahneman D, Knetsch JL, Thaler RH (1990) Experimental tests of the endowment effect and the Coase theorem. J Polit Econ 98(6):1325–1348

Kahneman D, Knetsch JL, Thaler RH (1991) Anomalies: the endowment effect, loss aversion, and status quo bias. J Econ Perspect 5(1):193–206

Knetsch JL (1989) The endowment effect and evidence of nonreversible indifference curves. Am Econ Rev 79(5):1277–1284

Knetsch JL, Sinden J (1984) Willingness to pay and compensation demanded: experimental evidence of an unexpected disparity in measures of value. Quart J Econ 99(3):507–521

Knetsch JL, Sinden JA (1987) The persistence of evaluation disparities. Quart J Econ 102(3):691–696

Knez P, Smith VL, Williams AW (1985) Individual rationality, market rationality, and value estimation. Am Econ Rev 75(2):397–402

Kőszegi B, Rabin M (2006) A model of reference-dependent preferences. Quart J Econ 121(4):1133–1165

Ledyard JO (1995) Public goods: a survey of experimental research. In: Kagel JH, Roth AE (eds) The handbook of experimental economics. Princeton University Press, Princeton, pp 111–194

List J (2003) Does market experience eliminate market anomalies? Quart J Econ 118(1):41–71
List JA (2004) Substitutability, experience, and the value disparity: evidence from the marketplace. J Env Econ Manage 47(3):486–509
Macmillan DC, Smart TS, Thorburn AP (1999) A field experiment involving cash and hypothetical charitable donations. Env Resour Econ 14(3):399–412
Morrison GC (1997) Resolving differences in willingness to pay and willingness to accept: comment. Am Econ Rev 87(1):236–240
Plott CR, Zeiler K (2005) The willingness to pay-willingness to accept gap, the. Am Econ Rev 95(3):530–545
Randall A, Stoll JR (1980) Consumer's surplus in commodity space. Am Econ Rev 70(3):449–455
Rowe RD, d'Arge RC, Brookshire DS (1980) An experiment on the economic value of visibility. J Env Econ Manage 7(1):1–19
Shogren JF, Hayes DJ (1997) Resolving differences in willingness to pay and willingness to accept: reply. Am Econ Rev 87(1):241–244
Shogren JF, Shin SY, Hayes DJ, Kliebenstein JB (1994) Resolving differences in willingness to pay and willingness to accept. Am Econ Rev 84(1):255–270
Shogren JF, Cho S, Koo C, List J, Park C, Polo P, Wilhelmi R (2001) Auction mechanisms and the measurement of wtp and wta. Resour Energy Econ 23(2):97–109
Strahilevitz MA, Loewenstein G (1998) The effect of ownership history on the valuation of objects. J Consum Res 25(3):276–289
Thaler R (1980) Toward a positive theory of consumer choice. J Econ Behav Org 1(1):39–60
Thaler RH, Johnson EJ (1990) Gambling with the house money and trying to break even: the effects of prior outcomes on risky choice. Manage Sci 36(6):643–660
Tversky A, Kahneman D (1991) Loss aversion in riskless choice: a reference-dependent model. Quart J Econ 106(4):1039–1061
Vickrey W (1961) Counterspeculation, auctions, and competitive sealed tenders. J Fin 16(1):8–37
Willig RD (1976) Consumer's surplus without apology. Am Econ Rev 66(4):589–597

Chapter 11
Are Roads Public Goods, Club Goods, Private Goods, or Common Pools?

Bruce L. Benson

Abstract An examination of the history of roads in England demonstrates that roads are never Samuelsonian public goods, and that free access roads are really common pools. In some institutional environments, however, many roads were club goods maintained through reciprocal arrangements. Private toll roads arose where possible but collecting tolls on "public" roads was a government prerogative. Nonetheless, as government action undermined club-good arrangements, local groups petitioned for and received permission to finance maintenance with tolls. Turnpike trusts managed these toll roads, but they were not "private" roads because significant regulations including government-mandated tolls and exemptions were imposed, based on political rather than economic considerations. Profit taking also was not allowed so incentives for trustees to monitor workers were weak, corruption was rampant, and many trusts ultimately failed. In the absence of regulatory constraints there is little doubt that private roads would have been widespread.

11.1 Introduction

An answer to the question posed in the title is offered in the following presentation by differentiating between the four concepts in the question and the institutional conditions that might create them, and then by examining the historical evolution of road provision systems in the United Kingdom. One conclusion is that roads are never public goods in a Samuelson (1954, 1955) sense. This answer may be surprising, since highways and roads are frequently cited as "important examples of production of public goods," (Samuelson and Nordhaus 1985, pp. 48–49). The second conclusion is that specific roads and road systems can be, have been, and are club goods, private goods, or common pools, depending upon the institutional

B.L. Benson (✉)
Florida State University, Tallahassee, FL 32306, USA
e-mail: bbenson@coss.fsu.edu

J. Hall (ed.), *Explorations in Public Sector Economics*,
DOI 10.1007/978-3-319-47828-9_11

environment within which the roads are provided.[1] To support these conclusions, this presentation is divided into six sections including this introduction, beginning in Sect. 11.2 where the concepts of public goods, club goods, private goods, and common pools are described and compared.

An extensive system of voluntarily created and maintained roads existed in medieval Great Britain, but actions taken by various kings undermined the incentives to maintain the system. In order to understand both the voluntary arrangements and their breakdown, Sect. 11.3 begins with a theoretically-based discussion of institutional characteristics that apply for successful voluntary provision of a club good. The analysis is subsequently employed in Sect. 11.4 to describe the early history of voluntary community-level road provision in Great Britain, as well as the coercive institution-changing actions by kings that broke down the incentives for members of some communities to cooperate in road provision. In order to make up for the reduction in voluntary road provision, the state was forced to create new institutions. The first was a mandated-contribution system which attempted to force local parishes to maintain roads. When this failed it was followed by decisions to allow regulated private entities to produce and maintain roads and control access so politically determined tolls could be charged.[2] These institutional arrangements are examined in Sect. 11.5, but in order to better appreciate their weaknesses and ultimate failure, Sect. 11.3 also offers a theory-based analysis of such coercive institutions. Section 11.5 applies this analysis to consider the actual rise and fall of Great Britain's mandated parish system and then its toll road arrangement. Despite initial success and widespread use of toll roads, the political manipulation of institutionalized incentives and tolls led to significant inefficiencies within this system and its ultimate demise. As a consequence, public financing of free access roads evolved, but the result is a common pool, not a Samuelsonian public good. Conclusions in Sect. 11.6 briefly note the substantial mix of club and private roads that still exist around the world alongside publicly provided common pool road systems. This section also contends that similar analysis applies to many other so-called "public goods".

11.2 Public Goods Versus Private Goods, Club Goods, and Common Pools

The seminal analysis of Samuelson (1954, 1955) indicates that the key **characteristics of public goods are: (1) non-excludability, and (2) non-rivalrous consumption, which combine to produce (3) free riding, and therefore, (4) "private**

[1] Minasian (1964) explains, in his criticism of public goods theory, that different institutional arrangements create different incentives for the allocation of resources. He discusses television signals, but consideration of other allegedly public goods has lead to similar criticism (e.g., Coase 1974; Benson 1994, 1998, 2010).

[2] Alternatives may exist (e.g., public roads fully financed by tolls, toll roads provided by private entities that are allowed to set tolls and retain profits), but these two dominated in the United Kingdom.

provision of these public goods will not occur" (Samuelson and Nordhaus 1985, p. 713) because coercive power is required to collect from non-paying free riders (McNutt 2000, pp. 927–928).[3] As a contrast, private goods are often characterized as being completely rivalrous in consumption in that one individual's use of the good means that it is completely gone so no other individual can use it. However, excludability (through force or bargaining) produces the same consequence, as the owner can dictate use and prevent others from using the good. Indeed, "private" generally refers to sole ownership and therefore control of access, so the key characteristic of a private good as the term is used here, is that it is owned by a single economic entity (e.g., an individual, a firm, an organization) with a right to exclude all other potential users. Thus, a private good need not be entirely consumed as the result of a single user's consumption (e.g., a road on a privately owned farm with a locked gate which can handle more traffic than it does; the viewing of a movie in a theater with several seats). In such circumstances, the owner can either use the good repeatedly, perhaps but necessarily ultimately depleting it (e.g., the farm road), or allow access by others if the fully internalized benefits of doing so exceed the fully internalized costs (e.g., the movie), but the good still can only be non-rivalrous to those individuals who are given access permission by the owner (e.g., those who pay to see a movie). Therefore, in comparison to a public good, a **private good's characteristics are: (1) excludability, (2) possibly, but not necessarily, rivalrous consumption (possibly non-rivalrous consumption for those who obtain permission to access), (3) non-owners must get permission (e.g., pay) for use, and (4) private provision occurs if it is allowed and profitable.**

Define a club to be a voluntarily-formed close-knit group of individuals who have a multi-dimensional web of mutually beneficial interactions. Since the club is voluntary, individuals who do not cooperate with others in the club are not likely to be accepted as members (i.e., individuals voluntarily accept membership and are voluntarily accepted for membership).[4] A club good is one that is produced within (or purchased by) a club and then consumed by all the members of the club. That is, access is free to members of a club (e.g., residents of a gated development whose homeowners' association owns the roads in the community), but not for non-

[3] Non-rivalrous consumption means that even though one person consumes the benefits of the good, everyone else can consume the same undiminished benefits. Non-excludability means that, not only unlimited numbers of people consume the benefits, but no one can be prevented from consuming them even if they do not pay their share of the costs. Free access to a non-rivalrous good creates "free-rider" incentives: individuals recognize that they can consume the benefits without paying, so they will not voluntarily pay for the good, and this means that private producers will not produce the good because they cannot collect revenues to cover costs (or at least, that they will not produce enough of the good because, while everyone can free ride, some may not).

[4] Buchanan (1965) contends that the dichotomy between pure private goods and pure public goods suggested by Samuelson and others is inappropriate. Instead, he suggests that all goods should be considered as club goods where the size of the club depends on the good. In theory, clubs can, in this sense, be as small as one person, or as large as infinity. A large theoretical literature on clubs has developed since the seminal work of Buchanan (1965). See McNutt (2000) for a review. Here, the issue is not how many people can consume a good, but rather, it is how many share ownership and the right to exclude others. This is more in line with the analysis of Ostrom (1990, 2005).

members (McNutt 2000, p. 928). Thus, a club good differs from a private good in that it "belongs" to and is used by a limited voluntary association of decision-makers who face collective decision-making costs. If high decision-making costs prevent an agreement, the good may not be produced, since no authority in the club has coercive power to mandate that individuals contribute. Therefore, if the good is produced (or purchased), it is done so voluntarily by the cooperating club members. Furthermore, for individuals outside the club, access requires obtaining permission from the club (members may agree to allow access by at least some non-members, depending on the costs and benefits of doing so in the context of their collective decision-making process). Such a good can be rivalrous or non-rivalrous in consumption for those with access (McNutt 2000, p. 928), **given** the size of and restrictions on access created by the club, so in this sense it can look like a public good.[5] If non-rivalry applies it is **because** access is limited, however. If congestion and overuse arises for club members, they can establish rules that limit use by members too (e.g., quotas). For comparison then, **a club good is: (1) non-excludable for club members but excludable for outsiders, (2) possibly but not necessarily** (McNutt 2000, p. 928) **non-rivalrous for those with access, (3) subject to collective decision-making costs, but within a voluntary close-knit group that can exclude free riders, and (4) voluntarily produced if high decision making costs do not prevent it.**

The common pool terminology usually is applied to a natural resource such as a fishery, but it also can describe many goods and services that are freely provided for some reason, often by the state (see Shoup 1964; Neely 1982; Benson 2010; Benson 2011, pp. 97–101; Rasmussen and Benson 1994, pp. 17–37), but also perhaps by a private entity – e.g., consider a shopping mall parking lot before Christmas or a private charity. A common pool exists when all people (or simply a substantial number of people who face high transactions costs in cooperating either as a club or in recognizing and bargaining over private property rights) have free or "common" access to a scarce good or resource that is subject to rivalry in consumption because one individual's use diminishes the benefits that another user gains. This diminution often involves crowding (congestion/overuse) and a deterioration in quality for all users (e.g., for roads, highway travel time rises, surface damage increases) as the result of over use. In fact, the incentives for individuals with access to the common pool are to consume as many benefits as possible before they are consumed by others. Therefore, an individual has incentives to rush into the commons in order to capture benefits that will dissipate quickly as others with access have the same incentives. This has been called the "tragedy of the commons" (Hardin 1968), of course, and it arises as a negative externality because no user is fully liable for the cost of his or her over use (e.g., congestion costs, excessive consumption).

It is important to recognize that crowding and rapid quality deterioration are not the only consequences of common access to a rivalrous good or resource. The dete-

[5] A club good can be rivalrous, in that use by one individual reduces the value somewhat, but not completely, for other users (e.g., congestion). If such costs are born by club members, and therefore, internal to the collective decision-making process of the club, then the club may still provide the good.

rioration in quality often can be offset with appropriate investments in maintenance, but individuals with common access do not have incentives to make such investments because they cannot charge others who consume the benefits or prevent them from doing so (other drivers will add trips on the highway if quality increases), thus creating a positive externality problem in the form of underproduction of maintenance. Indeed, while it might be contended that "non-excludable public goods" and "free-access common pools" are simply two terms for the same concept because the under-investment implications are the same, this inference is inappropriate.[6] Free riders are not paying for something they consume. Under-investors in maintenance are not paying for something that others will consume. Incentives to pay one's own way may be different that incentives to pay for others' benefit. Furthermore, as Minasian (1964, p. 77) explains, the public goods terminology often is "asserted" to imply that non-excludability is an intrinsic problem that cannot be resolved without coercing free riders into paying for the good. In contrast, the common pool terminology emphasizes that incentives arise because of the legal or customary definition of property rights, and therefore, that another rights assignment can alter incentives. To emphasize the distinction, **a common pool is characterized by: (1) non-excludability, (2) rivalrous consumption as a result of congestion and the resulting negative externalities, (3) under-maintenance due to positive externalities, and (4) either production by nature (a resource) or by someone with incentives to provide it free of charge** (often the state, as discussed below).

With these concepts in mind, let us consider the institutional environment that creates the potential for these different types of goods, and then turn to the actual development of such roads in Great Britain.

11.3 Collective Decision-Making Costs, Clubs, Coercive Mandates, Privatization, Rent-Seeking, and Public Provision of Common Pools

11.3.1 Small Clubs and Bargaining

Assume that two individuals, Dick and Jane are the only members of a club that involves joint consumption of a club good. Let the club good be non-rivalrous in consumption and freely accessible to Dick and Jane, while other potential consumers are excluded (e.g., the club has enforceable property rights). Further assume, for simplicity, that both Dick and Jane own resources (e.g., land and labor hours) dictating their individual capacities to produce both private goods and the club good. To maximize utility each individual decides how to allocate resources between production

[6]They can be related because initially a good or resource can have the characteristics of a public good but given the inevitable congestion that arises with free access, it will become a common pool.

of the club good and private goods, but the amount of the club good each consumes is the sum of the total amount produced by both.

Consider the Cournot–Nash non-cooperative outcome as a benchmark for comparison. This solution arises when Dick and Jane independently adopt a set of strategies which establish the best response that each individual can make to the other individual's allocation decision. As is widely known, the Cournot–Nash solution generally is not welfare maximizing. If Dick and Jane cooperate they can produce a combination of the club good and private goods that generates more utility. The "prisoners' dilemma" (Cournot–Nash) model is often used to demonstrate market failure, including under production of public goods. The Cournot–Nash solution is not likely to arise in a small club, however, as both game theory and experimentation demonstrates that cooperation can be the dominant strategy when repeated interactions occur. Repeated dealing creates both a willingness to cooperate and a potential to punish non-cooperative behavior through strategies like tit-for-tat, and others discussed by Ridley (1996, pp. 53–84). A repeated-game does not guarantee unconditional cooperation, as the dominant strategy still depends on expected payoffs, frequency of interaction, time horizons, and other considerations (Ridley 1996, pp. 74–75), but if a small number of individuals (two in this case) are in a club, they have already cooperated because they expect sufficient payoffs given their expectations about the frequency of interaction over their time horizons. In fact, the club may well form because a non-cooperative solution is recognized as Pareto inferior, so if promises are enforceable at low cost within the club (perhaps through tit-for-tat strategies or other sanctions discussed below) and the cost of bargaining is low, these two individuals should achieve a "trading equilibrium" (negotiate and establish an efficient allocation of resources). The result is a Pareto solution for the club members (Dick and Jane), of course, but not necessarily for society as a whole wherein access to the club good by non-members might be Pareto improving. In order to consider this issue let us explore the potential for expanding the club.

11.3.2 Large Clubs and the Cost of Bargaining

While circumstances under which a few (e.g., two) individuals achieve a Pareto equilibrium involving a club good might be envisioned (e.g., low bargaining and enforcement costs, perhaps due to a repeated game), as the number of individuals increases, the transactions costs of bargaining and enforcement rise. Therefore, the size of a club may be constrained by such costs to the degree that there might be external benefits that the club members do not fully recognize and internalize (McNutt 2000).[7] Perhaps a knowledgeable and benevolent coercive authority could impose a system that would improve on a completely voluntary club arrangement?

[7]Key theoretical conclusions from the two person game can hold for an N person model (Milleron 1972; Bergstrom et al. 1986; Bernheim 1986; Shitovitz and Spiegel 1998, 2001, 2002). Specifically, both a unique Cournot–Nash and a unique trading equilibrium can exist, and furthermore: (1) a trading equilibrium must be a Pareto optimum for the N individuals involved in the game

Key questions then become: (1) are the limits on size sufficiently binding so that a club is likely to be too small relative to the efficient scale of production of its club good(s), and (2) if this is the case, will a coercive authority implement a more efficient arrangement?[8] The first question is considered in Sects. 11.3.3 and 11.3.4, and the second in the remaining subsections of Sect. 11.3.

11.3.3 Community Norms as Substitutes for Bargaining

If there are potential net social benefits from expanding production of a club good beyond its current scope, then their clearly are incentives to do so. Thus, as Demsetz (1967) suggests, when externalities become large enough so that the benefits of internalization exceed the costs of doing so, institutional (e.g., property rights) changes are likely. Direct bargaining is not the only voluntary means for internalizing externalities, including achievement of efficient levels of production of a club good. Two alternative institutional developments are discussed here because they appear to be particularly relevant for the historical evolutions examined below: (1) substitution of norms or customs for bargaining, and (2) development of higher order cooperative clusters that allow limited kinds of inter-group interaction (e.g., produce a specific club good for a relatively narrowly focused club made up of smaller clubs).[9]

Let "rules" refer to behavioral patterns that other individuals expect a person to adopt and follow as all individuals pursue various interdependent activities and actions. The rules one individual is expected to follow influence the choices made by other individuals: like prices, rules coordinate and motivate interdependent behavior.

(Footnote 7 continued)
(the summation of the equilibrium Marginal Rate of Substitution for all N individuals equals 1); (2) the Cournot–Nash solution is not a Pareto Optimum (the Marginal Rate of Substitution = 1 for some individuals, so the sum of Marginal Rates of Substitution over all N consumers cannot equal one); (3) the total amount of the club good produced in the trading equilibrium is greater than the amount produced in the Cournot–Nash equilibrium; (4) each individual contributes more to club good production in the trading equilibrium than in the Cournot–Nash equilibrium; and most importantly, (5) the trading equilibrium **is strongly preferred by all** N **individuals** over the Cournot–Nash equilibrium. Therefore, in theory at least, a non-rivalrous club good can be efficiently produced as a result of the voluntary decisions by the members of a club who then have free access to use the good. This does not mean that the trading equilibrium arises in any institutional setting, however. Institutions that facilitate achieving this are discussed below.

[8]Whether there are net social benefits from expansion depends in part on the consequences of expansion on characteristics of club good itself. At some point, as club membership expands, congestion sets in, so further increases in membership has offsetting marginal effects on members' utility: the quantity of the club good can increase as membership increases, while its "quality" decreases due to congestion costs, and higher maintenance costs. Thus, the optimal membership is likely to be finite, and for at least some club goods, quite small, although optimal membership for other club goods may be very large Buchanan (1965).

[9]These institutional developments clearly do not exhaust the possibilities, however, and others may be equally or even more important for other times (e.g., where technologies differ), places (e.g., where different social, political or economic conditions apply) or club goods.

In this context, note that much of game theory implies that individuals calculate the best strategy in each interaction, but in reality, it is often rational for individuals to adopt rules that guide their behavior under many circumstances, in order to reduce decision-making costs (Hayek 1973). Rules of thumb might be adopted, for instance, by individuals who are not able to use conscious reasoning to evaluate every option in the array of available alternatives because there are significant limits on abilities to reason and absorb knowledge (O'Driscoll et al. 1985, pp. 119–122). Community-wide rules also can be adopted to reduce the need for bargaining or other collective decision making activities. Therefore, while the transactions costs of negotiating an agreement rise as potential club membership increases, members may find it useful to economize on such costs by voluntarily adopting rules and avoiding the need to renegotiate with many changing conditions such as an increasing group population. New members voluntarily adopt the club's existing rules.

While the most obvious rules may be "positive laws" created by legislation, there are many other types of rules that may actually be more important guides to most behavior (Ellickson 1991). In fact, within close-knit groups the rules that dominate tend to arise as "customs" or "norms" which do not require explicit codification or backing by coercive threats. As Nee (1998, p. 87) suggests, "Norms are implicit or explicit rules or expected behavior that embody the interests and preferences of members of a close-knit group or community."[10] A key distinguishing characteristic of community-wide customary norms is that a rule of obligation is initiated voluntarily by an individual's decision to behave in particular ways under particular circumstances. Fuller (1981, pp. 227–228) explains that

> 'Where customary [norms do] ... in fact spread we must not be misled as to the process by which this extension takes place. It has sometimes been thought of as if it involved a kind of inarticulate expression of group will... This kind of explanation abstracts from the interactional process underlying customary law and ignores their ever-present communicative aspect.

Habits or conventions often arise as individuals attempt to economize on time and effort required in calculating tradeoffs in similar circumstances (Hayek 1973), for instance, and many of these habits or conventions are repeatedly observed by others who begin to anticipate the behavioral patterns under similar circumstances, and take these expectations into account in various decisions. Such behavioral patterns also can be emulated by others so the norms and corresponding expectations spread through the community (Mises 1957, p. 192). Rules also can be explicitly and voluntarily created through contracting. The resulting contractual rules only apply for the negotiating parties and for the term of the contract, but others may voluntarily

[10] Buchanan (1994, p. 132) refers to the resulting arrangement as "ordered anarchy" and explains that "Much of human activity takes place in a setting described as 'ordered anarchy,' by which I refer to the simultaneous presence of apparent order and the absence of formal law governing behavior. How is such ordered anarchy possible? ... The answer suggested by my argument here is that interacting parties choose to constrain their separate choices in such fashion as to create non-intersecting and therefore non-conflicting outcomes." Others use the term, "customary law" to characterize such arrangements (e.g., Pospisil 1971; Fuller 1981; Benson 1988, 1989, 1999a, 2011).

emulate the behavior by adopting the same rule, resulting in a community wide norm. In other words, customary norms evolve spontaneously from the bottom up, and they are voluntarily accepted.

Consider a hypothetical example. Vanberg and Congleton (1992, p. 420) note that most forms of interaction are not actually characterized by game-theoretic models which assume that the individuals must play. In reality, people often have an "exit" option, and the exit threat can be more powerful than strategies like tit-for-tat under some circumstances. Specifically, Vanberg and Congleton (1992, p. 420) explain that "In practice, the net benefits of exit depend on the availability of alternatives (or more specifically, on the expected payoffs from those alternatives), whether such alternatives exist in the form of potential interactions with other players or in solitary activity." The exit threat is likely to be credible, for instance, when each individual is involved in several different games with different players, in part because the same benefits of cooperation may be available from alternative (competitive) sources. And of course, even in a very primitive setting, individuals are generally involved in at least one close-knit club "community" as described by Taylor (1982, pp. 26–30), wherein "the relations between members are direct and ... many-sided" (also see Bailey 1992 and Ellickson 1993) – i.e., a club.

Given the availability of competitive alternatives, all members of the club have a refuse-to-play option, so they may cut off all relationships with someone who they know has been untrustworthy in dealings with anyone else in the group. And importantly, to the extent that information, or "truthful negative gossip" (Ellickson 1991, pp. 180–182), can travel from one bilateral game to another, the negative consequences on reputation can limit the non-cooperative player's ability to enter into other games with other individuals. This means that an attractive strategy may be adopt a rule of thumb: **unconditional cooperation whenever an individual chooses to enter into some form of interaction, along with exit and the spread of information about any non-cooperative behavior**. Vanberg and Congleton (1992) refer to this response as "prudent morality," and given that reputation information spreads quickly within a group (club) and everyone spontaneously responds to information, the non-cooperative individual is excluded from all interaction with any member of the community. Such spontaneous ostracism can be a very significant punishment, creating strong incentives for individuals in a club to behave cooperatively in every game with other members, whether that game is one-shot or repeated.[11] The boycott response to information becomes a behavioral norm, as everyone is expected to ostracize a non-cooperative individual.

With regard to roads, Ellickson (1993, p. 1372) notes, in examining the historical development of property rights in land, that "affirmative covenants that impose duties" typically evolve as norms. Imagine an agricultural community for instance, where no system of roads exists. Two neighbors may find it beneficial to interact

[11] Vanberg and Congleton (1992, p. 421) suggest that another strategy is unconditional cooperation until or unless non-cooperative behavior is confronted, and explicit punishment of the non-cooperative player as exit occurs. They label this strategy "retributive morality," but such violence is risky, so with competitive options and the ability to spread information, prudent morality tends to be a superior strategy.

on certain dimensions (socially, religiously, and/or economically), so they begin to travel back and forth between their locations. Neither prevents the other from doing so, and a mutual obligation to respect rights of passage arise (e.g., an easement is recognized). These individuals may develop similar relationships with other neighbors and a network of such "easements" develops. Perhaps a central location becomes attractive as a market, a site for a religious structure, or a meeting place (clubs generally meet, after all), and/or perhaps individuals find it beneficial for some parcel of land to be used by the community as a whole (e.g., a community pasture, a hunting area) and everyone travels to it. In addition, perhaps individuals own dispersed plots of crop land in order to reduce their risks or take advantage of different types of land that have comparative advantages in different crops (Dahlman et al. 1980). In order to travel to the central location and/or the common property from the outlying farms, and/or to travel to dispersed plots, individuals have to cross other individuals' land, but since everyone benefits from the interaction that takes place at the central or common location or as a result of dispersed plots, each land owner has incentives to routinely allow others in the community to pass over their land, although probably only along certain routes (easements). People in the community come to expect such rights and customary obligations to allow passage arise. Indeed, as Ellickson (1993, p. 1381) concludes, "a human group invariably opens a significant portion of its territory to public use," recognizing that "public" denotes access privileges only for community members, not state ownership or free access to people outside the community, unless such access is recognized for some outsiders too. The alternatives, bilateral bargains between every traveler and all land owners whose land is crossed, or a community-wide multilateral bargain, involve very high transactions costs.

11.3.4 Club Hierarchies

No community evolves in complete isolation. Parallel localized communities develop that are geographically proximate, making inter-group competition and cooperation possible. Anthropological and historical evidence suggests that inter-group conflict has been an almost ubiquitous characteristic of human history, of course, but cooperative arrangements and inter-group norms also can and often do evolve between members of different groups (Pospisil 1971; Benson 1988, 1999a). Hardin (1982, p. 184) suggests that "Large-group Prisoner's Dilemmas might be resolved as a byproduct of smaller subgroup interactions. But this could be strictly a spontaneous voluntaristic by-product ..." Importantly, however, communities need not formally "merge" and accept an entirely common set of rules governing all types of interaction. Individuals from different communities only have to expect each other to recognize common rules pertaining to the types of inter-group interactions (e.g., trade, road access) that evolve. Given the importance of frequent interactions and reciprocity, trust relationships, and reputation effects, there clearly is a limit to the size of a single close-knit community, but a much larger web of communities can develop for certain particularly beneficial functions if it is "overlaid by a network of much

smaller subgroups, each concerned with its own conventional behaviors with respect to specific subgroup goals" (Hardin 1982, p. 184).[12] Indeed, Llewellyn and Hoebel (1961, p. 53) point out that the traditional western bias of trying to delineate some all-embracing system of governance for a society as a whole can be very misleading (also see Pospisil 1971; Benson 1988; Benson 1999b; and Benson 2011).[13]

Suppose, for example, each localized group's norms regarding travel across other members' lands can continue to govern its own members, while a different set of rules apply for access to roads for certain members of neighboring communities, and even for people from vary distant communities (e.g., merchants, courting swains, religion officials, individuals on religious pilgrimages, allies engaged in joint defense against a common enemy). Even in primitive societies, entrepreneurs establish extensive trade networks that cross community boundaries, for instance Ridley (1996, pp. 195–211), and trade between some members of different local communities may require linking a portion of each community's road system. Perhaps traveling traders could negotiate access on each trip and pay tolls, but incentives arise to encourage members of each group to recognize rights to access to key linkage roads (although perhaps not to all roads) for those engaged in trade or other value-generating activities. Hoebel (1954, p. 122) provides an interesting example, explaining that if an Ifugao (a primitive tribal society in the Philippines) left his home district he would move through a "neutral zone" into a "feudal zone" where, "Permanent feuding relations with certain families in the area are the thing," and then into a 'war zones" where, "Anybody in the area is killed on sight. Head-taking expeditions make their stealthy raids in such areas whenever heads are wanted for purely prestige or religious reasons." Yet, a "courting swain" had customary immunity from attack when traveling outside his home district, perhaps because one way to end a feud was through intermarriage and the entire Ifugao society recognized the advantages of peace (Hoebel 1954, pp. 122–124).

---

[12] Indeed, people often live and interact in many different communities (clubs) with voluntary governance from many different overlapping arrangements (Benson 1999b; Ostrom 2007, 2005).

[13] Various levels of custom can have different content and procedure (Fuller 1981, pp. 241–241):

> That the family cannot easily organize itself by a process of explicit bargaining does not mean there will not grow up within it reciprocal expectancies... Indeed the family could not function without these tacit guidelines to interaction... At the midrange, it should be observed that the most active and conspicuous development of [custom]... in modern times lies precisely in the field of commercial dealings. Finally, while enemies may have difficulty in bargaining with words, they can, and often do, profitably half bargain with deeds... That [customary norms are]... at home across the entire spectrum of social contexts does not mean that [they retain]... the same qualities.... At the terminal point of intimacy [custom]... has to do, not primarily with prescribed acts and performances, but with roles and functions.... In the middle area, [custom] ... abstracts from qualities and disposition of the person and concentrates its attention on ascribing appropriate and clearly defined consequences to outward conduct. Finally, as we enter the area of hostile relations.... the prime desideratum is to achieve - through acts, of course, not words - the clear communication of messages of rather limited and negative import; accordingly there is a heavy concentration on symbolism and ritual.

11.3.5 Implications from Sects. 11.3.1–11.3.4

(1) Voluntary production of club goods is likely for a small group because transactions costs are likely to be low (e.g., due to repeated dealing incentives); (2) larger groups can produce club goods if the group is made up of individuals involved in a multi-dimensional web of mutually beneficial relationships (e.g., due to exit threats and resulting ostracism sanctions); (3) even larger close-knit groups can produce club goods by lowering transactions cost through the substitution of customary rules or norms for repeated bargaining; and (4) if various club goods have different efficient sizes, a hierarchical linking of clubs can evolve, with functionally-focused norms to support production of those club goods which have large efficient scales compared to localized clubs while local clubs develop those goods that are smaller in scale.[14] None of this guarantees that the provision of club/public goods will be universally efficient in a Pareto sense, of course. It simply means that inefficiency is also not inevitable because incentives always exist to internalize externalities by developing a system of interrelated and overlapping clubs. Such voluntary (private) arrangements certainly could take considerable time to evolve, so even if they ultimately will arise, a "market failure" may appear to exist. Institutions might be created more quickly through coercion, so perhaps a political solution can be superior even if the voluntary process might ultimately succeed.

Mises (1949, p. 692) explains that market-failure justifications for coercive government (state) actions "ascribe to the *state* not only the best intentions but also omniscience." He then points out that neither assumption is valid: the state is not purely benevolent since both those who are employed by the state and those who demand state actions have subjective self-interests which may be achieved through the use of coercive power. Furthermore, the state is not all knowing since state decisions are made by individuals, knowledge is widely dispersed across individuals, and the cost of coordination is infinitely high, particularly without market profits and prices as coordinating mechanisms (Hayek 1973). These two assumptions are both relaxed below, in order to get a clearer picture of how and why the United Kingdom's system of road provision evolved as it did.

[14] Transactions costs also imply that there are limits to how extensive an inter-group network of cooperation can be, but there are other reasons to expect that these limits can be broken down if it is desirable. After all, as Mises (1957, p. 257) explains, "Man is not the member of one group only and does not appear on the scene of human affairs solely in the role of a member of one definite group. In speaking of social groups it must be remembered that the members of one group are at the same time members of other groups. The conflict of groups is not a conflict between neatly integrated herds of men. It is a conflict between various concerns in the minds of individuals." For example, a medieval merchant generally was simultaneously a member of the merchant community, a religious organization, and perhaps an urbanized community or neighborhood association, and the geographic dimensions of each varied. Thus, he was in fact familiar with the behavioral rules of several different groups and was in a position to facilitate the development of inter-group ties.

11.3.6 Benevolent Mandates by a Coercive Authority with Incomplete Knowledge

Suppose that an authority has the power to mandate that members of a community contribute to the production of a good that is non-rivalrous in consumption for community members and that all contributors are able to consume the good.[15] Further assume that the authority is benevolent and desires that an efficient level of the good be produced. The task for the authority is to decide how much of the good should be produced and how much each citizen should be required to contribute to its production. Suppose that the authority decides that the "fair" way to pursue production is a Lindahl "public good" allocation process as defined by Samuelson (1954), Foley (1970) and others. The Lindahl equilibrium requires that the total per-unit contribution made (resources contributed to production or prices/taxes paid) by each individual in the community equals the total per unit cost of the non-rivalrous good. These "Lindahl prices" or "Lindahl taxes" mean that every person's consumption decision is based on the share of the cost they must bear in order to obtain the club good. The resulting Lindahl equilibrium is Pareto efficient. A significant problem faces the benevolent authority, however, because in order to set the appropriate Lindahl prices/taxes the authority must know what each individual's demand functions is for all private and public goods. Since the demand determining preference functions are subjective, knowledge of actual demand functions is held exclusively by each individual. Everyone's incentives are to report a lower evaluation of the good than they actually have in order to "free ride." This free-rider problem is the justification often given for government provision, or at least government taxation to fund provision of, goods that are non-rivalrous in consumption, of course, because private producers will be under-paid by free riders, but the problem also undermines efficient public provision since the authority cannot know what individual preferences and valuations are. This preference revelation problem means that the authority cannot determine the efficient level of production or the appropriate Lindahl taxes. It is much more likely that too little of the good (e.g., if the authority accepts what individuals report to be their preferences) or too much of the good (if the authority assumes that individuals are lying, charges higher taxes, and produces more of the good) will be produced than that the Pareto Efficient solution will arise.

Since the actual decision by a benevolent authority cannot be based on the Lindahl process, some sort of approximation is required (e.g., equal taxes, taxes based on wealth which creates incentives to hide wealth). There is an alternative, however, which is much more likely to achieve an efficient equilibrium in which the members of the community reveal their relative preferences. Therefore, a benevolent authority

[15]The good could be rivalrous in the sense that congestion occurs with increased use, thereby reducing the benefits for all consumers, as explained below, but for now, assume that it is non-rivalrous, given the size of the community.

who recognizes the information problem that stands in the way of an efficient solution through coercive taxation should prefer this alternative.[16]

11.3.7 An Alternative for the Benevolent Leader: Privatization

If a benevolent authority really wants to establish a policy that might produce a Pareto superior outcome, one obvious option is to recognize private property rights in the free-access good. Private property rights create the ability to exclude non-payers even for non-rivalrous goods. Movies, concerts, and plays (and numerous other goods and services) are non-rivalrous in consumption, at least up to a point, but they are being provided through the market because people have the right to build theaters and exclude non-payers. In the case of roads, the *existing* roadways could be privatized so that each individual with exclusive ownership rights can charge a toll to any traveler and exclude those who refuse to pay. In this case, rather than relying on customary norms and their associated ostracism sanctions, or on coercively imposed state sanctions, to induce individuals to provide and maintain roads, they could rely on market forces to do so. If private property rights are complete and transactions costs associated with enforcing these rights are not prohibitive, a trading equilibrium should emerge as equilibrium prices are determined and each traveler pays each road supplier.

It should be noted that tolls are not the only way to pay for privately provided roads. A business community (club) may not charge tolls because it builds roads in order to attract customers (e.g., consider Disney World or a mall parking lot) who pay prices for goods or services that cover road costs (i.e., a non-rivalrous good can be bundled with rivalrous goods). Similarly, residential developers may build roads to make their lots more valuable, thereby covering the cost of the roads through the prices charged for those lots. Limits on access also can be achieved by means other than money prices. Thus, for instance, some private residential communities (clubs) discourage through traffic by limiting access to one or a few entrances and/or by installing traffic control devices like speed bumps (Newman 1980). Others place gates at their entrances with either coded locks or security guards. Privatization can lead to the development of club goods if relative net benefits of the market process are less than those of a club arrangement.[17]

---

[16]The knowledge problem facing a benevolent authority is actually much broader than suggested here (Hayek 1973; Mises 1949). The authority also does not have the knowledge required for efficient production, for instance, as such knowledge is generally dispersed through large numbers of individuals.

[17]Note the analogy to Coase's theory of the firm (1937) wherein allocations can occur through markets or through firms if the transactions costs of market allocation are higher than the costs of making allocations within a firm hierarchy. This discussion suggests a third choice - cooperation within a club.

One response to this suggestion of privatization as a method for road provision is that if a private-property-rights-based system of roads is efficient then why did it not arise in England (or elsewhere) rather than club or government provision? One answer is that institutional evolution is path dependent (Benson 2005) so the development of cooperative joint production and non-price rationing (e.g., club-goods or government provision) arrangements were less costly, given the institutional environment, than development of a market system. Essentially, the transactions costs of establishing and maintain private rights in roads and relying on a market system (or mixed market and club good system) for creating a network of roads could be very high. With this fragmentation of land holdings, for instance, the transactions cost arising from creating a linked system of toll roads (even to link local club provided local roads) could be quite high. Even if the transactions costs prevent private creation of efficient road networks, however, once the road network exists, privatization could result in a relatively efficient system of allocation and of maintenance. Given true private property rights to existing roads, mergers between road owners should occur to take advantage of any scale economies in their provision and reduce the transactions costs of actual provision (e.g., a highly fragmented system tends to be costly as travelers must constantly stop to pay small tolls, but such costs would fall with mergers). If excess roads were produced prior to privatization, some would fail to attract sufficient revenues and go out of business. If the network has too few roads, once a market is in place, the profit motive is likely to lead existing road owners to expand and/or new road owners to enter. Therefore, a relatively efficient network is likely to arise. This did not occur, however, so the question posed above still remains. The actual answer appears to be that, just as state action is not likely to be able to achieve the Lindahl equilibrium, state action is not likely to produce a property rights arrangement that allows an unhindered market solution (i.e., pure private property rights). Indeed, there are no true free markets once government becomes involved. Regulations limit entry, mergers, pricing options, location choices, and so on. Subsidies, tax breaks, and bailouts are paid to some producers but not others, some markets are declared to be illegal, and so on. The primary reason is that in general government decision makers are not benevolent.

11.3.8 Relaxing the Benevolence Assumption: Pursing Self-interest Objectives Through Manipulation of Property Rights

As Coase (1960) and Demsetz (1967) emphasize, one motivation for creating property rights (or more accurately, rules of obligation to respect property claims) is to eliminate externalities and facilitate voluntary interaction. Coase (1960) also explains that these institutions determine the distribution of bargaining power and therefore the distribution of wealth, however, and while he does not focus on this issue, these distributional consequences also create incentives to assign and reassign property

rights by manipulating rules. Indeed, as Oppenheimer (1914, pp. 24–25) observes, an understanding of the formation and development of the state requires recognition of the fact that

> There are two fundamentally opposed means whereby man ... is impelled to obtain the necessary means for satisfying his desires. These are work and robbery, one's own labor and the forceful appropriation of the labor of others.... [T]he warriors' trade ... is only organized mass robbery... Both because of this, and also on account of the need for having, in the further development of this study, terse, clear, sharply opposing terms for these very important contrasts, I propose ... to call one's own labor and the equivalent exchange of one's own labor for the labor of others, the "economic means" for the satisfaction of needs, while the unrequited appropriation of the labor of others will be called the "political means."

Importantly, when rules arise through coercive power they can facilitate the pursuit of either the economic or the political means of personal wealth enhancement. In fact, positive law (state made law) almost always involves conflicting efforts to achieve both objectives (Benson 1999a). To illustrate this, consider the development of a kingship, such as those in Europe.

An entrepreneurial tribal leader skilled in organizing joint production of raiding often recognizes that an attractive way to gain wealth is through organized aggression against another community (e.g., the Viking raids in Europe). Plunder tends to produce relatively small long-term returns, however, compared to the wealth that might be extorted over time if productive people are subjugated and allowed to continue their productive efforts in exchange for payment of "protection money." Therefore, an entrepreneurial war leader may advocate invasion and occupation of the territory of the other community, rather than repeated plunder. Oppenheimer (1914) contends that the origins of the earliest states trace to precisely this situation, as nomadic hunting and/or herding communities from the relatively unfertile mountains, desserts, or sea coasts, invaded and subjugated those who had settled in fertile valleys.[18] Successful war leaders who conquer other territories often asserted that they are kings, although the result was not necessarily permanent. After all, conquered subjects' promises to honor an invader are credible because of the fear of violence, so the king has to be forever vigilant in policing existing claims, even as he attempts to legitimize his claim and expand his domain. The internal dynamics of such a coercive wealth transfer system appear to be relatively unstable (Levi 1988, p. 44), but there are ways to reduce internal resistance (Levi 1988, p. 11).

[18]Groups relying on hunting tended to develop improvements in technologies for hunting which enhanced their wealth in the short run, but the long-run effect was often quite different. Many migratory animals were hunted into extinction by primitive groups (Ridley 1996, pp. 227–247), for instance, because ownership could not be established until an animal was killed. However, because the members of the group relying on hunting developed new weapons and other inputs to hunting (e.g., domesticated horses, ships), and became skilled in the use of those inputs, they developed a comparative advantage in violence. Carneiro (1970) agrees but adds that successful creation of relatively permanent states of this type occurred where exit by those being subjugated was very difficult due to the surrounding hostile environment (e.g., oceans, desserts, mountains, other hostile communities).

11.3.9 Public Provision Leads to Common Pools

Kings do not simply try to create a monopoly in violence; they also attempt "to act like a discriminating monopolist, separating each group of constituents and devising property rights for each" (North 1981, p. 230).[19] By transferring property rights to those who might be in a position to threaten their control (e.g., reduced taxes; provision of subsidies, franchises, exclusive licensing, privileges such as access to "public" property and services; granting use and even partial ownership of land), kings can buy their support, while taking property rights from those without power (e.g., setting high taxes; regulating the use of so-called private resources by withholding licenses or franchises; limiting or preventing access to public property and services; and/or explicitly taking resources that private entities had owned prior to conquest). Furthermore, in a dynamic setting where relative power can change (e.g., individuals can organize into groups with collective power), the king has incentives to redistribute wealth as changes occur. Given the kings' use of wealth transfers (changes in property rights) as a low cost mechanism of insuring against competition, groups have incentives to compete for favorable treatment from kings. In cases where a coalition of groups could actually threaten the king's power, he actually has incentives to encourage such "rent-seeking" competition (Levi 1988, p. 12), since by keeping sub-groups divided into adversarial political camps the possibility of a strong coalition forming to challenge for control is reduced. By focusing such competition in "advisory councils" or "representative assemblies," the transactions cost of interacting with various powerful groups is lowered (North 1990, pp. 49–51), and powerful groups also see their interests linked to the interests of a "sovereign" as they have a more direct say in the decision-making process (note that roads and communications networks connecting the king's central location with the outlying locations of his potential rivals are also important in this context (Levi 1988, p. 28)). Significantly, however, in such a political environment property rights are never "given": they are permanently in play, because as the relative power of groups change, some property rights are reallocated. Furthermore, one imposed change in property rights inevitably sets off a long chain of reactions (Benson 1984, 2005). For instance, the king may claim large amounts of property (e.g., William claimed ownership of all land in England after his successful invasion in 1066) but then grant "privileges" in the form of various access and use rights to powerful allies or potential enemies. Because of the instability in the relative power structure, more and more groups have incentives to organize and enter the competition (Benson 1984), but reducing existing privileges undermines support from already active and powerful groups. Thus, more individuals and groups obtain access to public property and services which becomes increasingly plagued by common pool problems. When property rights are subject to authoritarian alterations the result is a continually spiraling race for rents (Anderson and Hill 1990; Benson 1984, 2005) which dissipates wealth as resources are used up in the competitive process of trying to influence the coercive power.

[19] Also see Levi (1988, pp. 10–14).

Another closely related reason for the emergence of common pools arises when rights are significantly altered, or when they become sufficiently tenuous due to frequent changes. As Leoni (1961, p. 17) emphasizes, this has "a negative effect on the very efficacy of the rules and on the homogeneity of the feelings and convictions already prevailing in a given society .... [T]he fact that the very possibility of nullifying agreements and conventions through supervening legislation tends in the long run to induce people to fail to rely on any existing conventions." In other words, the fact that property rights are in play creates uncertainty about the stability of existing rules (obligations), including the security of whatever the property rights assignments might be at any point in time.[20] When this happens, individuals may quit performing previously worthwhile functions, such as road maintenance.

If the function is demanded by powerful groups, the king may try to force the previous behavior, and if that fails then the king (through a bureaucracy) is likely to use tax revenues to directly producing the function. Once the king (or a designated bureaucracy) begins to produce such "public services," access becomes something the king can hand out to supporters. Again, as competition for rents expands, many people obtain access to these services. The result is a common pool unless the service is non-rivalrous in consumption (e.g., see Benson 1994, 1998, 2010 regarding public policing, public prosecution, public courts, and public prisons). As those with free access to a publicly provided good or service increase in number, congestion (crowding) and overuse is inevitable. Rationing of the publicly provided service by first-come-first-serve (queuing) and its resulting congestion also typically leads to in rapid deterioration in quality. Other rationing methods can arise (as they do with club goods when overuse occurs), with quotas, limits on use through licensing, merit allocation, and so on, but when these non-price rationing mechanisms arise through coercion rather than voluntary agreement, the rationing process itself consume resources, in part because they must be enforced.

11.3.10 Implications from Sects. 11.3.6–11.3.9

(1) Even if the authority making decisions about the taxes to collect in order to produce a non-rivalrous non-excludable good is benevolent it is not omniscient; it is not possible to set the efficient levels of taxes and production because of the knowledge problem; (2) if the authority truly is benevolent and wants to establish a policy that might produce a Pareto superior outcome, the authority should allow the creation of private property rights in the free-access good(s) so owners can legally exclude non-payers (even for non-rivalrous goods) and market forces can determine prices and quantities of the product, moving the

[20]The resulting incentives for the use of tenuously owned property are similar to those with common pools. The incentives are to consume the property quickly before the government takes it away. Furthermore, incentives to invest in maintenance and improvement are very weak for the same reason - the current owner cannot be confident that he/she will be able to consume the benefits of these investments because the property may be transferred before the benefits are fully realized.

outcome towards a trading equilibrium; (3) authorities with coercive power are not benevolent, however, so the distributional consequences of coercive property rights assignments/reassignments create incentives to assign and reassign property rights in order to increase wealth of the authority and/or his/her supporters; and (4) more and more groups have incentives to organize and compete for privileges and benefits, and as the population with access to publicly provided goods and services increases, common pool problems arise. All of this suggests that government actions to replace private processes is not likely to achieve anything close to a Pareto solution, so the existence of what appears to be a private-sector (club or market) failure does not justify government intervention, at least from an efficiency perspective. Now let us turn to the early system of road provision in Great Britain where such arrangements actually developed, not once, but twice.

11.4 Roads as Club Goods in Medieval Great Britain

The Romans,[21] who arrived in 43 A.D., built "great military highways" in Britain in order to move their legions into the remote regions of the Island (Jackman 1966, p. 1).[22] While there is little doubt that the Roman roads were important transportation arteries for centuries, they were "by no means so good nor so complete" that a much larger system of other roads were not needed (Jackman 1966, p. 4).[23]

11.4.1 Roads in the Hundreds System

Direct knowledge of the process of development and maintenance of roads in Britain before the twelfth or thirteenth century is almost non-existent, but a good deal can be inferred by considering evidence of the kinds of travel that occurred, and by examining the system of roads and customary arrangements for road maintenance that existed shortly thereafter. With the fall of Rome, Europe moved into a period dominated by localized and largely self-sufficient agricultural communities. Nonetheless, the fact that at least some parts of the road networks were in good condition is evidenced by the records of military marches, some averaging as much as fifty miles per day (Gregory 1931, p. 94). Furthermore, Royal income for Anglo-Saxon kings was mostly in the form of the agricultural output of the royal estates, and in order to consume this income a King and his household had to travel from estate to estate throughout the year (Benson 1998, p. 204). Thus, there is substantial evidence that

[21] This section draws from and expand on Benson (2006).

[22] There is evidence of a network of roads in Britain prior to the Roman conquest, although much of it is indirect (Jackman 1966, p. 3; Gregory 1931, pp. 45–55).

[23] Also see Gregory (1931, p. 94). These other roads were not funded or maintained by the state (Roman roads that survived into the Middle Ages were not maintained by the state either).

these kings and their courts "moved incessantly around the kingdom, occasionally with the army" (Hindle 1982, p. 193), requiring passable roads to carry a "very sizable company" (Stenton 1936, p. 6). While such long distant travel occurred, however, indicating that a system of passable roads linked various parts of the Island, the fact is that the vast majority of road use involved local people traveling short distances (Beresford and Joseph 1979, p. 273):

> Journeys to markets, churches and courts are the principal exceptions to the generalization that most medieval roads were entirely local in purpose with an ambition no higher than to serve the villagers' immediate wants. There was need for lanes to provide access to holdings in the fields; to take loaded wagons to the windmill or to the watermill in the meadows; to reach the woodland with its timber, its fruit and its pannage for swine; to take the flock to the common pastures and heaths. The course of the roads with a purpose so narrow would be determined only by local needs.

Thus, almost all of the benefits of roads were internal to the members of local-close knit communities, and local institutions and customs determined how those roads were created and maintained.

By the tenth century, there was a clearly recognized hierarchical institutional arrangement in Anglo-Saxon England. Blair (1956, p. 232) points out that two of the primary purposes of these organizations were to facilitate cooperation in rounding up stray cattle and in pursuing justice.[24] When a theft occurred, for example, the several "tithings" that made up a "hundred" were informed: they had a reciprocal duty to cooperate in pursuit. A tithing was apparently a group of around ten neighboring families, many of whom probably were kin, while a hundred was a group of around ten neighboring tithing. A primary reason for recognizing reciprocal duties was that these organizations produced a number of valuable benefits more efficiently than individuals could, such as the return of stray cattle, deterrence, restitution to victims of law violations, some forms of credit, and so on.[25] These clearly were close-knit communities with multi-dimension webs of mutually advantageous interactions.[26]

Many functions beyond rounding up stray cattle and policing were performed by the tithing and hundred, including dispute resolution *and* road maintenance. Representatives of each tithing traveled to the hundred court, for instance, which met regularly to resolve disputes (Blair 1956, p. 233). When an individual was charged with an offense against someone in a different tithing, his tithing also had a customary obligation to bring him to the site of the meeting of the hundred for the trial

[24] Both can probably be characterized as club goods since cattle were generally held in community pastures to capture scale economies in herding (Dahlman et al. 1980), and policing presumably produced community wide deterrence effects.

[25] See Benson (1998, pp. 198–203) for more detailed discussion of these organizations.

[26] The hundreds were described in some of the king's early codes, so Lyon (1980, pp. 67, 84) argues that as kingdoms grew kings needed a way to organize local government; thus, they presumably established the tithings and hundreds as local judicial administrative units. However, as Blair (1956, p. 235) points out, such an interpretation is erroneous because it "mistake[s] the nature of Anglo-Saxon legal codes which were not so much concerned with promulgation of new law as with codification of established custom. There is little doubt that the hundred [and tithing] was functioning as a unit" before it appeared in any code.

(Stephen 1883, p. 71). Furthermore, higher order jurisdictions apparently existed, as a dispute between individuals who were not in the same hundred went to a shire court.[27] Importantly, in the context of this presentation, under Anglo-Saxon custom, representatives of each tithing were obligated to travel to the various courts, so local road systems clearly were linked and at least some rights to passage over some of the roads of one local community were recognized for members of other local communities (i.e., interconnecting roads were club goods that were extensive in order to produce another large scale club good: peaceful dispute resolution).

While there is no actual documentation of road maintenance and production before records began to be produced in the twelfth and thirteenth centuries (Webb and Webb 1913, p. 5), several inferences can be drawn regarding what was done. First, land over which a road passed actually "belonged" to the owner of the land on either side of the road, in the sense that if a road was abandoned (e.g., because travelers began beating a different path), it would revert to that landowner (Pawson 1977, pp. 65–66). However, under Anglo-Saxon custom, one of the rights to part of the land (i.e., use of the road) was assigned to the extended community (hundred) as an easement: "the right of passage was a communal right" (Pawson 1977, p. 66). Indeed, the concept of the "highway" initially referred to customary rights-of-passage rather than to the roadway or path itself (Jackman 1966, p. 5). Second, road construction and maintenance did not involve anything like modern highway construction. Individuals had customary obligations to other members of a tithing and hundred to remove any impediments to travel such as overhanging trees, hedges, logs, and perhaps water, through a drainage ditch (Webb and Webb 1913, pp. 6–7; Jackman 1966, p. 4), not to build roads. In fact, the word "road" apparently comes from the Anglo-Saxon word "ridan" (to ride) which may derive from the verb "rid," meaning to free or clear away any obstruction. Third, the members of the hundred had a customary obligation to make sure that all members maintained the roadways over their lands (Jackman 1966, p. 33). The actual need for enforcement was rare (Bodey 1971, p. 14), however, due to the multiple dimensions of reciprocities that existed within these close-knit communities.

The road system of the hundreds was in place through the middle of the eleventh century and it proved to be adequate enough to make "possible a centralization of national government to which there was no parallel in western Europe" (Stenton 1936, p. 21) following the Norman Conquest in 1066. After all, the dramatic increase in centralization required a substantial amount of travel by royal officials such as tax collectors and judges, and by armies when rebellions arose, as well as by politically connected citizens (e.g., Barons, representatives of the major church institutions

[27] Above the shire court there was, apparently, a third level of courts "which were, so to speak, hundreds in themselves" (Stephen 1883, p. 67). Note, however, that the higher level courts were not anything like modern courts of appeal. They were simply increasingly inclusive with jurisdictional rules requiring that a dispute be handled by the least inclusive group that encompassed the parties in the dispute. Also note that the rules applied in these courts were customary rather than royal in origin. A panel of jurors with equal numbers of representatives from the separate tithing supervised the trial, which generally involved oath taking (reflecting the important role of reputation in these communities) or ordeal (reflecting the strong religious beliefs of the time).

such as abbeys and monasteries) who had to visit the royal court. At the same time, however, many of the incentives underlying the hundreds were undermined. For instance, William seized virtually all of the land in England, and while he held many large estates for his own use, he also granted use of large tracts to Barons and the Church in exchange for support. Enclosure of some land which had been controlled by local agricultural communities as open fields and common pastures soon followed (Darby 1973, p. 85). In particular, land granted to the aristocracy, called the demesne, could be enclosed (other types of land were controlled by freeholders who paid rent to the lord, and by the villiens who provided labor to the lords). The Statute of Merton (1236) also permitted the lords to enclose large portions of the "waste," the high woodlands and unimproved pastures that lay in clumps around the arable lands, at the expense of the freeholders and villiens who used such areas as pasture, and as noted in Darby (1973, pp. 98–99), grazing was also significantly restricted in the vast royal forests and parks "in the interest of the chase." With increasing enclosure, the potential for straying cattle was diminishing so the value of this cooperative function of the tithing was also declining. Then, in the 1400s, as wool prices rose relative to grain prices, the landed aristocracy evicted large numbers of tenants and enclosed additional large tracts of land, converting it to sheep pasture from crops and stubble fields upon which cattle had grazed. Hundreds of local villages were abandoned (Darby 1973, pp. 210–211). Many of the remaining kinship groups and tithing were broken apart as people were driven from their traditional homes. In addition, the Normans replaced the Anglo-Saxon restitution-based "man-price" system (wer) with a criminal law system involving fines to and confiscations by the king along with corporal and capital punishment (Pollock and Maitland 1959, p. 53; Benson 1998, p. 205). This withdrawal of the right to restitution had significant implications for the tithing and hundred because it substantially reduced incentives to maintain the reciprocal arrangements for protection, pursuit, prosecution, and insurance, and to participate in the local court system. Indeed, the king's expectation that the local communities would continue to provide policing and prosecution in order to collect revenues and property for the crown, without compensation, proved to be unfounded, leading to a long series of institutional changes (Benson 1998, pp. 205–223). Thus, for various interrelated reasons the hundreds became ineffective or disappeared.

While members of local communities that remained probably still had incentives to maintain roads for local use, the breakdown of the voluntary hierarchical tithing-hundred-shire system apparently produced a growing problem of under-maintenance for some long-distance connections. Local freemen were probably less likely to travel between communities, at least voluntarily, so they had weaker incentives to maintain those arteries that were predominantly for long-distant (inter-community) travel, at the same time that the demand for long distance travel was growing from other sources.[28] As noted above, the demand for long-distance travel due to the activities of the king and his court increased dramatically under the Normans. In addition, representatives of the church with its widespread land holdings also traveled

[28]The Normans did mandate that local freemen travel to royal courts, although there was considerable resistance to such requirements (Benson 1998, pp. 210–212).

extensively, as explained in more detail below. Trade also was expanding throughout eleventh and twelfth centuries (Benson 1989). Most commercial retailing activities took place at fairs during this period, and merchants traveled from fair to fair in order to sell their wares and buy others (Benson 1989), thus requiring increasingly intensive use of some roads (Gregory 1931, p. 95; Willan 1976, p. 13).

Kings claimed royal rights to free passage for themselves and their courts to travel anywhere in their kingdom, as well as for anyone traveling to his court on royal business, and expected roads to be provided for these purposes. This claim was reinforced after William's seizure of land, because, even though he granted fiefs of land to his supporters and others that he wanted support from, he retained a claim of absolute authority over the use and disposition of the land granted to these individuals. Landholders controlled land only as long as they performed their required duties and paid their required fees. Successes in putting down rebellions (e.g., against William's successor William Rufus in 1088 and 1095, and against Henry I in 1101) tended to strengthen this property rights arrangement, so the Norman kings' claim of free passage, was simply, in their minds, a right to pass over their own lands. Not surprisingly, then, of the three groups demanding access to roads for long distance travel, it was not the royal government that took up the road-provision task after the breakdown of the hundreds; it was the religious and merchant communities.

11.4.2 Replacing the Hundreds: Merchants, Parishes, and Monasteries

Numerous examples of merchants and merchant organizations contributing to the construction and/or maintenance of roads, and especially bridges, can be found (Jackman 1966, pp. 15–16; Gregory 1931, pp. 97–98).[29] Some guilds were particularly active in this regard, especially when much of the business of the country was conducted at local fairs, and this was the case until the establishment of more permanent markets. However, some guilds and wealthy merchant benefactors continued supporting bridges and roads well into eighteenth century; as Pawson (1977, p. 73) explains:

> Many private improvements were, of course, carried out purely in self-interest. New roads were built to promote the exploitation of mineral wealth within estates, and to enable landowners to divert existing highways ... Sometimes an economic interest led to improvements in the surrounding area, benefiting everyone.... However, when there was little direct return to those involved in private schemes, there efforts were primarily for the social good. It was illegal for a toll to be charged on a public highway without the consent of parliament

[29] Note that the "commercial community" of this period can also be characterized as a club ruled by customary norms, as merchants established their own participatory dispute resolution forums at each market and fair (Benson 1989). The earliest merchant guilds also arose spontaneously, both to provide protection for foreign merchants who were away from their homes, and to protect against unknown foreign merchants who might take advantage of a local merchant and then never return (Milgrom et al. 1990, p. 4).

> so it was not possible to charge those who benefited from such works except by voluntary means.

Nonetheless, there were actually some very important rewards for such local benefactors. After all, roads played a very significant role in determining the success of a market town (Hindle 1982, p. 207) and those trading within it, so other members of both the local community and the merchant community tended to be very grateful to someone who aided the two communities in this way. Thus, building and maintaining roads and bridges was an investment in reputation. And for Christians, even more significant personal benefits were anticipated.

The medieval Church probably had greater demands for long distant travel than the royal court. For one thing, the Church was a major trader (Bewes 1923, p. 9), and in addition, many of the important fairs were held at priories and abbeys. Furthermore, the Church encouraged pilgrimages (e.g., the road from Winchester to the shrine of Thomas Beckett in Canterbury became known as the Pilgrims Way). The Church also maintained frequent tours by peripatetic preachers and friars, but perhaps the most significant source of Church-related travel was the monasteries, whose scattered estates required constant visits (Gregory 1931, p. 95; Jackman 1966, p. 8). Therefore, the Church promulgated the belief that care of the roads was "a work of Christian beneficence, well pleasing to God" (Jackman 1966, p. 8). This created incentives for private citizens within the Christian community to aid in the maintenance of roads and bridges, and the Bishops' registers throughout the United Kingdom provide ample evidence of such activity (Jackman 1966, p. 16). Indeed, such religious beliefs explain the development of the long-lasting customary obligation that members of local parishes accepted for road maintenance (Jackman 1966, p. 30) after the decline of the hundreds.[30] That is, for the purpose of road maintenance (but not for many other functions that had been performed by the hundreds), parishes replaced the hundreds system in the production of this club good, with the aid, encouragement, and where necessary, supervision of the monasteries and bishops of the church. Indeed, and importantly in this context, the monks also accepted a customary obligation to maintain roads, willingly taking on the task because it "was a pious work highly to be commended" (Jackman 1966, pp. 30–31). Thus, the merchants, **and** especially the monks, tended to supplement the parishes where local incentives to maintain roads were relatively weak due to substantial long-distant traffic.

The various local, religious, and merchant communities who established and maintained roads in the United Kingdom prior to 1500 were apparently quite effective, given the technology available. Indeed, the "essence of a modern road pattern existed in the early fourteenth century" and transportation of goods and passengers "could be easily and efficiently undertaken by road" at least throughout southern England and the Midlands (Darby 1973, pp. 174, 287). This system of voluntary road maintenance was also ultimately undermined, however, as a consequence of the almost continuous struggle for power between the English kings and the Church. Henry VIII finally dissolved the monasteries in 1536-39, divided their properties,

[30] See also Pawson (1977, p. 68).

and transferred them to "a class of rapacious landlords who would be slow to recognize any claim upon their rents for the maintenance of roads .... The inevitable result would be a rapid decadence of many highways which had hitherto been in common use" (Jackman 1966, p. 29), also see Gregory (1931, p. 96), and Parkes (1925, p. 7). Local parishes continued to maintain roads in many areas, particularly for local travel (probably 80–85 % of the actual roads in Great Britain), and various merchants and guilds also continued to provide support for some roads and bridges near market towns, but the elimination of the monasteries was apparently quite significant with regard to the roads used for long-distant travel. These roads began to deteriorate (possibly 15–20 % of the roads). Indeed, Jackman (1966, pp. 30–31) contends that the seizure of the monasteries was the primary factor leading to passage of the "Statute for Mending of Highways" in 1555 which mandated that parishes establish a very specific institutional arrangement for maintenance of **all** roads in each parish.

11.5 Evolving Road Policy in the United Kingdom

After seizure of the monasteries, parishes continued to maintain many roads used by parish members but without the help and encouragement of the monks they were often unwilling to maintain the heavily traveled arteries, at least at a level that was satisfactory to many who wanted to use them, including representatives of the state.[31] Therefore the Statute for Mending of Highways (1555) simply ordered the parishes to do, by themselves, what they had been doing with the help and encouragement of the monks.

11.5.1 The Mandated Parish System

Local justices-of-the-peace (JPs) were ordered to appoint two parish surveyors of highways, chosen from a list provided by each parish.[32] The surveyors were ordered to travel the parish at least three times a year to inspect the roads and bridges, see to it that landowners were keeping roads and ditches clear of impediments and announce before the church meeting any violators of the statute. They were also required to collect and account for the fines, compositions and commutations that arose as a result of the failure of individuals to contribute their required inputs (discussed below) to highway maintenance. The JPs were to audit the surveyors' accounts, hear pleas of excuse for non-fulfillment of the statute's input-contribution requirements, levy fines

[31] This section also draws from and expands on Benson (2006).

[32] The JP office was created in 1326 with a mandate "to keep the peace" (Stephen 1883, p. 190). Appointed by royal commission for each county, JPs were to pursue their duties without monetary compensation. Over thirty statutes instituted between the late fourteenth and the middle of the sixteenth centuries establishing additional functions for JPs, including those dealing with the road maintenance.

and order seizures for violations, and when necessary, collect a tax from the parish residents to cover an extraordinary expense. Both the JPs and the surveyors were to perform their tasks without compensation. All of the manual labor, tools, horses and carts needed for repairing the roads were to be provided by the parishioners, also without any compensation. Specifically: "Every person for every plough-land in tillage or pasture" and "every person keeping a draught (of horses) or plough in the Parish" had to provide a cart with oxen or horses, the necessary tools, and two men annually to work four eight-hour days (raised to six days in 1563) in road maintenance on the days chosen by the surveyors. Those households which did not own farm land, horses, or a plough were also required to provide labor, either in person or hired, for the same period.[33] As Parkes (1925, p. 8) notes, however, "Though an elaborate system, it neither sought to introduce any effective method of repair nor took heed of the frailty of human nature."

The mandated obligations of the highway statute of 1555 were largely unnecessary for the roads over which travel remained largely local, and for roads that were heavily used by travelers who did not live in the local community (particularly in the area of London), they were largely unsuccessful.[34] On these heavily traveled roads, traffic by government officials, by freighters using heavy wagons or long pack trains, and by cattle herds being driven to markets "kept the roads in a perpetual slough" (Parkes 1925, pp. 6–7). The burdens placed on the parishioners seemed to them to be very inequitably distributed (Webb and Webb 1913, p. 29).[35] As a result, many did not show up for the mandated work, others sent children or some other substitute instead, and those who did present themselves for work, "often poor men who could ill afford wageless days – would spend most of their time in standing still and prating, or asking for largesse of the passers-by ... so that they became known as The King's Loiterers, in derision of their earlier title, the King's Highwaymen" (Parkes 1925, p. 9). Therefore, JPs were obliged to collect large numbers of fines (Willan 1976, p. 3).

A long series of additional statutes attempted to create sufficient negative incentives to induce the parishioners and surveyors to do their mandated duties. Ultimately none worked and the system of fines evolved into commutations to be collected from individual parishioners that relieved their obligations to perform the statutorily mandated duties and allowed the JPs to hire laborers to work under the supervision of the surveyors (Pawson 1977, p. 71; Webb and Webb 1913, pp. 20–21). These funds also proved inadequate for the heavily traveled arteries, however: "Indeed, what with the lack of any definite valuation roll or fixed assessment, the complications and uncertainty of the law, and the unwillingness of both Surveyors and Justices to be at the trouble of legal proceedings against their neighbors, it is plain that under the commutation system the greatest inequality and laxness prevailed" (Webb and Webb 1913, p. 36). Thus, commutations were supplemented with a general highway tax

[33] See Webb and Webb (1913, pp. 14–26) for more details on this statute and others which followed.

[34] See Albert (1972, p. 8), Darby (1973, pp. 290, 372), and Pawson (1977, pp. 68–69).

[35] Indeed, these cost were often made even higher because the best time of the year for road repairs was also the busiest time of the year for most parishioners since they were engaged in agricultural production (Parkes 1925, p. 9).

from the mid-seventeenth century onward. However, an even more important source of funds was generated through the criminal law with fines levied by the royal courts through presentment or indictment of the parish as a whole for the non-repair of its highways (Webb and Webb 1913, pp. 51–61). Some parishes were perpetually under indictment, and "At varying dates in the different Counties, but eventually ... nearly all over England, it became the regular thing for a parish periodically to find itself indicted at the Sessions for neglecting to keep its highways in repair" and to pay a substantial fine rather than repair the roads (Webb and Webb 1913, pp. 53–54). Despite these sources of revenues, however, the quality of road and bridge construction and repair on the major arteries did not compare to what had been done under the supervision and encouragement of the monks in the previous centuries (Parkes 1925, p. 30).[36] Part of the problem was that surveyors, typically farmers who served for a single year, had no expertise in organizing road repairs and no incentives to see that it was done well (after all, some other farmer would be responsible for taking care of the problems next year if they were not completed), in contrast to the monks who had specialized in such activities and considered them to be long-term obligations to God. In addition, the mandated repair procedure (e.g., periodic large scale efforts rather than ongoing repairs as damage began to appear) was not an efficient way to carry out the task (LaMar 1960, pp. 8–10).

The failure of the mandated parish system to maintain the major long-distance arteries left parliament with relatively few options. One that was tried was a long series of regulations defining "unreasonable" uses of the roads, establishing limits on weight and the number of horses, and so on (Pawson 1977, pp. 74–75). That was an attempt to ration the commons through various restrictions on how it could be used. Surveyors and JPs were expected to enforce these laws, but they were reluctant to do so. A second and more important approach was loosening central government control over and claim to tolls.

11.5.2 Toll Roads and Turnpike Trusts

The right to charge a toll in the United Kingdom had been severely restricted, in part because tolls were an important source of royal revenues (Jackman 1966, p. 11). Kings had long required that tolls be collected from travelers who crossed certain bridges or used some roads. These revenues were not earmarked for road maintenance, however, so they went into the general treasury. Officials who collected tolls also retained a portion for their own purposes, but those purposes rarely included road maintenance. Kings (and later parliament) had the power to grant the right to collect tolls to private individuals or organizations, although they were reluctant to do so for fear of losing this source of revenues. Nonetheless, there is evidence that burgesses (merchants who formed local governments in market towns) in several politically important communities had petitioned for and been granted the right to

[36] Also see Jackman (1966, pp. 48–49).

collect tolls as early as 1154 (Jackman 1966, pp. 9–11). Furthermore, there was one situation under which tolls could be collected by a private citizen without getting government permission: land owners could charge for passage through private land as long as an easement (customary or mandated) had not already been established. Not surprisingly, there is considerable evidence that enterprising land owners began to establish new "private roads" that allowed travelers to avoid the "ill-repaired public highways" (Pawson 1977, pp. 73–74), charging tolls for access. This option was severely limited, however, both by the fragmentation of land and by the fact that easements through many feasible routes already existed. Members of several parishes recognized that these private toll roads suggested an alternative way to finance their required road maintenance activities, however, and the early market-town toll roads provided clear precedent for granting the right to limit access and charge tolls. Thus, politically influential individuals and groups in some parishes began petitioning parliament for the right to collect tolls in order to finance maintenance of certain heavily traveled arteries, and a long series of parliamentary acts were passed beginning in 1663 enabling the establishment of local *ad hoc* bodies known as "Turnpike Trusts." It must be emphasized that these were not parliamentary innovations. The initiative was always at the local level (Albert 1972, p. 12), as parishioners had to petitioned parliament for establishment of a trust for each segment of road on which they wanted to exact tolls.[37]

After about 1700 the turnpike-establishment process became fairly standardized. A group of local landowners and/or merchants would accumulate the money necessary to fund pursuit of a Turnpike Act in parliament and to carry the cost of the trust through its start-up period (Moyes 1978, p. 406). Each Turnpike Acts established a Turnpike Trust and granted it an exclusive right to operate a road (generally for 21 years), fundamentally altering the customary rights-of-passage for most travelers (but see discussion of exemptions below). Trustees did not have complete private rights to the roads they were to operate, however. They were responsible for erecting gates to collect tolls, appointing collectors and a surveyor to supervise repairs and a Clerk and Treasurer to administer the trust, but the trustees were required to be unpaid. The tolls to be charged for various types of traffic were often specified in the legislation, and the funds collected could only be applied to the road named in the Act. No revenues could be diverted to other uses or retained as profit. If start-up costs were significant a Trust could mortgage its future tolls, and many did so, with long-term loans. If the tolls were insufficient to cover costs at particular times (e.g., up front), trusts were allowed to borrow more at a rate of interest fixed by the Act.

Turnpikes were usually existing roadways, although new roads were also built, particularly after 1740, and more importantly, the extent of "usable" roads for heavy traffic expanded significantly through turnpike creation (Webb and Webb 1913, p. 144). The early turnpikes were maintained using the same techniques as the monasteries and parishes had employed before (Darby 1973, p. 374), but much more intensively (Pawson 1977, p. 107). Thus, the quality of the turnpikes tended to be sub-

[37]For detailed discussions of Turnpike Trusts, see Pawson (1977), Webb and Webb (1913), and Albert (1972).

stantially higher than the previously free-access roads they were established on. Furthermore, trusts employed paid surveyors who, thorough specialization, developed expertise in road maintenance, and after about 1750 there is considerable evidence of experimentation and innovation in construction and maintenance by some of these specialists. Webb and Webb (1913, p. 144) note, for instance, that

> Between 1750 and 1770, when the number of Turnpike Trusts was actually trebled, the contemporary self-complacency over the new roads rises to dithyrambic heights, "There never was a more astonishing revolution accomplished in the internal system of any country,' declares an able and quite trustworthy writer in 1767, 'than has been with the compass of a few years in that of England. The carriage of grain, coals, merchandize, etc., is in general conducted with little more than half the number of horses with which it formerly was. Journeys of business are performed, with more than double expedition.... *Everything wears the face of dispatch* ... and the hinge which has guided all these movements and upon which they turn is the reformation which has been made in our public roads [the turnpikes]."

Innovations in surfacing, road widening and banking (Webb and Webb 1913, pp. 133–134), and later, improvements in administration (primarily through the combination of small turnpike trusts into larger administrative units supervises by professional road managers/surveyors, as discussed below) all made travel in the United Kingdom faster and less expensive.

As the preceding quote suggests, Turnpike formation really accelerated during the 1750s (and actually, during the 1740s as well), so by 1770 Trusts controlled almost 16,000 miles of turnpikes (Moyes 1978: 407). In this regard, note that the period of rapid expansion in turnpikes (1740–1830) involved a dramatic increase in heavy long-distance traffic due to the industrial revolution. Indeed, the early period of the industrial revolution was supported by turnpike road (and to a degree, by water) transport, rather than by the railroad system that often seems to get credit for supplying the transportation needs of the revolution (Pawson 1977: 338).[38] As

[38]The correspondence between the timing of the turnpike era and the beginnings of the industrial revolution is more than accidental. As Webb and Webb (1913, pp. 143–144) explain,

> With the coming the Industrial Revolution, with a rapidly increasing population, with manufactures ready to leap from the ground, with unprecedented opportunities for home and foreign trade, improvement of communication between different parts of the kingdom became, from the standpoint of material property, the most urgent requirement. Today, the railway and the tramway, the telegraph and the telephone, have largely superseded roads as the arteries of national circulation. But, barring a few lengths of canal in the making, and a few miles of navigable river estuaries, it was, throughout the eighteenth century, on the King's Highway alone that depended the manufacturer and the wholesale dealer, the hawker and the shopkeeper, the farmer, the postal contractor, the lawyer, the government official, the traveller, the miner, the craftsman and the farm servant, for the transport of themselves, and the distribution of their products and their purchases, their services and their ideas.... And all contemporary evidence indicates that, what with the surface-making and embanking, widening and straightening, leveling and bridging, the mileage of usable roads was, by the eighteenth-century Turnpike trusts, very greatly extended.

Indeed, the tremendous increase in economic activity that began during the mid to late 1700s could not have occurred without the simultaneous improvements in transportation. Furthermore, the development of the British railroad system did not really begin until the 1820s as the turnpike

explained below, the growth of railroads did not begin until the turnpike system peaked. Indeed, aspects of the industrial revolution also helped lead to the demise of the turnpike system as advocates of competing modes of transportation, including the developing railroads, and shippers who wanted reduced their own transport costs (part of which was the tolls they had to pay), manipulated the political process to undermine the ability of turnpike trusts to form, operate efficiently, and/or collect sufficient revenues to maintain their roads. Thus, turnpike activity peaked in about 1830 when there were 1,116 Turnpike Trusts operating 22,000 miles of roads (Roth 1996, 176), and declined thereafter, but the decline was due to political factors rather than purely economic factors.

11.5.3 The Decline of the Turnpike System

The Turnpike era came to an end due to a combination of at least three factors. First, the politically mandated structure and characteristics of the trusts created significant principal-agent problems. The Trustees were not allowed to be paid or earn profits, so other income generating activities (farms, businesses) commanded most of their attention, and they generally were not interested in the day to day operation of the road. Toll gates were farmed out, and while trustees were supposed to monitor the gate-keepers and surveyors, their incentives to do so were very weak. Furthermore, there was no threat of takeover when a trust was operated inefficiently, so the competition for control that regulates managerial behavior in modern for-profit corporations was not at work. With little monitoring and no competitive threat, corruption increased, becoming widespread, "and only a small part of the money collected for the upkeep of the road was in fact used for that purpose" (Hindley 1971, p. 63). Many small trusts had borrowed excessively using long-term mortgages of toll revenues, but then because of inefficient management and increased competition (political and market), they were unable to meet their debt payments.

Second, the political limitations on trusts also led to significant complaints by shippers and travelers. While they probably did not want to pay tolls at all, that may not have been the most significant cost imposed by the turnpike system. A serious complaint was that there were too many toll booths, requiring too many stops, thereby slowing transportation services unnecessarily. Gregory (1931, p. 193) suggests, in fact, that this was the most important complaint against the turnpikes, concluding that: "Road users declared that they would rather pay twice the amount if they could be saved the annoyance of the delay." This problem resulted from the fact that most of the turnpike trusts controlled only short sections of roadway within a parish, so travelers had to pay new tolls each time they left one trust's road and entered another (Webb and Webb 1913, p. 177). While consolidation of small trusts was desirable, the

(Footnote 38 continued)
system was nearing its peak (Pawson 1977, p. 8), and well after the beginnings of the industrial revolution.

trusts operated at the prerogative of parliament, and any formal consolidation required parliamentary approval. Some efforts were made to obtain parliamentary approval to combine small trusts into larger organizations, particularly after the reason for doing so was articulated by John Loudon Macadam, beginning around 1810, but parliament did not respond with necessary enabling legislation that might have led to widespread consolidation, choosing instead to deal with such merger proposals individually and quite slowly (Webb and Webb 1913, pp. 177–180). The cost of influencing parliament combined with political resistance to consolidation (e.g., by local trust employees such as toll collectors who did not want to lose jobs, and by alternatives modes trying to reduce competition from more efficient turnpikes, as explained below) meant that the vast majority of the small trusts remained independent until their bankruptcy and demise.

Third, there was significant political opposition to the trusts themselves. Opposition came from those involved in competitive transportation modes such as the river and canal barges and railroads (see the discussion below), from the trade centers that already had effective transportation connections and feared competition from other centers if their road connections were improved, from some landowners and farmers who feared that better roads would make it easier for their low-wage laborers to be attracted away, from farmers who supplied local markets and feared that improved roads would bring in competition from distant suppliers, from heavy road users who did not want to pay tolls for access even though they wanted the roads to be maintained, and so on. Therefore, in order to gain sufficient support for passage, Turnpike Acts always had to reflect significant political compromise, including long lists of toll exemptions for powerful individuals and groups who opposed each Act (Albert 1972, pp. 12, 14–29). Agricultural interests and in some areas, industrial groups, were particularly effective at obtaining exemptions (Jackman 1966, pp. 260–261). Often those with exemptions were some of the worst abusers of what to them remained a common pool resource. Exemptions also grew over time (individual Trust Acts were annually renewed, with revisions possible), seriously reducing trust revenues (Jackman 1966, p. 261). Politics, rather than economic considerations, also determined the tolls that could be set. Thus, for instance, "There was no invariable relation, and no necessary connection, between the amount that it cost to keep a particular mile of road in repair, and the amount that could be collected in tolls" (Webb and Webb 1913, p. 216). Indeed, just as some road users who did considerable damage to roads were exempted, prohibitively high tolls were established for some types of transportation that did little damage, if that transportation option threatened the market for other politically influential road users or other transport modes such as railroads.

11.5.3.1 Politics and the Inefficiency of Transportation Systems: An Example

The inefficient allocation of transport services resulting from political manipulation of tolls can be seen by examining parliamentary treatment of the steam powered carriages which began to appear on the roads of the United Kingdom in the early

1800s (Fletcher 1891).[39] Indeed, while it is widely believed that the general use of mechanically propelled road vehicles began in the late nineteenth century, the fact is that sophisticated steam powered road vehicles had both commercial and technical success sixty years earlier.[40] These vehicles could maintain high sustained speeds relative to horse drawn carriages (24 miles per hour over four miles, and an average of 12 miles per hour over longer distances) and carry more passengers (up to 14 in 1831). Estimates of relative operating costs suggest that steam carriages could run at about a half to a third of the cost of horse-drawn stage coaches,[41] and in the absence of discriminatory tolls, per passenger fares were apparently about one half those of stage coaches.[42] These vehicles were also much safer as they were much less likely to overturn, and steam engines did not "run away with" passengers the way horses could (Gurney 1831, p. 20). Steam carriages also threatened railroads, which generally were granted monopolies over particular routes (Dalgleish 1980, p. 117), allowing them to charge relatively high prices for passenger services. After all, the steam carriages could compete in terms of speed and they were not limited by the need for rail lines. These competitive threats were not allowed to develop, however, as Parliament responded to political demands from railroad and horse carriage related interests by limiting the potential for competitions from steam carriages.

Where parliament allowed steam carriages they mandated tolls that were at least six times higher than those on horse-drawn stage coaches (Gurney 1831: 22; Dalgleish 1980: 117).[43] Furthermore, parliament imposed outright prohibition of steam carriages in a large number of Turnpike Acts (Dance 1831, p. 48). These very high tolls and prohibitions were imposed despite the fact that "highway engineers were unanimous that injury to the road surface from the action of horses' feet exceeded that caused by the wheels of traffic by a factor of three" (Dalgleish 1980, p. 119). Steam carriages had innovative braking systems that did not lock and drag, as well as one driving wheel with the potential of engaging a second to prevent slippage, both of which did less damage to roads than horse-drawn carriages. Furthermore, the wheels on horse-drawn vehicles were necessarily made narrow to reduce the effort required of the horses, and these narrow wheels caused considerable rutting. Steam carriages, on the other hand, had very wide tires in order to give them greater traction, and these wide tires did virtually no damage to road surfaces, according to engineers such as Thomas Telford (a leading engineer and famous road builder who co-founded the Institute of Civil Engineers and was its first President) who testified before a Parliamentary Select Committee convened in 1831 to consider the exorbitant tolls on steam carriages and to consider the potential future use of mechanical (steam and petroleum powered) vehicles (Dalgleish 1980, pp. 118–119).

[39] Fletcher (1891) provides a very detailed discussion of the development and technological advances in steam powered road vehicles, including information about both the successful and unsuccessful entrepreneurs and inventors involved.

[40] See Gurney (1831, p. 12), Dance (1831, p. 45), and Dalgleish (1980, p. 117).

[41] See Gurney (1831, p. 18) and Dalgleish (1980, p. 122).

[42] Gurney (1831, p. 12) and Dance (1831, p. 45).

[43] See, for example, Gurney (1831, p. 22) and Dalgleish (1980, p. 117).

In light of their safety, cost advantages, speed, capacity, and reduced road damage, the 1831 select committee recommended that the tolls on steam carriage be dramatically reduced (Select Committee on Steam Carriages 1831; Gurney 1831), and if this had occurred, there is "little doubt that a network of good toll roads would have soon been built to take the new vehicles" and that a substantial part of the United Kingdom's railway system would not have been built (Dalgleish 1980, p. 128). However, as Dalgleish (1980, p. 125) notes: "we can well imagine what happened. The many interests – corn merchants, harness makers, horse-copers, railway promoters, iron masters hoping to make rails, and those who were simply against change – would unite against steam carriages. It was only necessary for parliament to do nothing for them to be killed off, and nothing is what it did."[44] As a result, the use of mechanical vehicles on Britain's roads was delayed for some 60 years.[45]

The success of the railroad- and horse-carriage-related interests allied against steam carriages appears to be an important reason for the demise of the turnpike

[44]Dance (1831, p. 46) also notes that coach proprietors, coachmen, and postboys were in the opposition.

[45]The steam carriage industry did not give up. For instance, at least one group including Thomas Telford initiated an effort to run steam-carriage services on their own improved road between London and Birmingham, with intentions of extending the services beyond this route (Dalgleish 1980, pp. 125–128). This group organized the "Steam Company," surveyed the route, and gained support from innkeepers and canal operators (who hoped to compete with railroads by connecting with the steam carriages). The railway serving the route objected strongly, but the group apparently was relying of Telford's prestige to carry them through parliamentary approval. Telford died in September, 1834 however, and the project was abandoned. Yet another initiative by the advocates of steam-powered road travel was the formation of the "Institute of Locomotion for Steam Transport and Agriculture" for the purpose of pursuing the application of steam power to transportation, agriculture and other economic purposes through both economic and political means (Gordon 1833, p. 1). Their political efforts to alleviate the restrictions on steam carriages clearly continued after the 1831 Select Committee report. See for instance, the report of the Select Committee on Mr. Goldsworthy Gurney's Case in 1834; Gurney was an active advocate and promoter of steam carriage transportation (Gurney 1831), but to no avail. There were also numerous additional efforts to develop steam-powered transportation (Fletcher 1891) but politically imposed limitations also continued to be established, even as tolls were eliminated, as discussed below. A significant blow to the development of horseless road transportation in England came with passage of the Locomotive on Highways Act of 1865, for instance, often referred to as the "Red Flag Law". The Act stipulated that all self-propelled vehicles on public highways in country areas be limited to a maximum speed of four mph (two mph in towns) and that they be preceded by a man on foot carrying a red flag or lantern. However, this was just one of the many actions taken to prevent the introduction of steam carriages in England. See Fletcher (1891, pp. 279–288) for details. Indeed, he laments that "All the high-speed engines of recent times have been built for service in foreign countries - our foolish and meddlesome laws prohibiting sensible speeds in this country - hence Russia, Greece, Turkey, India, Ceylon, France, New Zealand and Germany are all ahead of Great Britain in this matter" (Fletcher 1891, p. 257). England was far ahead of the rest of the world in the development and improvement of road vehicles (using steam power) at the beginning of the 18th century and the advantage continued for some time, but the political resistance to horseless transport on roads undermined these advantages, leading to a shift in innovative activity to other countries where the internal combustion engine was favored over steam. The early 1900s saw around 125 different manufacturers producing steam cars in the U.S., however, and one, the Stanley Motor Carriage Company, remained very competitive with internal combustion cars in the U.S. until Henry Ford's development of mass production methods (StanleyMotorCarriage.com 2016, p. 1):

trusts, because it meant that the highways were not allowed to encourage the development of an option that could be competitive with the developing railroads. Indeed, the horse-drawn alternatives proved to be inferior to railroads in head-to-head competition. With the development of the short lines between Stockton and Darlington in 1825 and then between Liverpool and Manchester in 1930, for instance, stage coaches, postchaises and private horse-drawn carriages passenger traffic by turnpike between these points declined dramatically. The turnpikes had come to depend on such passenger traffic for revenues, in part because so many other forms of traffic had more significant toll exemptions or limitations. Ironically, for the horse-drawn passenger service and its supporters who joined the railroads to prevent the development of the steam carriage industry, the advantage going to the railroads quickly led to the decline of their industry as well. Without the steam carriage as a more effective competitor for the railroads, "The transfer of this business was instantaneous and complete. Every coach had to be taken off the road the moment the railway was open to the towns along its route" (Webb and Webb 1913, p. 215). As the railroads spread, road traffic declined. The last stage coach between London and Birmingham went out of business in 1839, for instance, with other routes from London ending their runs over the next few years (to Bristol in 1843, Plymouth in 1847, Bedford in 1848). Thus, turnpike toll revenues fell by one third between 1837 and 1850 as railroads spread through the country while steam carriages were not allowed to offer competition. More and more Trusts were unable to maintain their financial solvency,

(Footnote 45 continued)

> While their car looked similar to most others, their real advantage was simplistic automation. Their early steam engine boasted 13 moving parts with the count for the entire car at 37. It was light, quiet, and perhaps the most powerful vehicle of its time; definitely the fastest. Once lit, the car automatically generated steam to meet demand with little additional attention required except perhaps watching the water level. All that was required of the driver was to set the throttle to a comfortable speed and to move the tiller for steering.
> Steam engines are the only engines (or motors) that generate maximum power from rest. Through the simple movement of a lever the power was precisely controlled. With their finicky ignition systems, balky carburetors, and gear-grinding transmissions the "internal explosion engines" as the Stanley's called them were no match for the simplicity, reliability, and power of steam. The thrashing, banging, clattering, and smell of the internal combustion car was no match for the discernable hiss of a Stanley burner and the sound of tires rolling over stones on the dirt roads of the era. The smell of raw gasoline, partially burned hydrocarbons, along with a mechanical complexity of hundreds of parts for the engine and transmission not to mention the car's audible noise further tarnished the early image of the gasoline-powered automobile. In America the steam car gained popularity, especially with the rich, and Stanleys became the premier steam cars to own.

Stanleys were the most popular car in the U.S. from 1900 to 1904, and they set numerous land-speed records in January 1906, reaching 127.6 miles/hour on January 26. 86 major models were produced with six different body styles. About 11,000 Stanleys were sold over roughly 25 years, with peak production of 750 in 1907. The company consciously chose not to compete with Ford, however, and in 1914 twice as many Model Ts were produced in a day than the annual output of Stanley Steamers. Model Ts also sold at about 25 % of a Steamers price. The company was sold in 1817 and production ended in 1924.

forcing defaults on debt payments (for many, creditors took immediate possession of all revenues to cover interest on bonds, leaving no funds to cover maintenance costs).

11.5.3.2 Public Takeover of Roads

Rather than recognizing the underlying incentive problems and lifting the imposed constraints that created them (e.g., allowing trusts to earn profits and charge market determined tolls for steam carriages as well as other forms of road transportation, allowing competition for ownership to develop and mergers to take place) parliament began to empower the trusts to draw on "statute labor" (i.e., the labor that the parishioners were mandated to provide under the 1555 highway statute). Initially the trusts were required to pay wages fixed by parliament, but later a portion of the labor was required without payment (Hindley 1971, p. 62). Some trusts were even given parliamentary authority to appropriate materials without payment. Furthermore, under the law, the parishes had never been dissolved of their liability for road maintenance, and while they were supposedly able to recover any money they spent from the turnpikes' revenues, the trusts that failed in their road maintenance were generally so far in debt that parishes had little chance of repayment. As chronic insolvency spread, the burden of maintenance for more and more turnpikes was shifted, once again, onto the parishes. The same incentives were at work in the parishes that existed at the beginning of the turnpike era, of course, so this simply led to resistance (shirking, etc.), which actually flared into a "conclusive popular rebellion known as the Rebecca Riots" in South Wales during 1842–1843 (Webb and Webb 1913, p. 217). This resulted in a royal Commission to inquire about the grievances in South Wales, and finally to the dismissal of all Turnpike Trustees throughout the area and the merger of all trusts into "County Road Boards" which took over the roads, their debts, and the tolls of the former trusts. A "General Superintendent of County Roads in South Wales" was appointed by the central government, putting the roads in the area "Under what was virtually Government control" which also loaned 218,000 Pounds to the counties so they could pay off creditors and consolidated debts at low interest rates (Webb and Webb 1913, pp. 219–220). These county organizations were able to substantially reduce the number of toll gates as well as the level of tolls that had led to the revolts. Furthermore, they actually were able to operate efficiently enough to pay off the accumulated debts over the next 30 years, suggesting that if Parliament had responded earlier to the need for consolidation, the riots and subsequent government control might have been avoided. In fact, if consolidation of a similar "sort could have been done with the English Turnpike Trusts in 1844, they might have been spared the long-drawn-out agony of the ensuing half-century. But every attempt at legislation was defeated .... So far as the government was concerned, under the timid and unresourceful advice of the Home Office, and the refusal of successive Cabinets to trouble themselves about the subject, the Turnpike Trusts were allowed to go on just as before, annually getting their expiring terms renewed by Parliament, as a matter of course, falling, most of them, progressively further and further behind

their task, and many of them, deeper and deeper into insolvency" (Webb and Webb 1913, p. 220).

A Select Commission of the House of Commons was formed in 1864 to consider the question of how to end all tolls. The Commission's report concluded that the tolls were "unequal in pressure, costly in collection, inconvenient to the public, injurious as causing a serious impediment to intercourse and traffic," all of which arose because of the politically created constraints on the trusts, of course. However, the Commission concluded that the trusts should be abolished, and that the roads should be turned over to a government authority, as in South Wales (Quoted in Webb and Webb 1913, p. 221). Again, Parliament did not respond to these recommendations by establishing a general policy. Instead, a gradual abolition of more and more tolls began. Most Trusts were renewed each year, although from 1864 onward, 20–30 trusts were dissolved annually, with the roads turned over to a local parish or a highway district. Dissolution accelerated over time, however, as more and more trusts became insolvent, and in 1871 all tolls were ended in the London area. The number of trusts was down to 854 in 1871, 588 in 1875, 184 in 1881, 71 in 1883, 15 in 1887, and 2 in 1890. The last trust ended operations in November of 1895.

The increasing rate of dissolution of the trusts rapidly placed thousands of miles of roads back into the care of the parishes, leading to increased local resistance. In order to mitigate some of the local opposition the central government began giving Grants in Aid in 1876 to help pay for maintenance. Then in 1878 the Highway and Locomotive Act ordered the counties to contribute half of the annual cost of maintaining the former turnpike roads. The Local Government Act of 1888 granted more aid from the central treasury to counties for road maintenance but required the counties to take over full maintenance obligations from the parishes for all of the "main" roads. County governments were becoming the local road authority, and the parishes were finally formally dissolved of their road maintenance liability in 1895. "Thus at length the British road system was placed under control of elected public authorities each representing larger areas" (Gregory 1931, p. 196). Funding shifted from tolls to county (or in places, borough or other local government) taxes along with subsidies from the central government, and government expenditures on roads increased rapidly.[46] The average government expenditure per mile of county roads rose from 43 pounds in 1890 to 69 pounds in 1902, for instance while expenditures on urban roads increased from 49 to 207 pounds over the same period (Gregory 1931, p. 196).

[46]The Local Government Act of 1858 authorized any parish to become an "Urban Sanitary District," and these districts could not be included in any larger Highway District (Gregory 1931, p. 195). Parishes whose local officials wanted to maintain control of their roads used this process to do so, and as a consequence, a number of small districts avoided political consolidation.

11.5.3.3 The Commons Problem

Free access to roads led to new types of road users. By the end of the nineteenth century many highways had more bicycle traffic than horse drawn traffic, for example, and virtually every country home had a bicycle. As Webb and Webb (1913, pp. 204–211) report, "What the bicyclist did for the roads, between 1888 and 1900, was to ... accustom us all to the idea of our highways being used by other than local residents. It was the bicyclist who brought the road ... into popular use for pleasure riding." Pleasure riding is pretty attractive when one does not have to pay for it, of course. In addition, while excessive tolls or outright prohibitions had kept the steam carriage off the roads, the end of tolls and the development of light internal combustion engines in 1885 led to the introduction of the first petroleum driven motor cars in England in about 1894. The commons problem quickly became evident as bicycles and automobiles alarmed horses and pedestrians, and raised dust due to their speed (roads were largely still surfaced with gravel at this time): "the turning loose on our roads of tens of thousands of heavy vehicles, often traveling at speed of an express train, amounted to a real aggression on the safety and comfort of all the other users of the roads" (Webb and Webb 1913, p. 214). Accidents increased dramatically, generally at the cost of those who did not enjoy the benefits of the new transportation methods (pedestrians, users of horses), and road damage significantly increased maintenance costs.[47] Those who wanted to use the roads for traditional horse-drawn traffic protested loudly, but with no turnpike trusts or tolls to manipulate, raising political barriers to road use by bicycles and automobiles proved to be difficult. Efforts to prevent their use were successfully resisted by the growing political influence of motor vehicle owners, although several actions were taken to limit motor traffic access. The Motor Car Act of 1903 required new vehicles to be registered and licenses and to carry "conspicuous identification numbers back and front," for instance, and drivers were also required to be licensed even though other types of road use did not face such costs. The costs of automobiles were also increased due to requirements of lights and alarms that could be sounded, and speed limits were established. These limitations were clearly not sufficient, of course, and the use of motor cars expanded, with their accompanying negative externalities.

The cost of road maintenance and improvement due to these "new users" rose rapidly during the first decade of the twentieth century and the central government was continually pressed to provide relief to local taxpayers. It was felt that the users of motor vehicles should share in the costs of maintenance but the local road authorities (counties, boroughs, etc.) were not able to collect from most of them who traveled through their areas because they could not charge tolls. Finally a national tax on petrol was established in 1909 along with increased licensing fees paid to the central government. These new road revenues were administered by a new Road Board with the power to subsidize local road authorities, not for general maintenance, but for

[47] Pounding of the road surfaces when dry created an unanticipated problem of "waviness" (Gregory 1931, p. 257), the use of "armoured tires" with iron studs on automobiles to prevent side slipping further damaged road services, and ruts were created during rainy weather.

specific types of road improvements and new roads. The new road users demanded a very different type of road than horse-drawn transport required, after all, and as the political power of the owners of motor vehicles increased, more pressure was brought to bear on Parliament to provide roads suitable to such traffic.

The Road Transport Board was created in 1918 to coordinate all work on roads during the First World War, and it continued to function after the war as a Department of the Ministry of Transportation. Its role was one of centralized supervision of road development, but it also was given the power to allocate grants from the central government's Road Fund (Gregory 1931, p. 248). Roads were divided into three classifications based on width (60 feet for first class and 50 for second), and up to 50 % of the cost of maintenance and improvement of first class roads was to come from the national government. Second class roads were to get up to 25 % and third class roads were to receive no aid. At the time, 26,000 miles of Britain's 179,000 total road miles were classified as first class, and 15,800 miles fell into the second class. As Gregory (1931, p. 257) notes, however, "It is a natural desire of local road authorities to transfer roads from the grant-less third class," so local authorities began finding ways to claim that more of their roads were wide enough to warrant subsidization. This did not mean that roads actually had to be widened, however: "Each class must have a minimum carriage way of 20 feet, with one footpath. Therefore a third class road which has a footpath and a grass verge on each side, can be promoted to the second class and made eligible for a grant, by the simple process of absorbing a footpath and part of the verge into the carriage way, to the detriment of the safety of the pedestrian and the appearance of the road" (Gregory 1931, pp. 248–249). Grants could then be obtained to surface formerly graveled third class roads with tar. Thus, more and more of the funding of roads was being shifted to the central government.

11.6 Conclusions

The contention that roads are not public goods, but that they can be club goods, private goods, or common pools, depending on the institutional arrangement that exists, is supported above by an examination of the evolution of road provision in Great Britain. Indeed, recognition of the fact that public roads are really common pools and that other arrangements have existed in the past under different institutional environments reinforces the point by Minasian (1964, p. 79) that the outcomes we observe are a result of the property-rights/institutional arrangements that exist and "alternative exclusion and incentives systems" produce very different results. This opens up a much larger set of policy options beyond government taxation and financing of roads. After all, the analysis presented here is not just supported by the historical evolution of road provision in Great Britain, as similar systems are prevalent in the history of other nations as well.[48] Furthermore, the fact is that both private and club provision of roads

[48]For example, the British experience with toll roads has an American counterpart. The first toll road company in the U.S. was chartered by Pennsylvania in 1792, in order to provide a highway

is evident today. Roth (1996, pp. 180–197) documents several private road projects in developing countries, for instance, where governments have found that the private sector can respond to demands for highways more effectively and more quickly than the state can.[49] Roth also cites examples in developed countries, including two privately built highways in Great Britain (the Dartford River Crossing Ltd.'s toll bridge crossing the Thames, and Midland Expressway Ltd.'s M6-Toll Road, a 27 mile expressway to relieve congestion in one of England's busiest urban areas). Privately-financed highway projects are also appearing in the U.S. (e.g., the Dulles "Greenway" project in Northern Virginia and the State Road 91 project in Orange County, California).[50]

Large proprietary business operations also provide streets and even highways. Consider Disney World, for instance (Foldvary 1994). Smaller scale business operations also provide streets (or substitutes for streets: mall parking lots and corridors are substitutes for public streets that reduce the costs of shopping relative to business districts crisscrossed by public streets, for instance). In addition, many residential developments actually include streets that are private or club goods (Foldvary 1994). The actual number of these arrangements is not known, but estimates of the number of United States residents living in gated communities ranged up to 8,000,000 in 1995 (Benson 1998, p. 93) and this is only a small portion of the communities that do not rely on the government for streets. Many streets are built by developers who factor consideration of the cost of streets in the prices of lots or homes. As residents move in they often form home ownership associations, and at some point these associations generally join with and ultimately replace developers in coordinat-

(Footnote 48 continued)
between Philadelphia and Lancaster. 1,562 turnpike companies established over 10,000 miles of roads in the Eastern United States between 1792 and 1845 (Klein and Fielding 1992). As Gunderson (1989, p. 192) notes, "Relative to the economy at the time, this effort exceeded the post -World War II interstate highway system that present day Americans assume had to be primarily planned and financed by the federal government." Similarly, Klein and Yin (1996) point out that about 150 private toll roads were opened in California between 1850 and 1902. Klein (1990) explains, however, that numerous government mandated toll exemptions for powerful interest groups tended to undermine the incentives to build private toll roads (government regulations often explicitly prevented profit taking, just as in Great Britain).

[49]Many developing countries are franchising roads to private firms which construct the roads and them operate them, charging tolls to earn the costs of construction and operation, **and** to cover franchising fees paid to the government (Pereyra 2002). Indeed, providing such roads are so attractive, in part because of their impact on real estate values, that it is becoming increasingly common for governments to auction franchises (Engel et al. 2002).

[50]The Intermodal Surface Transportation Efficiency Act of 1991 actually attempted to stimulate privately provided toll roads, bridges and tunnels in the U.S. (as long as they are not part of the interstate highway system) by making them eligible for a 50 % grant from the Highway Trust Fund, and in an effort to take advantage of these available funds a number of states have passed their own legislation to allow private provision of roads. However, some of the largest and most important private highway projects have refused federal funds in order to avoid the added complications that accepting such funding entails.

ing street maintenance. Thus, these streets become local club goods. There also are cases where streets were actually owned by a local government that deeded them to local homeowners associations (Newman 1980). This mix of private and club goods clearly demonstrates that roads are not necessarily Samuelsonian public goods, and an examination of the traffic situation on many modern public roads demonstrates that even publicly provide roads are not public goods in a Samuelson sense. This is made quite apparent by a simple examination of the level of traffic congestion in most urban areas of the United States, the United Kingdom, and numerous other parts of the world, but note that negative externalities arise whenever multiple drivers want to travel at different speeds at the same time on the same road.

The focus in this presentation has been on roads, but the same analysis appears to apply to numerous other so-called public goods. Consider education for instance, with its mix of private schools and home schooling, community based (club) schools such as those provided by various religious organizations, and crowded inefficiently run and expensive public schools. Or consider policing. In the United States today there are roughly three times as many private security personnel as there are public police, and there are also large numbers of community level crime control activities, such as crime watch and neighborhood patrols (Benson 1998, 2010, 2014; Benson and Mast 2001). Furthermore, public police find their files crowded with cases that cannot even be investigated due to the excess demand for their services, and victimization surveys suggest that most crimes against persons and property are not even reported as victims are, essentially, crowded out of the system. Indeed, the evolution of the institutions of policing in Great Britain is very similar to the evolution of road provision described above (Benson 1994). Similarly, public court backlogs and delay force many disputes to be dropped while many more disputants choose to go to private or club alternatives: direct negotiation, or mediator/arbitrators service, some of which are offered for a price while others are provided by communities such as trade associations (see Bernstein 1992 and Benson 1995, 1999c for instance). In fact, a claim that roads, education, policing, or courts (or most (all?) other publicly financed goods or services) are public goods, even up to the point where crowding sets in, is of little real relevance for public policy because, as Minasian (1964, pp. 79–80) explains, "the theory [of public goods] generates economic analysis which is not based on the opportunity cost notion." Rationing of scarce goods cannot be avoided by declaring that no one can be excluded; this simply means that first-come-first-serve and its accompanying congestion costs determine who gets what, or that regulations limiting use must be passed. Individuals and close knit-communities look for alternative means to produce the desired services, however, and a mix of club and private goods inevitably exist. And importantly, these alternative sources of the desired goods would provide a lot more of them if property rights were not constrained by the state in order to meet the demands of special interest groups.

References

Albert W (1972) The Turnpike road system in England: 1663–1840. Cambridge University Press, Cambridge
Anderson TL, Hill PJ (1990) The race for property rights. J Law Econ 33(1):177–197
Bailey MJ (1992) Approximate optimality of aboriginal property rights. J Law Econ 35(1):183–198
Benson BL (1984) Rent seeking from a property rights perspective. South Econ J 51(2):388–400
Benson BL (1988) Legal evolution in primitive societies. J Inst Theor Econ 144(5):772–788
Benson BL (1989) The spontaneous evolution of commercial law. South Econ J 55(3):644–661
Benson BL (1994) Are public goods really common pools? Considerations of the evolution of policing and highways in England. Econ Inq 32(2):249–271
Benson BL (1995) An exploration of the impact of modern arbitration statutes on the development of arbitration in the United States. J Law Econ Organ 11(2):479–501
Benson BL (1998) To serve and protect: privatization and community in criminal justice. New York University Press, New York
Benson BL (1999a) An economic theory of the evolution of governance and the emergence of the state. Rev Austrian Econ 12(2):131–160
Benson BL (1999b) Polycentric law versus monopolized law: implications from international trade for the potential success of emerging markets. J Priv Enterp 15(1):36–66
Benson BL (1999c) To arbitrate or to litigate: that is the question. Eur J Law Econ 8(2):91–151
Benson BL (2005) Regulation, more regulation, partial deregulation, and reregulation: the dynamics of a rent-seeking society. Adv Austrian Econ 8(1):107–146
Benson BL (2006) The rise and fall of private roads in England. In: Roth G (ed) Street smart: competition, entrepreneurship and the future of roads. Transaction Publishers, New Brunswick, pp 245–276
Benson BL (2010) The allocation of police. In: Benson BL and Zimmerman PR (eds) Handbook on the economics of crime. Edward Elgar, Northampton, pp 184–217
Benson BL (2011) The enterprise of law: justice without the state, 2nd edn. The Independent Institute, Oakland
Benson BL (2014) Let's focus on victim justice, not criminal justice. Indep Rev 19(2):209–238
Benson BL, Mast BD (2001) Privately produced general deterrence. J Law Econ 44(2):725–746
Beresford MW, Joseph JS (1979) Medieval England: an aerial survey. Cambridge University Press, Cambridge
Bergstrom T, Blume L, Varian H (1986) On the private provision of public goods. J Pub Econ 29(1):25–49
Bernheim BD (1986) On the voluntary and involuntary provision of public goods. Am Econ Rev 76(4):789–793
Bernstein L (1992) Opting out of the legal system: extralegal contractual relations in the diamond industry. J Legal Stud 21(1):115–157
Bewes WA (1923) The romance of the law merchant: being and introduction to the study of international and commercial law with some account of the commerce and fairs of the middle ages. Sweet & Maxwell, London
Blair PH (1956) An introduction to Anglo-Saxon England. Cambridge University Press, Cambridge
Bodey H (1971) Roads. BT Batsford Limited, London
Buchanan JM (1965) An economic theory of clubs. Economica 32(125):1–14
Buchanan JM (1994) Choosing what to choose. J Inst Theor Econ 150(1):123–135
Carneiro RL (1970) A theory of the origin of the state. Science 169(3947):733–738
Coase RH (1960) The problem of social cost. J Law Econ 14(1):1–44
Coase RH (1974) The lighthouse in economics. J Law Econ 17(2):357–376
Dahlman CJ, et al (1980) The open field system and beyond: a property rights analysis of an economic institution. Cambridge Books
Dalgleish A (1980) Telford and steam carriages. In: Penfold A (ed) Thomas Telford: engineer. Thomas Telford Ltd, London

Dance SC (1831) Postscript to steam carriages on common roads. A pamphlet possibly prepared as evidence to the parliamentary select committee on tolls
Darby HC (1973) A new historical geography of England. Cambridge University Press, Cambridge
Demsetz H (1967) Toward a theory of property rights. Am Econ J 57(2):347–359
Ellickson RC (1993) Property in land. Yale Law J 102(6):1315–1400
Ellickson RC (1991) Order without law: How neighbors settle disputes. Harvard University Press, Cambridge
Engel E, Fischer R, Galetovic A (2002) Highway franchising and real estate values. NBER working paper no 8803 57(3):432–448
Fletcher W (1891) History and development of steam locomotion on common roads. E. & FN Spon, London
Foldvary F (1994) Public goods and private communities: the market provision of social services. Edward Elgar Publishing, Chelthenham
Foley DK (1970) Lindahl's solution and the core of an economy with public goods. Econometrica 38(1):66–72
Fuller L (1981) The principles of social order. Duke University Press, Durham
Gordon A (1833) Institute of locomotion for steam transport and agriculture. Report regarding this organization
Gregory JW (1931) The story of the road, from the beginning down to A.D. A. Maclehose, New York
Gunderson G (1989) Privatization and the 19th-century turnpike. Cato J 9(1):191–200
Gurney G (1831) Steam carriages on common roads. A pamphlet possibly prepared as evidence for presentation to a select committee of the house of commons
Hardin G (1968) The tragedy of the commons. Science 162(3859):1243–1248
Hardin R (1982) Collective action. Johns Hopkins University Press, Baltimore
Hayek F (1973) Law, legislation, and liberty, vol 1. University of Chicago Press, Chicago
Hindle P (1982) Roads and tracks. In: Cantor L (ed) The English medieval landscape. Croom Helm, London, pp 191–210
Hindley G (1971) A history of roads. Peter Davies, London
Hoebel EA (1954) The law of primitive man. Harvard University Press, Cambridge
Jackman WT (1966) The development of transport in modern England. Augustus M. Kelley Publishers, New York
Klein DB (1990) The voluntary provision of public goods? The turnpike companies of early America. Econ Inq 28(4):788–812
Klein DB, Fielding GJ (1992) Private toll roads: learning from the 19th century. Transp Q 46(3):321–341
Klein DB, Yin C (1996) Use, esteem, and profit in voluntary provision: toll roads in California, 1850–1902. Econ Inq 34(4):678–692
LaMar VA (1960) Travel and roads in England. The Folger Shakespeare Library, Washington
Leoni B (1961) Freedom and the law. Nash Publishing, Los Angeles
Levi M (1988) Of rule and revenue. University of California Press, Berkeley
Llewellyn KN, Hoebel EA (1961) Cheyenne way. University of Oklahoma Press, Norman
Lyon BD (1980) A constitutional and legal history of medieval England, 2nd edn. WW Norton and Company, New York
McNutt P (2000) Public goods and club goods. In: De Geest G, Bouckaert B (eds) Encyclopedia of law and economics. Edward Elgar, Chelthenham, pp 927–951
Milgrom PR, North DC et al (1990) The role of institutions in the revival of trade: the law merchant, private judges, and the champagne fairs. Econ Polit 2(1):1–23
Milleron JC (1972) Theory of value with public goods: a survey article. J Econ Theory 5(3):419–477
Minasian JR (1964) Television pricing and the theory of public goods. J Law Econ 7(1):71–80
Mises LV (1949) Human action: a treatise on economics, 3rd edn. Contemporary Books, Chicago
Mises LV (1957) Theory and history: an interpretation of social and economic evolution. Yale University Press, New Haven
Moyes A (1978) Transport 1730–1900. In: Dodgshon R, Butlin R (eds) An historical geography of England and Wales. Academic Press, New York, pp 401–429

Nee V (1998) Norms and networks in economic and organizational performance. Am Econ Rev 88(2):85–89
Neely R (1982) Why courts don't work. McGraw-Hill, New York
Newman O (1980) Community of interest. Anchor Press, Garden City
North DC (1981) Structure and change in economic history. Norton, New York
North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
O'Driscoll GP, Rizzo MJ, Garrison RW (1985) The economics of time and ignorance. Basil Blackwell, Oxford
Oppenheimer F (1914) The state: its history and development viewed sociologically. Bobbs-Merrill Company, Indianapolis
Ostrom E (1990) Governing the commons: the evolution of institutions for collective actions. Cambridge University Press, New York
Ostrom E (2005) Understanding institutional diversity. Princeton University Press, Princeton
Ostrom V (2007) The political theory of a compound republic: designing the American experiment. Lexington Books, Lanham
Parkes J (1925) Travel in England in the seventeenth century. Oxford University Press, London
Pawson E (1977) Transport and economy: the turnpike roads of eighteenth century Britain. Academic Press, London
Pereyra A (2002) Auction theory and road franchising. Universidad de Uruguay-Departamento de Economia-Documento de trabajo en edicion
Pollock F, Maitland FW (1959) The history of English law, vol 1. Lawyers' Literary Club, Washington
Pospisil LJ (1971) Anthropology of law: a comparative theory. Harper and Row, New York
Rasmussen DW, Benson BL (1994) The economic anatomy of a drug war: criminal justice in the commons. Rowman & Littlefield, Lanham
Ridley M (1996) The origins of virtue: human instincts and the evolution of cooperation. Viking Penguin, New York
Roth G (1996) Roads in a market economy. Avebury Techinal, Aldershot
Samuelson P, Nordhaus W (1985) Economics. McGraw-Hill, New York
Samuelson PA (1954) The pure theory of public expenditure. Rev Econ Stat 36(4):387–389
Samuelson PA (1955) Diagrammatic exposition of a theory of public expenditure. Rev Econ Stat 37(4):350–356
Select Committee on Steam Carriages (1831) Report from select committee on steam carriages; with minutes of evidence and appendix
Shitovitz B, Spiegel M (1998) Cournot-Nash and Lindahl equilibria in pure public good economies. J Econ Theory 83:1–18
Shitovitz B, Spiegel M (2001) Stable provision vs. Cournot–Nash equilibria in pure public good economies. J Public Econ Theory 3(2):219–224
Shitovitz B, Spiegel M (2002) Trading equilibrium in the public good economy. University of Haifa working paper presented at the association of private enterprise education meetings
Shoup CS (1964) Standards for distributing a free governmental service: crime prevention. Public Financ 19:383–392
StanleyMotorCarriage.com (2016) The stanley steamer, why the fascination?
Stenton FM (1936) The road system of medieval England. Econ Hist Rev 7:1–21
Stephen JF (1883) A history of the criminal law of England. Burt Franklin, New York
Taylor M (1982) Community, anarchy and liberty. Cambridge University Press, Cambrige
Vanberg VJ, Congleton RD (1992) Rationality, morality, and exit. Am Polit Sci Rev 86(2):418–431
Webb S, Webb B (1913) The story of the king's highway: English local government, vol 5. Archon Books, Hamden
Willan TS (1976) The inland trade: studies in English internal trade in the sixteenth and seventeenth centuries. Manchester University Press, Manchester