HARVARD UNIVERSITY
THE GRADUATE SCHOOL OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division

Department **Government**

Committee

have examined a thesis entitled

"Casual Inference in Political Science and Law"

presented by **Daniel En-Wenn Ho**

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance

Signature.. *Gary King (Chair)*

Signature.. *James E. Alt*

Signature.. *Ian Ayres*
Yale University

Signature .... Kevin Quinn

Date: **March 25, 2004**

Causal Inference in Political Science and Law

A thesis presented by

Daniel E. Ho

to

The Department of Government in partial fulfillment of
the requirements for the degree of Doctor of Philosophy
in the subject of Political Science

Harvard University
Cambridge, Massachusetts
March, 2004

UMI Number: 3131866

Causal Inference in Political Science and Law

By Daniel E. Ho

Committee: Gary King (Chair), James E. Alt, Ian Ayres, Kevin M. Quinn.

Abstract

This work presents four independent essays applying statistical methods of causal inference to examine puzzles of interest in law and political science.

Part One, consisting of the first two essays, examines the robustness of causal claims in observational studies. The first essay reconsiders a claim in positive political economy on the causal effect of electoral systems on price levels. It illustrates that traditional parametric methods may identify causal effects solely on functional form, and shows how matching methods may provide more robust estimates.

The second essay assesses the impact of war on U.S. Supreme Court decisionmaking. It finds that war is associated with a substantial decrease in the probability of a liberal decision in civil rights and liberal cases. Contrary to received wisdom, the war affects cases not directly related to the war, but little evidence suggests that war impacts war-related cases.

Part Two, consisting of the last two essays, takes an experimental approach, capitalizing on randomization to draw credible causal inferences for a natural experiment and a survey with a randomized instrument. The third essay examines the natural experiment of the California alphabet lottery, which has randomized the ballot order of candidates for statewide offices since 1975. It finds that ballot order substantially affects all candidates in

primary elections, and that ballot order may have changed the winner in over ten percent of all primary races from 1978 to 2002.

The final essay illustrates a Bayesian generalization of instrumental variable estimation to assess the effect of racial perceptions on political knowledge surveys. It finds that contrary to previous analyses, the race of an interviewer in a telephone survey affects political knowledge answers, irrespective of the stated perception of race, and that this effect is more substantial for white respondents than for black respondents.

To My Parents

Contents

Acknowledgements **xi**

1 Introduction **1**

I Observational Approaches 7

2 Electoral Systems and Consumer Power **8**
2.1 Introduction . . . 8
2.1.1 The Causal Logic . . . 9
2.1.2 Assessing the Price-Level Effect . . . 11
2.2 Extrapolation in the RK Data . . . 13
2.3 Assessing the Causal Effect of Electoral Systems . . . 21
2.3.1 The Causal Effect . . . 21
2.3.2 Majoritarian Versus PR Countries . . . 22
2.4 Matching on the Propensity Score . . . 22
2.4.1 Findings . . . 26
2.5 A Test with New Data on all Democracies . . . 29
2.5.1 Matching . . . 30
2.6 Conclusion . . . 33
2.7 Appendix . . . 35

3 The Crisis Thesis **39**
3.1 Introduction . . . 39
3.2 Data . . . 49
3.2.1 The Cases . . . 49
3.2.2 The Crises . . . 52
3.2.3 Confounding Factors . . . 57
3.3 Empirical Results on the Causal Effect of War . . . 71
3.4 Conclusion . . . 79

II Experimental Approaches 80

4 California Alphabet Lottery 81
4.1 Introduction . . . 81
4.2 Elections and Ballot Order . . . 84
4.3 The California Alphabet Lottery . . . 89
4.3.1 Lottery Procedure . . . 89
4.3.2 Are Alphabets Really Random? . . . 93
4.4 Causal Effects of Ballot Order . . . 95
4.4.1 Causal Inference and Treatment Assignment . . . 97
4.4.2 Identifying Causal Effects of Ballot Order . . . 99
4.4.3 Estimated Causal Effects of Ballot Order . . . 105
4.4.4 Margin of Victory and Ballot Order Effect . . . 114
4.5 Alternative Approaches . . . 116
4.5.1 A Multinomial Model . . . 116
4.5.2 Distribution-Free Randomization Inference . . . 118
4.6 Policy Implications for Ballot Reform . . . 122
4.6.1 Cost and Benefits of Randomization . . . 122
4.6.2 The Cost-Effectiveness of Randomization . . . 125
4.6.3 Existing Policies and Possible Reforms . . . 128
4.7 Concluding Remarks . . . 132
4.8 Appendix: Assessing Balance of Covariates . . . 133
4.9 Appendix: Conditional Effects . . . 136

5 Bayesian Instrumental Variables 138
5.1 Introduction . . . 138
5.2 The Political Knowledge Data . . . 140
5.3 Framework of Causal Inference . . . 142
5.4 A Mixture Model . . . 147
5.5 Results . . . 148
5.5.1 ITT Analysis . . . 149
5.5.2 IV Analysis . . . 150
5.6 Conclusion . . . 152

Bibliography 162

List of Figures

2.1 The convex hull of GDP per capita data 16
2.2 Density estimates of GDP per capita and arable land for SMD and PR countries 17
2.3 OLS fitted values for (a) model regressing prices on arable land and dummy variable of SMD and (b) model regressing prices on arable land separately for SMD and PR countries 18
2.4 Density estimates (smooth versions of histograms) of estimated propensity scores for SMD and PR countries in original specification of Model 1.4 (left panel) and matching model (right panel). 27
2.5 Matched and unmatched density estimates of propensity scores 30
2.6 Comparison of average treatment effect of majoritarian system on price levels from RK and matching model 34
2.7 Seats-Vote elasticity as a function of vote share in three electoral systems . 36

3.1 The proportion of U.S. Supreme Court decisions supporting rights, liberties, or justice claims, 1953-2001 terms 51
3.2 Facts relevant to the Supreme Court's adjudication of Fourth Amendment Cases 60
3.3 The political ideology of the median justice on the U.S. Supreme Court, 1941-2003 68
3.4 Kernel Density Estimate of Ideology Conditional on War, 1941-2002[1] . . . 69
3.5 The Preponderance of Left-of-Center Courts During Times of War 70
3.6 Proportion of U.S. Supreme Court decisions in the areas of rights, liberties, or justice that are salient, 1941-2001 terms 71
3.7 Histogram of the effect of wars civil rights and liberties 75
3.8 The effect of wars on the outcomes of Supreme Court Cases in the four areas of rights, liberties, and justice 76
3.9 The effect of wars on the outcomes of supreme court cases in the areas of rights, liberties and justice: a comparison of matched pairs[2] 78

4.1 1992 California Assembly Districts 92
4.2 Simple Random and Systematic Random Assignment under Populations with Monotonic and Periodic Trends 102

4.3 Candidate-Specific Average Relative Gain due to Being Listed in First Position on Ballots for 1998 and 2000 Elections 107
4.4 Comparison of Estimated Average Ballot Order Effect for Second-highest Vote-getter and Margins of Victory from 1978 to 2002 115
4.5 Distribution of T-statistics from Covariate Means Tests 135

5.1 Simulation scatterplot of the joint posterior distributions 155

List of Tables

2.1 Summary Rogowski-Kayser findings 11
2.2 Number and percentage of counterfactuals falling within convex hull of data 20
2.3 Means tests of single-member district countries (SMD) and proportional representation countries (PR) on selected covariates 23
2.4 Means tests of matched countries 28
2.5 Estimated average treatment effects with one-to-one nearest neighbor matching 28
2.6 Means tests of matched countries with new data 31
2.7 Estimated average treatment effects (ATE) of majoritarian electoral systems on prices with one-to-one nearest-neighbor matching 33
2.8 Sensitivity of OLS estimates to specification 37
2.9 Variable description and data sources of new dataset 38

3.1 Crises and U.S. Supreme Court Cases Involving Rights, Liberties, and Justice, 1941-2001 Terms[3] 57
3.2 Time series assessments of the effect of international crises on the severity of case characteristics 62
3.3 Summary statistics of overall sample. 72
3.4 Summary statistics of matched cases 73
3.5 Estimated causal effects of war 74
3.6 Direct effects of other binary case variables for warcases and non-warcases, matching exactly on all other covariates. Ideology here is dichotomized by whether it takes on the highest observed value of the Segal-Cover score or not. N_1 and N_0 represent the number of observations matched. 79

4.1 Randomized Alphabets Used for the California Statewide Elections Since 1982. 94
4.2 Number of Candidates Running in All Races Examined 96
4.3 Party-Specific Average Causal Effects of Being Listed in First Position on Ballots Using All Races from 1978 to 2002 108
4.4 Absolute Gain due to Being Listed in First Position on Ballots using All Races from 1978 to 2002 112

4.5 Average Relative Gain due to Being Listed in First Position on Ballots using All Races from 1978 to 2002 113
4.6 Party-Specific Average Causal Effects of Being Listed in First Position on Ballots in 13 General Elections using the Multinomial Logit Model 118
4.7 Fisher's Exact Test by Party and Offices 121
4.8 Cost Effectiveness of Selected Ballot Reforms 126
4.9 Types of Ballot Order Rules for State Gubernatorial Races in General Elections as of 2003 129
4.10 Means Tests of Selected Covariates 134
4.11 Absolute Ballot Order Effects Conditional Whether Incumbents are Running 136
4.12 Absolute Ballot Order Effects for On or Off-Year Elections for Senate Elections 136
4.13 Relative Ballot Order Effects Conditional on the Number of Candidates . . 137

5.1 Political knowledge questions 142
5.2 Summary statistics of political knowledge survey 143
5.3 Intention to treat results 149
5.4 Summary statistics of posterior distributions 154

Acknowledgements

First and foremost, I'd like to thank my the members of dissertation committee, who patiently, kindly, and professionally guided me in my research. Gary provided uncountable hours and conversations,[4] continually offering detailed comments and professional advice, to which I am gratefully indebted. It was his confidence in me that transformed my graduate experience. Jim provided an academic home as director of the Center for Basic Research in the Social Sciences ("CBRSS"), reading drafts of my work with great detail and thoughtful encouragement. Ian has, since the first days of contracts class, inspired me to diversify and broaden my work to speak to a legal audience. And Kevin has been always been receptive to my ideas with an open door, patient hand, and sharp mind. I consider myself blessed beyond riches to have be able to work under the guidance of such a splendid committee.

Second, I'd like to thank other collaborators, with whom I have had the utmost pleasure to do research. Kosuke Imai and Liz Stuart, in particular, have been amazing colleagues, from whom and with whom I have learned a tremendous amount. Lee Epstein and Jeff Segal have provided the perspective of judicial politics, with an openness for new findings true to scientific principle.

[4]Taking a random sample of 30 emails from roughly 1,000 over the last year, not one was answered with more than a 24-hour delay.

Third, I am thankful for generous financial support from CBRSS, the Project on Justice, Welfare and Economics at the Weatherhead Center for International Affairs, and the Center for American Political Studies.

Fourth, I'd like to thank faculty, colleagues, and friends at Harvard University, who have entertained my ideas with steadfast support, including Scott Ashworth, Elizabeth Cobb, Jim Greiner, Matt Hindman, Dale Jorgenson, Michael Kang, Sophie Langeskiöld, Hanna Lee, Wendy Pearlman, Alison Post, Jas Sekhon, and Elina Treyger. CBRSS, with its amazing staff and computing support, facilitated research substantially and brightened cold winter days.

Lastly, I'd like to thank my family for their unwavering support. My father has continually pushed me on the *science* of political science; my mother has always listened with caring ears; and my sister, who was concurrently in medical school in Boston, provided family when our parents were an ocean away in Europe.

Chapter 1

Introduction

Causal inference forms a central part of social scientific research (see, e.g., Pearl, 2000; King, Keohane and Verba, 1994; Epstein, 1987). Political scientists and legal scholars have become accustomed to examining empirical implications of their theories. Yet crucial identification assumptions of causal effects often remain *implicit* in actual applications. As a result, our knowledge of causal quantities of interest in the social world is often weak at best: incredible assumptions lead to fragile estimates of causal effects that are highly sensitive to model specification. This work seeks to take a step towards *credible inference* in political science and law, by clarifying and relaxing common assumptions imposed by studies purporting to assess causal effects.

This work is motivated by important developments in statistics, primarily in matching methods for causal inference and extensions from experimental designs to observational studies (e.g., Frangakis and Rubin, 2002; Holland, 1986; Imbens and Rubin, 1997a; Rosenbaum and Rubin, 1983b). These methods can greatly reduce the fragility of causal inferences in political science and law, by permitting researchers to relax distributional or functional form assumptions commonly imposed in studies. In conjunction with this research, we provide an open-source program implementing a wide range of sophisticated matching methods.[1] I demonstrate the broader methodological points with four examples using observational and experimental data from several fields in political science and law.

Part I, consisting of Chapters 2 and 3, examines causal claims in *observational data*, where the treatment of interest is not physically randomized (Rosenbaum, 2nd edition, 2002). As a result, identification stems largely from an assumption of conditional ignorability or exogeneity of treatment (i.e., that treatment is *random* after holding all other covariates constant). Yet compared to parametric strategies of identification in observational data, matching methods are less sensitive to functional form assumptions. This sensitivity of traditional approaches is particularly acute when treated and control units are quite distinct in pre-treatment covariates. When units self-select into treatment, for example, as is often the case in political science and law (e.g., the adoption of electoral rules), differences along observable covariates may be substantial. Hence, estimates may extrapolate wildly from the bounds of the data (King and Zeng, 2002).

To illustrate this, Chapter 2 examines the case of a widely-celebrated claim in political economy due to Rogowski and Kayser (2002, p. 526) that "systems of proportional representation . . . systematically advantage producers and disadvantage consumers." The Chapter finds that selection on observables looms large: proportional representation countries

[1] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart, MATCHIT: Matching Software for Causal Inference, available at `http://gking.harvard.edu/matchit/`.

are substantially different from majoritarian district countries along a host of background characteristics. As a result, the original study, relying on linearity, extrapolates severely from the data to estimate the causal effect of a majoritarian electoral system. Accounting for these differences by matching on the propensity score forces one to discard observations that extrapolate from the data. This in turn yields estimates of such high variance that we cannot find evidence for the price-level effect. Matching, similar to robust estimation techniques (Wilcox, 1997), discards or downweights high leverage outlier points, reducing model-dependency of causal estimates. The Chapter then reassesses the hypothesis by gathering a larger dataset. By including non-OECD democracies, this dataset increases potential observations. Yet even with this data the price-level effect remains undetectable. The conclusion of a price-level effect thereby rests on unjustified modeling assumptions.

In contrast, Chapter 3 illustrates a case in which matching methods detect a robust effect, which was obscured by misspecified parametric models at an earlier stage of the research. Since confounding covariates are largely categorical, we can match exactly on all covariates. The virtue of exact matching is that it yields estimates that are entirely invariant to parametric adjustments. Substantively, the Chapter provides the first systematic empirical evidence of the so-called "crisis thesis," which holds that U.S. courts uniformly restrict civil rights and liberties during times of war.[2] Examining all civil rights and liberties cases decided by the U.S. Supreme Court from 1941–2001, the Chapter finds that war substantially decreases the probability of a liberal decision by the court. Yet contrary to

[2]The chapter is collaborative work with Lee Epstein, Gary King, and Jeffrey A. Segal, and is a substantially abbreviated version of an article forthcoming in the *New York University Law Review*.

common wisdom, war exhibits no effect on cases that are a direct result of the war, such as cases involving war protests, military takings, or enemy nationals. Instead, the effect of war appears to be largest on cases not directly related to the war. These results are consistent with the notion that for war-related cases, where the threat of encroachment on the judiciary looms largest, the court rallies 'round the court. In unrelated cases, however, the balance between liberty and security swings towards security, as maintained in the literature.

Given the fragility of causal estimates to the assumption that treatment is random after conditioning on observable covariates, Part II, consisting of Chapters 4 and 5, examines causal claims from an experimental approach. This approach capitalizes on physical randomization of treatment or an instrument to draw inferences of interest with greater credibility and minimal assumptions. Unlike matching on observed covariates, randomization ensures balance along both observed *and unobserved* covariates. While it has long been recognized that randomization forms "a reasoned basis of inference" (Fisher, 1935), experiments still remain a small share of studies published in both law and political science. For many political and legal quantities of interest, pure laboratory experiments are often seen to lack external validity, while field experiments may be infeasible, expensive, and/or unethical. Part II therefore develops and illustrates methods to overcome these weaknesses, while still capitalizing on randomization where it exists.

Chapter 4 assesses a natural experiment, which does not suffer the external validity and pragmatic concerns of field or laboratory experiments. It answers the longstanding question of whether the name order of candidates on ballots affects election outcomes.[3] The natural

[3] The chapter is joint work with Kosuke Imai.

experiment derives from the fact that since 1975, California law has mandated randomizing the ballot order with a lottery, where alphabet letters would be "shaken vigorously" and selected from a container. Previous studies, relying overwhelmingly on observational data and parametric models to identify effects, yield conflicting results about whether ballot order effects even exist. Developing methods to test randomization and adapting techniques tailored to what we term as "systematic random treatment assignment," the Chapter analyzes statewide elections from 1978 to 2002. This analysis demonstrates that ballot order might have changed the winner in twelve percent of all primary races, including major and minor party races. The Chapter also illustrates the robustness and agreement of parametric and non-parametric methods when treatment is randomized. That is, when treated and control units are similar *by design*, functional form assumptions can be as innocuous as usually assumed. The Chapter concludes by proposing that all electoral jurisdictions randomize ballot order to minimize ballot effects. In fact, the Chapter shows that this proposal may be substantially more cost-effective at reducing voting bias than currently proposed voting technology reforms.

Lastly, Chapter 5 shows how researchers can employ a randomized instrument to assess a theoretical question of interest, where the treatment of interest itself would have been impossible to randomize. The Chapter illustrates a Bayesian framework developed in the statistical literature of analyzing, making explicit, and relaxing crucial assumptions that are implicit in virtually all instrumental variables analyses. The application assesses the causal effect of racial perceptions on answers to a political knowledge telephone survey.

It employs the fact that the race of the interviewer was randomly assigned to respondents. Contrary to previous resesarch, which failed to recognize the usefulness of interviewer randomization as an instrument, the results suggest that race affects survey answers irrespective of the perception of race, and that effects are larger on white respondents than on black respondents. This strongly indicates that previous research estimates are biased due to unobservable differences in respondents.[4] This type of research design, lying "between" perfect randomization of treatment and purely obervational studies (Gelman et al., 2004, p. 229), hence permits researchers to gain more leverage over theoretical quantities of interest with greater credibility.

In sum, this work takes a step towards a more credible and unified approach to causal inference. It uses new methods to reduce the fragility of causal inferences in political science and law.

[4]Selection on unobservables is what Heckman et al. (1998) defines as true selection bias.

Part I

Observational Approaches

Chapter 2

Majoritarian Electoral Systems and Consumer Power: A Matching Rejoinder

2.1 Introduction

In a substantial contribution to positive political economy, Rogowski and Kayser (2002, p. 526) (RK) finds that "systems of proportional representation (PR) systematically advantage producers and disadvantage consumers." If true, this previously unnoticed welfare effect of electoral systems might not only help to explain such puzzles as the "exceptional...frequency of electoral reform" in the 1990s (Persson and Tabellini 2003, p. 77) but would also seem to defy the complex systematization of electoral rules by political scientists such as Cox (1990) by its reduction of election rules into two discrete and ideal types (majoritarian and PR). Clearly, the validity of the study bears significant policy and intellectual consequences, as evidenced by the wide reception of the RK study in the field.[1] This paper concerns itself with the *empirical* validity of the RK claim.

2.1.1 The Causal Logic

The intuition of the causal logic in the formal model developed by RK is as follows: First, employing a Stigler-Peltzman model of regulation, politicians maximize political support which is a function of money and votes. While producers can provide both money and votes, consumers can only vote to influence political actors. Politicians thereby trade off support from producers and consumers, equating the marginal rate of substitution between the two groups.

Second, the seats-votes elasticity, defined as "the percentage increase in seats to be

[1]Evidence of the impact of the article abounds. In the short time since the original article was published, Scartascini (2002) has directly extended the Rogowski-Kayser model to examine business regulation; Milner and Judkins (2002) has found supporting evidence in the effect of majoritarian electoral systems on policies that affect price levels, specifically examining protectionism; Rosenbluth and Schaap (2002) has found empirical evidence on the effect of electoral rules on banking regulation; and Fiona (2002) analyzes inter-industry stock market price variation. The article has been widely cited, e.g., Gourevitch, Carney and Hawes (2003), Gourevitch (2003), Scartascini and Crain (2002), Pablo T. Spiller and Tommasi (2003), and Iversen and Soskice (2002) describe the tension between simple median voter frameworks and, *inter alia*, the Rogowski-Kayser findings as the "electoral system puzzle." Iversen and Soskice (2002, p. 2). Moreover, Rogowski, Chang and Kayser (2002) expands the empirical tests of the price-level effect to a time-series cross-sectional study spanning OECD countries from 1970-2000. However, the problems of extrapolation and causal inference discussed in this paper are arguably even more severe in a panel setting.

anticipated from a one percent increase in votes" (p. 530), is assumed to be greater in majoritarian systems than in PR systems. That is, since the marginal vote in a majoritarian system can more drastically change the allocation of seats, politicians should cater more towards consumers in majoritarian as opposed to PR systems. Intuitively, the marginal difference between 49% and 51% of the vote share in a majoritarian system can *determine* the winner, whereas in a pure PR system, the additional 2% of the vote simply allocates more legislative seats *proportionally*.[2] Therefore, in a majoritarian system politicians will allegedly care more about the marginal voter, or the consumer.[3]

Third, RK posits that price levels, or more accurately deviations from competitive world prices, reflect the tradeoff between producers and consumers, with consumers favoring lower prices. Therefore, majoritarian electoral systems systems should systematically favor consumers by exhibiting lower price levels than PR countries.

[2]Of course in a PR system the number of seats in a district determines whether just *how* proportional the vote gain is translated into seats.

[3]Note that this claim relies on an assumption of a two-party competition with both parties receiving close to 50% of the vote. It is certainly possible that in a two-party competition where one party *ex ante* expects a much larger vote share than the competitor, the seats-votes elasticity would actually be lower for a majoritarian than a PR system. This is easily verified by examining the graph of the seats-votes elasticity in note 15, p. 531, of Rogowski and Kayser (2002) (replicated in Figure 2.7 of the Appendix). In their stylized example, PR systems are more pro-consumer when the vote of any one candidate exceeds 0.684 (i.e., to the right of the intersection between the PR and SMD curves).

	Specification		
	Model 1.1	Model 1.3	Model 1.4
Effect of SMD on Prices	-13.99^{**}	-12.40^{**}	-10.45^{*}
	(4.71)	(3.83)	(4.96)
Controls Included			
GDP per capita	*Yes*	*Yes*	*Yes*
Trade Openness	*No*	*Yes*	*Yes*
3 Year Exchange Rate Appreciation	*No*	*Yes*	*Yes*
Log of Arable Land / Population	*No*	*No*	*Yes*
Log of Population	*No*	*No*	*Yes*
Log of Energy Production / Consumption	*No*	*No*	*Yes*
N	24	22	22

Table 2.1: Summary Rogowski-Kayser findings in OLS regression analysis from Rogowski and Kayser (2002, p. 533). Standard errors in parentheses. $*p < 0.1$; $**p < 0.01$. Yes/No indicates whether variable was included in regression model.

2.1.2 Assessing the Price-Level Effect

Accordingly, RK uncovers a never before observed effect in their empirical study: single-member district (SMD) electoral systems lead to a decrease in roughly 12% of national prices, plus or minus 7%, compared to proportional representation systems (hereinafter, the "price-level effect") (Rogowski and Kayser 2002, p. 533, Model 1.3, using 95% confidence interval). Table 2.1 summarizes their main results and model specifications using linear regression. In their own words, "[t]he clear finding is that – controlling for virtually every other relevant influence – prices of goods and services are systematically higher in PR countries" (p. 526).

What RK overlooks is the model-dependency of the price-level effect due to the fact that countries with single-member district electoral systems are *systematically different* from countries with proportional representation systems in background characteristics that affect price levels.[4] SMD countries import substantially less goods as a proportion of GDP, are geographically much larger, have far more arable land, and have significantly higher ratios of energy production to consumption. Despite the fact that these variables are included in regression analysis, the imbalance in these characteristics between SMD and PR countries makes any inference about the causal effect of electoral systems highly model-dependent.[5] Any inference about the price-level effect may face the problem of extrapolating too far from the data. Intuitively, it may be very difficult to assess what the price levels of PR countries would be *if* they were SMD countries, given that SMD and PR countries are so drastically different.

RK is aware of the potential sensitivity of the prive-level effect, and conducts exemplary robustness tests using casewise diagnostics (p. 536-37). Yet a more direct technique to assess the robustness of causal effects given systematic differences between SMD and PR countries, namely propensity score matching, might have allowed RK to perform a more exhaustive analysis of the price-level effect. Propensity score matching has been widely applied to problems of causal inference in various disciplines with non-randomized observational studies (see, e.g., Holland 1986, Rosenbaum and Rubin 1983b, Dehejia and Wahba 1999, King and Zeng 2002, and Imai 2004). While it was not widely used in political science at the time, had the technique had been available to RK it would likely have led

[4]PR countries in their dataset are Spain, Norway, Belgium, Turkey, Germany, Japan, Sweden, Greece, Denmark, Ireland, Portugal, Netherlands, Austria, Finland, Italy, and Switzerland. SMD countries are Australia, Canada, France, New Zealand, the United Kingdom, and the United States.

[5]By imbalance, I simply mean that PR and SMD countries are not comparable along these dimensions.

to a different conclusion as shown herein.

Section 2 employs a more formal definition of extrapolation to assess the severity of the problem in the RK data. Section 3 provides an overview of the problem of causal inference as embodied in the comparative study of electoral institutions, and provides a brief formal rationale for using propensity score matching to assess the effect of SMD systems. Applying this framework, Section 4 finds that RK's claim of a "significant and robust" price-level effect of majoritarian electoral systems is incorrect (Rogowski and Kayser 2002, p. 526). Matching exposes a bias-variance tradeoff arising from the small number of comparable SMD and PR countries in the RK data. Section 5 therefore expands the original dataset from OECD countries to all democracies, but again finds little support for the broad claim of price-level effect. Section 6 concludes.

2.2 Extrapolation in the RK Data

Extrapolation is often informally described as an inference made outside of the available data range (Manski 1995, pp. 1-20). For example, suppose we are interested in the effect of trade openness on price levels. First consider an inference made *inside* of the available data range, or an interpolation (see King and Zeng 2002). RK's Model 1.2 would suggest that an increase of 1% in Portugal's trade openness would lead to a decrease of 0.5% in Portugal's price levels, plus or minus 0.4%. This qualifies as an interpolation since there are countries with similar levels of trade openness to approximate the counterfactual. An inference based on extrapolation, however, would suggest that an increase in 100% in

Portugal's trade openness is associated with a decrease of 50% in Portugal's price levels, plus or minus 80%. While the regression model would easily provide this first difference calculation, no countries with such high values of trade openness exist in our data to allow us to examine this counterfactual. Any inference based on such similar counterfactuals extrapolates too far, relying critically on model-dependent assumptions. There may, for example, exist declining marginal price effects for extremely open economies, which would indicate that we overestimated the first difference by assuming a linear additive effect. While model-based variance estimates do increase slightly as we extrapolate, this is *conditional* on the functional form assumptions of the model that identify a quantity of interest. Relaxing these assumptions will increase the variance of the quantity of interest, forcing us to examine only the *informative* data that identifies a quantity of interest without extrapolation (i.e., interpolation).

To define extrapolation more formally, let X denote an observed data matrix of explanatory covariates, and x a vector of values of the covariates, in this case representing characteristics of a particular country. Anytime that x is not a row contained in X, we are making a counterfactual inference, common to most if not all comparative statics claims. King and Zeng (2002, p. 7) define extrapolation stating that "questions that involve interpolation are values of the vector x which fall in the *convex hull* of X," where the convex hull is the smallest convex set containing a set of points.

To provide some intuition, suppose we simply had two variables, GDP per capita and price levels in the RK data. Figure 2.1 depicts the convex hull of this set of points, represented by the gray-shaded polygon with vertices at extreme points of the data, where each observed data point is labelled by the country abbreviation. The graph also provides simple fitted curves for a linear and quadratic specification, regressing prices on GDP per capita. The black squares, denoted by A, B, and C, represent counterfactuals. Suppose we were interested in the price level of a country with a GDP per capita of US $15,000, represented by A. Using both linear and quadratic models we would roughly estimate that this country would have price levels of 101, roughly equal to that of the United States. Since A falls within the convex hull, it qualifies as an interpolation. Now consider our estimation for counterfactual of a country with GDP per capita of US $40,000, represented by B and C. This counterfactual clearly falls outside of the convex hull and is thereby an extrapolation. More importantly, we can see that the model specifications yield price estimates that differ by roughly 11%, the vertical distance between B and C, despite the fact that the in-sample fit of these models is virtually identical. This illustrates that extrapolations rely to an extent far greater than interpolations on model specification. Linear and quadratic predictions diverge the more we extrapolate.

Similarly, the counterfactual of what the price levels of countries would be if PR countries were SMD countries relies on the existence of *comparable* SMD countries. Without comparable countries, the counterfactual extrapolation relies too critically on modeling assumptions. Note that there is no easy way to get around counterfactual reasoning: in order to learn about the effect of SMD systems on price levels, we use countries that are the same in all respects except for the electoral system.

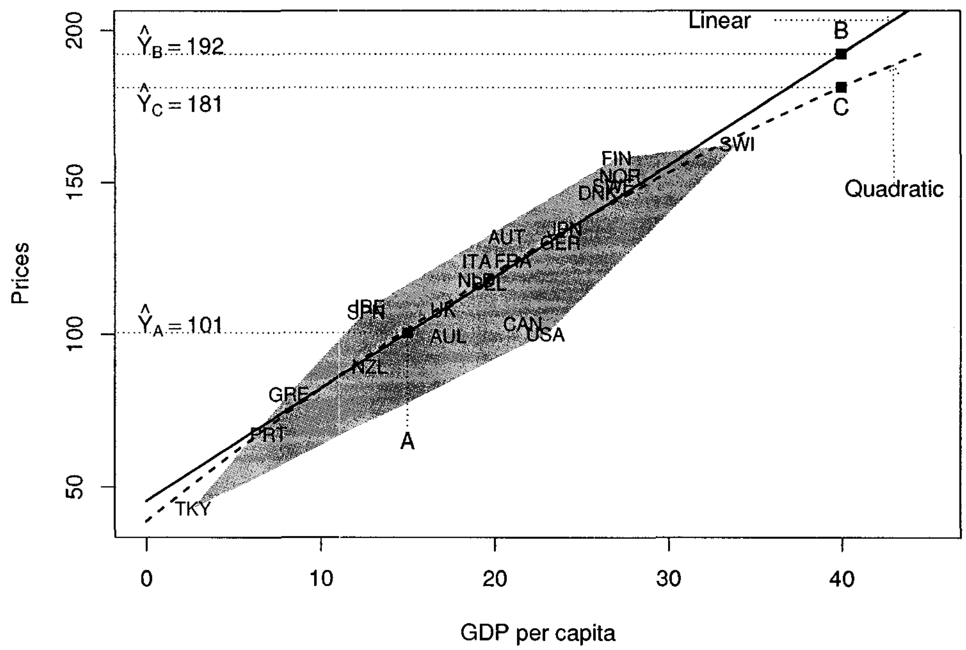


Figure 2.1: The gray polygon represents the convex hull of the data consisting of GDP per capita and price levels in the RK dataset. All countries are labelled. Dashed lines represent OLS estimations using a linear ($\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 GDP$) and quadratic ($\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 GDP + \hat{\beta}_2 GDP^2$) specification. Squares A, B, and C represent counterfactuals and $\hat{Y}_A$, $\hat{Y}_B$, and $\hat{Y}_C$ their predicted values.

To illustrate extrapolation in a simple setting, Figure 2.2 provides the density estimates (smooth versions of histograms) of SMD and PR countries for GDP per capita on the left panel x-axis and arable land on the right panel x-axis. The solid curves represent SMD countries and the dashed curves represent PR countries. In the left panel, note that PR countries have a far wider range of incomes. Unlike PR countries, there are no SMD countries with GDP per capita above US $30,000 or below US $8,000. An inference about the counterfactual of changing an SMD country to PR outside of this range where the

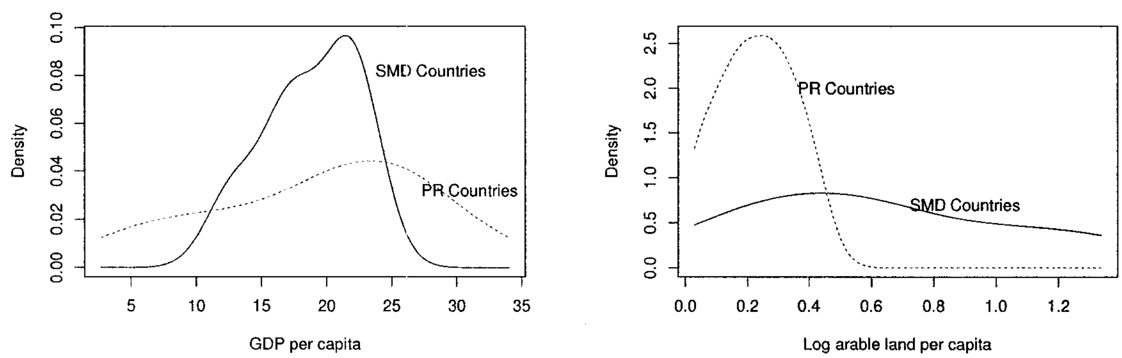


Figure 2.2: Density estimates of GDP per capita and arable land SMD countries (solid curve) and PR countries (dashed curve). In the left panel, the counterfactual of changing SMD to PR countries is outside of the convex hull for GDP per capita above US$ 30,000 and below US$ 8,000. In the right panel, the counterfactual of changing PR to SMD countries is outside of the convex hull above arable land of roughly 0.4.

densities overlap is an extrapolation. The right panel demonstrates that SMD countries have significantly more arable land per capita than PR countries. A counterfactual beyond arable land of roughly 0.4 constitutes extrapolation.

Consider now the implications for extrapolation in the latter instance of arable land. Figure 2.3 presents fitted values for two linear regression models using only arable land and SMD as variables. The left graph presents a reduced RK specification, where the electoral system is modelled as a dummy variable. The solid lines represent fitted values for SMD countries and the dashed lines represent fitted values for PR countries. With this specification we would estimate roughly an 8% price decrease at all values of arable land – simply the difference in the two parallel lines. While this model assumes linearity and additive treatment effects across the full range of arable land, note that there are only two SMD countries, signified by a box around the country label, that are within the range of

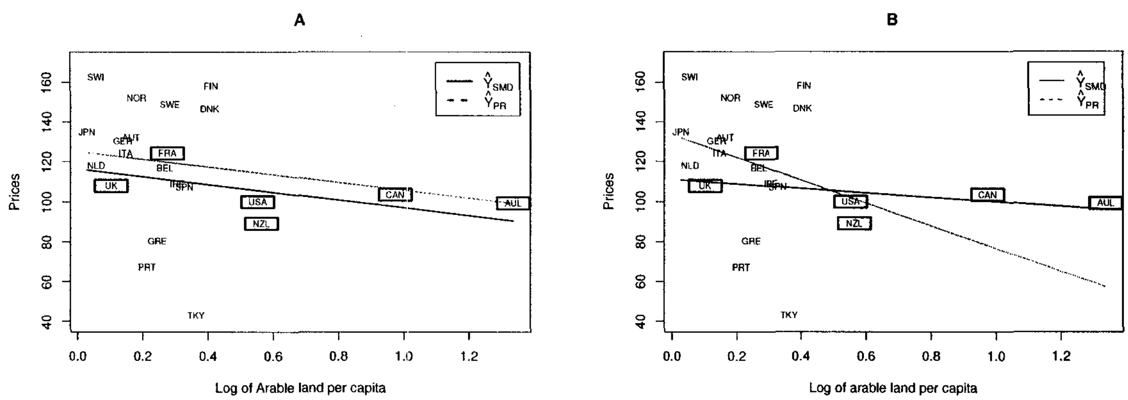


Figure 2.3: OLS fitted values for (a) model regressing prices on arable land and dummy variable of SMD (left panel) and (b) model regressing prices on arable land separately for SMD and PR countries (right panel). Solid lines represent fitted values for SMD countries and dashed lines represent fitted values for PR countries. Squares around country labels represent observed SMD country.

arable land for PR countries, signified by the country labels without boxes: only France and the United Kingdom are SMD countries that have arable land levels below 0.4. Any inference from our data about price levels of PR countries with *high* levels of arable land is highly model-dependent. The right panel presents the fitted values using separate regressions for PR and SMD countries. Note how the fitted lines now intersect roughly at arable land values of 0.45. In this specification, we would predict that at high ranges of arable land SMD countries have higher prices, wheras at low ranges of arable land SMD countries have lower prices than PR countries. In other words, the change in specification would lead to a complete reversal of the causal effect at high levels of arable land! Clearly, we should be cautious about extrapolation here. Note again here that interpolations are not as sensitive to the specification change – the estimate of the treatment effect within the range of overlapping values of arable land is quite similar.

While such simple two-dimensional examples illustrate the intuition of extrapolation analysis, a geometric visualization of the convex hull and the problem of extrapolation becomes more difficult as we increase the number of explanatory covariates. I employ King and Zeng's methods to calculate how many counterfactuals for the full set of covariates in RK's study are in the convex hull.[6] Each counterfactual is created by simply changing the electoral system from its observed value to its counterfactual (e.g., SMD countries to PR, or PR countries to SMD), while holding all other covariates at their observed values. Table 2.2 presents the results of this analysis. Even in RK's Model 1.1, which simply includes GDP per capita as a control, the majority of the counterfactuals are extrapolations. As Figure 2.2 illustrated, all six SMD countries are in the region of overlapping densities for GDP per capita whereas extreme values of the PR countries are not. Table 2.2 clearly reflects this, as 100% of the counterfactuals changing SMD countries to PR countries are in the convex hull, whereas only 28% of the counterfactuals changing PR countries to SMD countries are. This suggests that the data might better be able to explain the effect of SMD on SMD countries than on PR countries. If we define SMD as a treatment, this would be referred to as the treatment effect on the treated. Extrapolation becomes quickly more severe with more fully specified models. In Model 1.3, only 18% of the counterfactuals are inside the convex hull. Even worse, in Model 1.4 *none* of the counterfactuals are inside the convex hull of the data. In other words, in the fully-specified model, every counterfactual inference is an extrapolation! The consequences of such extrapolation are demonstrated by simple

[6]The software by King and Zeng to solve this linear programming problem is available at `http://GKing.Harvard.Edu`.

		Number of counterfactuals in convex hull					
		SMD → PR		PR → SMD		Total	
Model	N	No.	%	No.	%	No.	%
RK 1.1	24	6	100	5	28	11	46
RK 1.3	22	3	50	1	6	4	18
RK 1.4	22	0	0	0	0	0	0

Table 2.2: Number and percentage of counterfactuals falling within convex hull of data. SMD (PR) → PR (SMD) indicates a counterfactual of changing an SMD (PR) country to PR (SMD) while holding all other covariates at their observed values. RK refers to Rogowski-Kayser specifications.

changes in the model specification, the traditional way to conduct sensitivity analyses of model-dependent conclusions. For example, in Model 1.4, the substitution of a squared term for GDP per capita yields an insignificant coefficient for SMD (p=0.13). Similarly, adding a squared term for energy leads to an insignificant coefficient for SMD (p=0.13). Or the addition of an interaction term between SMD and energy results in no substantive effect of the electoral system. (See Appendix, Figure 2.8.)

Yet of course not all extrapolations are equally bad. An extrapolation just outside of the convex hull that has many data points close by may even represent a better inference than an interpolation that has an isolated datapoint bounding the convex hull. To go back to our earlier example of income and price levels, Switzerland has the highest income in the RK dataset, with a GDP per capita of roughly US $34,000. A counterfactual of what its price level would be if its income increased to US $35,000 is obviously a better inference than if its income increased to US $100,000, yet both are outside the convex hull. Therefore, we turn to a framework of causal inference that explicitly takes into account degrees of differences between SMD and PR countries.

2.3 Assessing the Causal Effect of Electoral Systems

2.3.1 The Causal Effect

A cause is a intervention that leads to some observable effect in our outcome variable as compared to no treatment (i.e., control). More formally, let i index the countries of our interest. Let $Y_i(1)$ represent the potential outcome of interest when the country is SMD and $Y_i(0)$ represent the potential outcome when the country is PR. The treatment effect for country i is defined as $Y_i(1)-Y_i(0)$. The average treatment effect for the treated population is defined as:

$$E(Y_i(1)|T_i=1)-E(Y_i(0)|T_i=1),$$

where T_i signifies whether country i was assigned treatment. The notation in this section largely trails that of Dehejia and Wahba (2002) and is consistent with the standard literature on causal inference (see, e.g., Holland 1986 and King, Keohane and Verba 1994). The fact that we cannot observe a control country when it has been assigned treatment, $Y_i(0)|T_i=1$ is known as the fundamental problem of causal inference (Holland 1986, p. 947, King, Keohane and Verba 1994, p. 79). Concretely, we can never observe the counterfactual of a country under a PR system when it in fact has been "assigned" a majoritarian system, or the counterfactual of a country under a majoritarian system when it in fact has been "assigned" a PR system. In experimental studies, randomization provides a way of estimating the treatment effect by balancing all countries on characteristics and thereby:

$$E(Y_i(0)|T_i=1)=E(Y_i(0)|T_i=0).$$

Randomization therefore plays the central role of balancing all other covariates. By randomly assigning countries a majoritarian electoral system, the effect of other factors (e.g., geography, GDP per capita, imports/GDP) balances out, and the difference that remains can be attributed to the treatment.

2.3.2 Majoritarian Versus PR Countries

Table 2.3 depicts simple means comparisons between PR and SMD countries. SMD countries are strikingly different from PR countries. Import penetration is 60% lower on average in SMD countries than in PR countries (p=0.008), and SMD countries on average trade 8% less as a proportion of GDP, though this is only borderline significant. Moreover, it appears that PR countries have significantly less arable land per population and have lower energy productions, both at significant levels (p=0.078 and 0.027, respectively). Such differences make it extremely difficult to assess what the causal effect of the electoral systems might be – as shown earlier, inferences based on OLS extrapolation are highly model-dependent.

2.4 Matching on the Propensity Score

Given that differences between SMD and PR countries exist on multiple dimensions, our goal is to match countries on observable characteristics so as to obtain an unbiased estimate of the treatment effect. The intuition is that we only care about countries that are closely aligned in observable characteristics such as import penetration and factor endow-

Variable	SMD Mean	PR Mean	T-Statistic
Per Capita GDP	18.81	19.41	0.23
Trade Openness	0.07	−0.01	−1.67
3 Year Exchange Rate Appreciation	6.68	−6.48	−0.98
Log of Arable Land / Population	0.64	0.23	**−2.18**
Log of Population	10.4	9.55	−1.29
Log of Energy Production/Consumption	0.88	0.51	**−2.44**
Imports as a proportion of GDP	0.18	0.30	**2.93**
Size (1,000 km^2)	4,754	244	**-2.26**

Table 2.3: T-statistics comparing single-member district countries (SMD) and proportional representation countries (PR) on selected covariates. Bold signifies $|p| < 0.10$. $N = 22$.

ments. The Netherlands, Greece, and Switzerland, for example, have far higher import penetration levels than all SMD countries. Using such countries that are highly unlikely to have SMD systems to estimate $Y_i(0)|T_i = 1$ may lead to significant bias. In short, the key objective is to assess the causal effect by examining countries comparable in all covariates but the electoral system.

In the simple case of one additional explanatory variable x_i, the intuition is simple: we match PR and SMD countries with similar values of x_i and the difference in outcome variables is our estimate of the treatment effect. Matching on the host of covariates in RK's models, however, proves difficult when exact matches are unavailable since it is unclear which covariates matter and how to weight them. Each additional variable leads us down to the curse of dimensionality, namely that each additional variable increases the number of parameters required to estimate $E(Y|X)$ geometrically in the number of covariates.

Therefore, to summarize these differences, I calculate the propensity score, $e_i(X_i) = Pr(T_i = 1|X_i)$, which represents the probability that a country has an SMD system conditional on other covariates X_i. The propensity score is akin to a one-dimensional summary of all relevant covariates, and Rosenbaum and Rubin (1983b) shows in a seminal result that conditioning on the propensity score is bias reducing in the same degree as conditioning on all the covariates. Matching on the propensity score thereby enables us to estimate the treatment effect, where:

$$E(Y_i(1)|T_i = 1, X_i) - E(Y_i(0)|T_i = 1, X_i) = E(Y_i(1)|T_i = 1, e_i(X_i)) - E(Y_i(0)|T_i = 0, e_i(X_i)).$$

In other words, once we have found countries that are equally likely to be majoritarian conditional on X, a comparison of SMD and PR countries provides a less biased estimate of the causal effect of SMD on SMD countries.

Specifically, King and Zeng (2002) discusses the sources of bias in estimating causal effects, generalizing Heckman, Ichimura and Todd (1997). Bias in causal inference is due to four sources: (a) omitted variable bias, (b) controlling for variables affected by the main causal variable, (c) non-overlapping bias, and (d) density difference bias.

Bias due to (a) is already well-known: omitting a variable that affects price levels and is correlated with the electoral system may bias the causal estimate. This is most evident in the RK Models 1.1 and 1.3, which is why this paper focuses on the fully-specified Model 1.4, though results hold across all models.

Post-treatment bias (b) occurs when a variable that is an effect of the cause is included as a control. Earlier studies on the effects of electoral systems on government expenditures, welfare spending, income equality, and stability might certainly support such a critique of RK. For example, if SMD affects GDP per capita (e.g., through any of the prior effects) we may not be able to isolate the effect due to SMD systems. Rogowski and Kayser 2002,

p. 537 generally recognizes that "[o]f course the choice of electoral institutions is itself ultimately endogenous," yet since their empirical models nonetheless assume exogeneity of these covariates, I do not relax that assumption here.[7]

Matching concerns itself with extrapolation due to *non-overlapping bias* (c), stemming from the fact that SMD countries take on certain values of X which no PR countries do, and *density difference bias* (d), stemming for example from the fact that while both SMD and PR countries have low levels of arable land, many more PR observations take on low values than SMD observations. These multidimensional difference are succinctly summarized by density graphs of the propensity scores by electoral system. In such graphs, ranges of where the density of the propensity score between SMD and PR countries differ indicate bias due to (d), whereas a nonzero density of SMD (PR) and a zero density of PR (SMD) would indicate bias due to (c).

Note also that this matching technique is semi-parametric. That is, at this point we make no functional form assumptions about effect of X on Y and we only care to get a balance of X, thereby addressing problems arising from extrapolation. In fact, this method makes uniformly fewer assumptions that regression analysis.[8] The key simply lies in obtaining

[7]If anything, allowing for cross-cutting effects would yield more difficulties identifying the causal effect. Hence, the standard errors here are, if anything, likely to be conservative.

[8]More formally, propensity score matching assumes that $(Y(1),Y(0)) \perp\!\!\!\perp T|X$ and $0 < Pr(T = 1|X) < 1$, where independence $\perp\!\!\!\perp$ indicates that $P(T,Y(1),Y(0)|X) = P(T|X)$. These are akin to the exogeneity assumption in regression analysis, namely that treatment is assigned based on observable differences (i.e., no omitted variable bias) and of course that X covariates are not a result of treatment, but without functional form assumptions. In order for regression analysis to converge to the matching estimates one must addi-

good matches, or more specifically, in obtaining matching treatment and control groups that have comparable propensity scores. The procedure is akin to matching in the earlier case of one covariate. We match SMD and PR countries that have similar propensity scores, and calculate the difference in means of the outcome variable to estimate the treatment effect.

2.4.1 Findings

The left panel in Figure 2.4 graphs the estimated propensity scores for all countries by electoral system, using the covariates from RK Model 1.4. The propensity scores are estimated using a logit model, where $P(T_i = 1|\beta, X) = \frac{1}{1+e^{-x_i\beta}}$. Since the primary role of the propensity score is to balance covariates, the actual β's in this step of the estimation are of little substantive significance. I present findings only for Model 1.4 because this is the fully specified model that is least likely to suffer from omitted variable bias, though similar results hold for the other specifications.

The panel summarizes the drastic differences we uncovered before. In fact only one PR country even falls within the range of SMD country propensities (Spain, $\hat{e}_i(X) = 0.773$)! Given this drastic difference, parametric estimates using this data are subject to non-overlapping and density difference bias.

The right panel of Figure 2.4 presents the distributions of SMD countries after having matched comparable PR countries using a nearest-neighbor matching algorithm with re-

tionally assume (1) constant additive treatment effect, $Y_i(1) - Y_i(0) = \alpha \forall i$, and (2) linearity in all covariates, $E(Y_i(0)|X_i) = X'\gamma$.

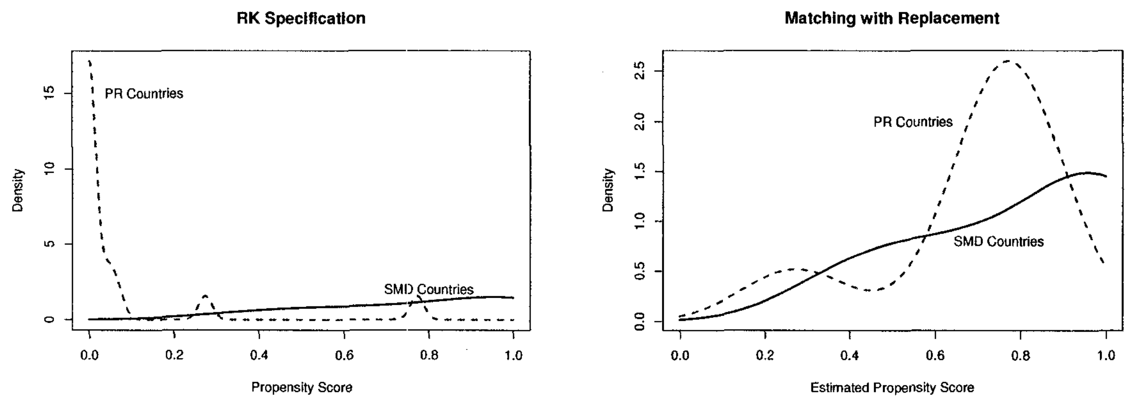


Figure 2.4: Density estimates (smooth versions of histograms) of estimated propensity scores for SMD and PR countries in original specification of Model 1.4 (left panel) and matching model (right panel).

placement.[9] While the distributions are still discernably different, some common support is at least obtained.

Recall though that the only purpose of the propensity score is merely to balance the pre-treatment covariates. Table 2.4 presents mean differences between SMD countries and matched PR countries after matching. Matching substantially reduces differences of SMD and PR countries arable land, energy, and import penetration.[10]

[9]The procedure for nearest-neighbor matching, the simplest of all matching algorithms, is as follows. Given X_i, I estimate $e_i(X_i) = Pr(T_i = 1|X_i)$. For an SMD country i, I then select some PR country j so as to minimize the absolute distance between propensity scores, $j(i) = \text{argmin}_j|(e_i|T_i = 1) - (e_j|T_j = 0)|\forall j \neq i$. Each SMD country is thereby matched with a PR country. To estimate the average treatment effect, we simply subtract outcomes of the matched SMD and PR countries. Depending on the bias-variance tradeoff in the data, other aspects of matching, such as whether to discard non-overlapping units or whether to match with replacement, become relevant. For a more general reference on matching algorithms, see Rosenbaum (2nd edition, 2002).

[10]Despite the fact that import penetration and size are specified in the Rogowski-Kayser models, they are

	T Statistics	
Variable	RK Model	Matching Model
Per Capita GDP	0.23	0.00
Trade Openness	−1.67	−0.98
3 Year Exchange Rate Appreciation	−0.98	−0.89
Log of Arable Land /Population	−2.18*	−1.98
Log of Population	−1.29	−0.71
Log of Energy Production /Consumption	−2.44*	−0.52
Imports as a proportion of GDP	2.93*	1.58
Size (1,000 km2)	−2.26*	−2.18*

Table 2.4: T-statistics comparing single-member district countries (SMD) and matched proportional representation countries (SMD) on selected covariates. * signifies $|p| < 0.10$. $N = 22$.

Method	Effect of SMD, ATE	Std Err
Matching - No Replacement	−11.3	12.5
Matching - Replacement	−15.6	15.7
RK Estimate	−10.5	5.0

Table 2.5: Estimated average treatment effects with one-to-one nearest neighbor matching, without discarding. All matching models match six SMD countries with six PR countries. Standard errors were calculated using 500 bootstrapped samples.

Table 2.5 summarizes the estimates of the treatment effects. Given that matching with replacement is preferable when treatment and control units are vastly different (see Dehejia and Wahba 2002), our best estimate yield a *highly uncertain* treatment effect if it exists at all: after accounting for systematic differences between SMD and PR countries, majoritarian systems lead to a 16% decrease in prices, plus or minus 31%! The standard errors are more than three times as high as in the RK findings.

component parts of trade openness and factor endowment instruments and still balance with inclusion of Model 1.4 covariates in the balancing score.

Matching on the propensity score exposes the small sample size problem in the RK data: only six matched PR countries remain after we discount incomparable PR countries, leaving us with highly uncertain estimates. Hence the only way to reassess the price-level effect is through a larger dataset.

2.5 A Test with New Data on all Democracies

As a preliminary matter, note that RK offers no theoretical reason to believe *a priori* that the price-level effect should be restricted only to OECD countries. Indeed if the RK argument is correct, we should observe this dynamic across all democratic countries. I thereby assemble a new dataset using the same sources and specifications of the original RK study where possible to investigate the price-level effect across all democratic countries, hopefully allowing us also to obtain better matches. Data was gathered for all 73 democratic countries in Perrson and Tabellini (2003), who based their selection of democracies on the Gastil and Polity indices. Data description and sources are provided in Table 2.9 in the Appendix. The only data source that differs from RK consists of the World Development Indicators for domestic energy production, which in the original RK source is only available for OECD countries. Combining these data sources changes none of the conclusions herein. The new data exhibits a substantial missing data problem, with so few observations that make even multiple imputation infeasible. Nevertheless, the new data allows to potentially improve the comparable matched countries to obtain a less biased estimate.

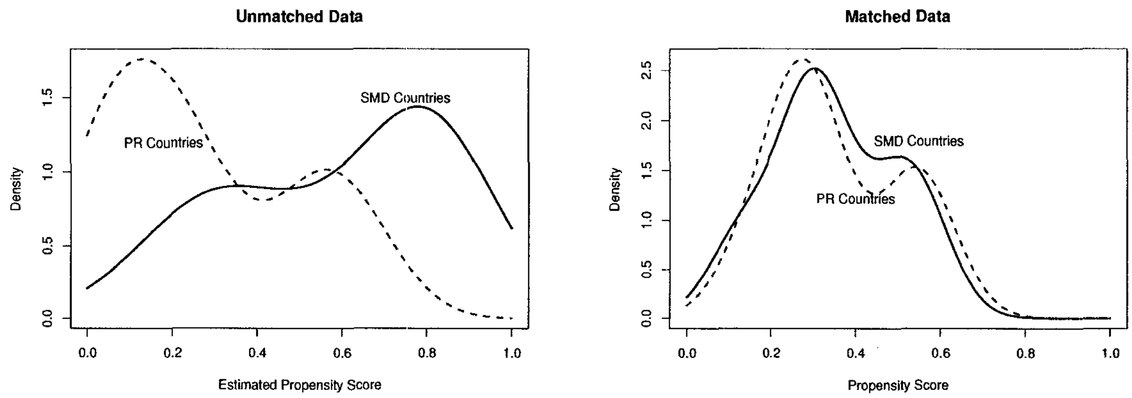


Figure 2.5: Density estimates of estimated propensity scores for the new dataset without matching (left panel) and with matching (right panel).

2.5.1 Matching

With an expanded dataset and new observations, we observe a substantial reduction in the extrapolation problem. The left panel in Figure 2.5 graphs the density estimates of PR and SMD countries, showing that there is significantly more overlap in the support (non-zero density).

Table 2.6 provides means tests for SMD and PR countries. SMD and PR countries appear to exhibit fewer systematic differences in this new dataset, largely reflecting the high variance in most covariates with the inclusion of the more heterogeneous non-OECD countries. For example, no longer are all the SMD countries in the dataset large and rich in factor endowments, as evidenced by the addition of SMD countries such as Malaysia and Malawi. The assumption that we can match well is therefore likely to hold with this new data.

After performing nearest-neighbor matching, the balance is substantially improved.

Variable	SMD Mean	PR Mean	T-Statistic	Matched
GDP per capita	6.10	10.54	1.87	−0.04
Trade Openness	0.21	0.25	2.06*	−0.18
Exchange Rate	−53.91	−127.00	−0.88	−0.64
Arable Land	0.18	0.23	0.98	−0.18
Population	9.66	9.19	−1.10	1.42
Energy	0.95	0.72	−1.01	−0.73
Imports	38.46	43.14	0.79	0.77
Size (1,000 km2)	340	1011	1.69	1.05

Table 2.6: T-statistics of new dataset comparing single-member district countries (SMD) and matched proportional representation countries (PR) on selected covariates. $N = 67$.

The right column of Table 2.6 summarizes the balance of covariates across all matched SMD and PR countries. There are no statistically significant differences between the matched PR and SMD countries, and all means differences except for those of population decrease with matching.

Most importantly, as this diagnostic already indicates, we now obtain better matches by matching with replacement and discarding non-overlapping units.[11] In the original dataset, discarding non-overlapping units would have left only two total countries. The right panel Figure 2.5 depicts the density estimates of the propensity scores for matched countries. Note that the densities are close – a drastic improvement over the matches using the original data. This being the crucial assumption of matching methods, estimating the causal effect

[11]Specifically, treatment units with propensity scores above the maximum propensity score of control units and control units with propensity scores below the minimum of propensity scores of treatment units were discarded. This addresses non-overlapping bias. As King and Zeng point out, these discarded units in themselves may be of substantive interest, allowing us to determine which counterfactuals are reasonable comparisons, akin to choosing case studies. Replacement reduces density difference bias by providing better matches when treatment and control distributions are quite different.

now becomes a simple difference in means calculation.[12]

Table 2.7 presents the estimates of the causal effect of a majoritarian electoral system on price levels. We obtain the best matches by discarding outliers and matching with replacement, thereby minimizing non-overlapping and density difference bias. This yields a treatment effect SMD on prices of roughly 2.1%, plus or minus 35% – again, a highly uncertain effect. From the left columns we can see that outliers appear to be driving the price-level effect, and even then not at any significant levels. In fact, with these specification, the price-level effect confidence interval ranges from as high as 37% to as low as -60% in the matching models. Again, we discover the bias-variance tradeoff: the RK price-level effect is an artifact of the functional form imposed by the regression, and by comparing similar PR and SMD countries (using the same exact datasources and covariates), we find little evidence for the broad price-level effect. Figure 2.6 graphs the substantive difference between the RK and the matching models, demonstrating the tradeoff.

Lastly, note it is possible that the new data introduces more *unobserved* heterogeneity by including non-OECD countries. Capturing these differences between OECD and non-OECD countries by matching on an indicator variable for OECD, however, yields the same inconclusive price-level effect. Nevertheless, to the degree that there are important omitted

[12]Estimation of the standard errors is not quite this simple, requiring us to bootstrap the entire sample, which is standard in the literature (see, e.g., Dehejia and Wahba (2002), Imai (2004)). The bootstrap procedure draws samples from the donor pool data with replacement, matches on the propensity score within each bootstrapped sample, then calculates the ATE for the matched units in the bootstrapped sample, and finally obtains the variance across the bootstrapped estimands.

	Discard Outliers					
	(1) No			(2) Yes		
Matching Method	ATE	SE	Matches	ATE	SE	Matches
No Replacement	-17.6	9.0	13	−0.1	14.9	6
Replacement	-19.0	20.3	13	2.1	17.5	6
RK Specification				−10.5	5.0	

Table 2.7: Estimated average treatment effects (ATE) of majoritarian electoral systems on prices with one-to-one nearest-neighbor matching. Standard Errors (SE) calculated using 500 bootstrapped samples. Matches refers to number of paired SMD and PR countries.

variables, a criticism that would just as well apply to the original dataset, it would be productive for researchers to gather this additional data.[13]

2.6 Conclusion

These results significantly question the "clear finding that . . . prices of goods and services are systematically higher in PR countries" (Rogowski and Kayser 2002, p. 526). Several comments bear note here. First, these findings do not necessarily undermine the important line of research of RK. Much to the contrary, this paper suggests locating the causal mechanism with a finer comb. Perhaps a more refined measurement of seats-votes elasticity (which could be district-specific) might pinpoint the price-level effect more specifically (see, e.g., King 1989). Perhaps the assumption that majoritarian systems have categorically higher seats-votes elasticities does not fit the empirical reality of multiple parties and

[13]It is of course possible to conduct a sensitivity analysis here due to omitted variables, as outlined in Rosenbaum and Rubin (1983a) and Imbens (2002), but such an analysis would likely serve only to widen confidence intervals. In other words, assuming exogeneity, the standard errors here are if anything *conservative*.

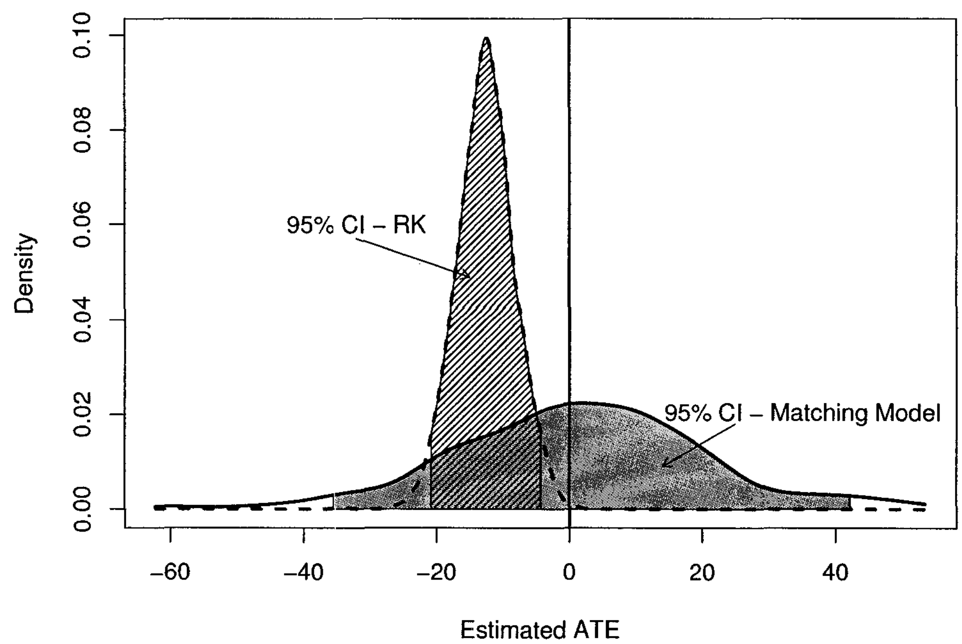


Figure 2.6: Bias-variance tradeoff: Comparison of average treatment effect of majoritarian system on price levels, simulated from RK Model 1.3 (holding other variables at median values) and matching model with new dataset (with replacement and discarding outliers). Shaded areas represent 95% confidence intervals, and vertical line represents no impact of SMD on price levels.

vote shares that might deviate widely from a 50% vote share. Recall that the two-party and roughly 50% vote share assumptions were crucial to this claim, which would suggest stark omitted variable bias due to exclusion of expected voteshares. Or perhaps prices are too rough an indicator of the tradeoff between consumers and producers. What is certain, however, is that the systematic differences between SMD and PR countries make a convincing empirical test of the price-level effect much more difficult than originally claimed. Nonetheless, any of these above refinements might help us pinpoint more accurately where and under what conditions the causal effect might exist. Rather than discretize electoral rules into SMD and PR systems we may need to examine the complexities of electoral rules that Cox (1990) for example has classified. Changes in the electoral system in Japan, New Zealand, the Philippines, and the Ukraine in the 1990s might also prove fruitful in narrowing down the effect of electoral systems on price levels. Such "natural experiments" arguably adhere more closely to the assumptions of randomized treatment assignment (see, e.g., Rosenbluth and Schaap 2002). Second, the methods developed should prove useful not only to quantitative researchers seeking to determine what their data can tell them, but also to qualitatively-oriented researchers seeking to formalize case selection criteria. Propensity score matching permits researchers to choose good cases, thereby aiding in the design of their study before even venturing into the field. Lastly, one could well retreat from the position that electoral systems have a *causal effect* on prices, but that arguably requires abandoning one, if not the, central claim of RK.

2.7 Appendix

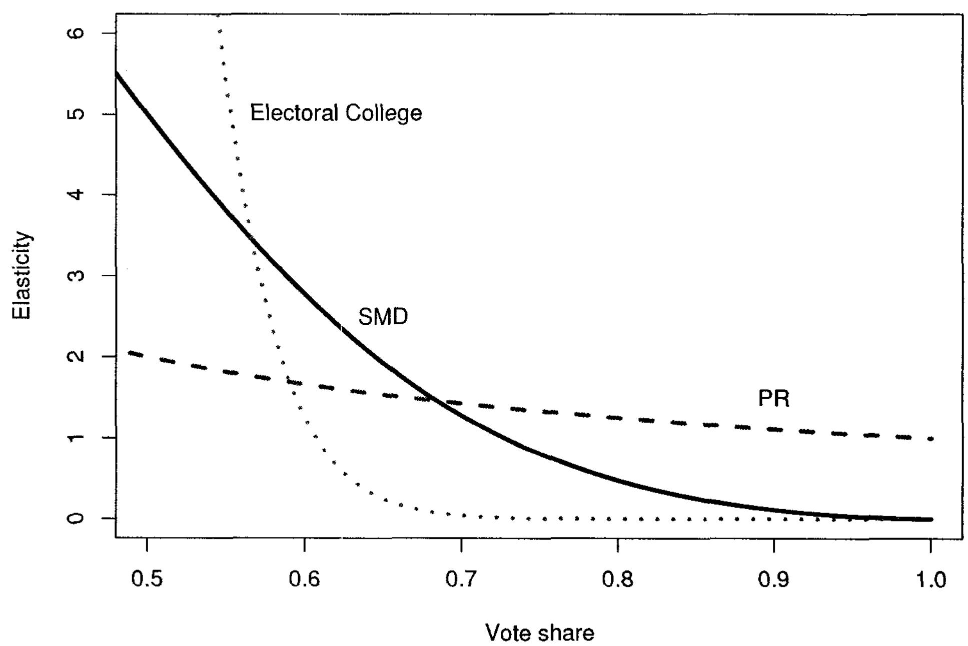


Figure 2.7: Seats-Vote elasticity as a function of vote share in three electoral systems. Replicated from RK Note 15, p. 531.

	Dependent Variable: GDP Price		
Constant	88.37***	100.97***	90.84***
	(16.41)	(19.61)	(16.48)
GDP per capita	2.81***	.	2.82***
	(0.29)	(0.49)	(0.29)
Majoritarian	**-6.93**	**-10.68**	**-9.88**
	(14.35)	**(6.58)**	**(6.04)**
Trade openness	-68.00**	-82.61**	-68.88**
	(23.06)	(28.89)	(23.01)
Exchange Rate Appreciation	0.12**	0.22**	0.12**
	(0.04)	(0.05)	(0.04)
Ln(Arable land/pop)	3.54	4.88	1.90
	(10.63)	(9.86)	(7.89)
Ln(Population)	-2.58	-1.30	-2.70
	(1.53)	(1.89)	(1.49)
Ln(Energy)	-3.08	5.42	-0.42
	(4.63)	(5.60)	(18.08)
Maj * Ln(Energy)	**-5.23**	.	.
	(19.88)	.	.
GDP^2	.	0.06***	.
	.	0.01	.
$Ln(Energy)^2$	.	.	1.40
	.	.	(7.80)
R^2	0.97	0.94	0.97
F	45.97	31.1	45.84
N	22.	22.	22.

Table 2.8: Implications of extrapolation – effects of interaction and squared terms on original estimates. Bolded figures indicate substantive impact of SMD on price levels. Standard errors in parentheses. $*p < 0.1$; $**p < 0.05$; $***p < 0.01$

Variable	Description	Source
Price Level	Price level GDP [%] (PPP GDP/ U.S. dollar exchange rate)	Penn World Tables, 5.6
Majoritarian electoral system	Dummy variable scored as 1 if all the lower house is elected under plurality rule, 0 otherwise	Perrson and Tabellini (2003)
GDP per capita	GDP per capita in US$1,000	GDP: IMF IFS rf..zf, Population: Penn World Tables, 5.6
Exchange Rate Appreciation	Percentage change in NC/USD exchange rate since 1987	IMF IFS rf..zf
Arable land	Log of ((arable land in hectares / population)+1)	World Development Indicators (2002)
Energy	Log of (commercial energy production (kt of oil equivalent) / commercial energy use (kt of oil equivalent))	World Development Indicators (2002)
Population	Natural log of population in millian inhabitants	Penn World Tables, 5.6
Openness	Measure of "Free trade openness" instrumented as $Openness = 0.528 - 0.026 log(area) - 0.095 log(dist)$, where *area* represents the size of the country in million square km and *dist* represents the average distance to capitals of world 20 major exporters, weighted by values of biliateral imports in 1000 km	Lee (June 1993)

Table 2.9: Variable description and data sources of new dataset including non-OECD democracies

Chapter 3

Assessing the Effect of War on the Supreme Court[1]

> The Constitution of the United States is a law for rulers and people, equally in war and in peace ... When peace prevails, and the authority of government is undisputed, there is no difficulty of preserving the safeguards of liberty ... but if society is disturbed by civil commotion ... these safeguards need and should receive the watchful care of those intrusted with the guardianship of the Constitution and laws.
>
> —*Ex parte Milligan*[2]

> We uphold the order [to exclude those of Japanese ancestry from the West Coast war area]... Compulsory exclusion of large groups of citizens from their homes, except under circumstances of direst emergency and peril, is inconsistent with our basic governmental institutions. But when under conditions of modern warfare our shores are threatened by hostile forces, the power to protect must be commensurate with the threatened danger.
>
> —*Korematsu v. United States*[3]

3.1 Introduction

Running through these quotes, taken from landmark U.S. Supreme Court decisions, is a common strand: In both, the justices seem to suggest that their institution ought play

[1]This is an abbreviated version of an article written in law review style.
[2]71 U.S. 2, 120 (1866).
[3]323 U.S. 214, 219-220 (1944).

a different role in times of "emergency and peril" than when "peace prevails."[4] But the cases stand for fundamentally different propositions about that role. *Milligan* implies that the justices must become especially vigilant in protecting rights and liberties during "commotions;"[5] *Korematsu* commends quite the opposite: that the justices ought be especially willing to subordinate rights and liberties when America is "threatened."[6] If *Korematsu* is testimony to the continued viability of Cicero's maxim *inter arma silent leges* ("during war law is silent"),[7] as many suggest that it is,[8] then *Milligan* provides a counter punch: During

[4]We invoke the terms "emergency and peril," "commotions," and "crisis" here to signify major international events, including (but not limited to) war, that threaten the security of the nation. In Part 3.2 we provide more precise definitions.

[5]This also corresponds with Oren Gross, *Chaos and Rules: Should Responses to Violent Crises Always Be Constitutional?*, 112 YALE L.J. 1011 (2003) at 1043-58 (noting that "[u]nder the Business as Usual model of emergency powers, a state of emergency does not justify a deviation from the 'normal' legal system ... Thus, Justice Davis could state in Ex Parte Milligan that the Constitution applied equally in times of war as well as in times of peace.") (citations omitted). See also Jules Lobel, *The War On Terrorism and Civil Liberites*, 63 U. PITT. L. REV. 767 (2002); (describing the Milligan perspective as "absolutist").

[6]See, e.g., Harry T. Edwards, *The Judicial Function and the Elusive Goal of Principled Decisionmaking*, 1991 WIS. L. REV. 837 (1991); Eugene V. Rostow, *The Japanese American Cases—A Disaster*, 54 YALE L. J. 489 (1945); Charles Fairman, *Law of Martial Rule and the National Emergency*, 55 HARV. L. REV. 1253 (1941-1942); John P. Frank, *Ex Parte Milligan v. The Five Companies: Marital Law in Hawaii*, 44 COLUM. L. REV. 639 (1944).

[7]To be precise, Cicero's phrase was "silent enim leges inter arma" (in battle, indeed, the laws are silent). CICERO, PRO MILONE16 (N.H. Watts trans., Harvard Univ. Press, 5th ed. 1972).

[8]Joel B. Grossman, *The Japanese American Cases and the Vagaries of Constitutional Adjudication in Wartime: An Institutional Perspective*, 19 HAW. L. REV. 649 (1997); WILLIAM H. REHNQUIST, ALL THE LAWS BUT ONE: CIVIL LIBERTIES IN WARTIME (1998); Martin S. Sheffer, THE JUDICIAL DEVELOP-

war the law speaks loudly.[9]

While the Court ignored *Milligan* in its *Korematsu* decision,[10] and subsequently has repudiated at least the "racist basis" of *Korematsu*, it has overruled neither decision;[11] both, in the eyes of the justices, apparently remain valid law.[12] But not so in the eyes of many members of the legal community. To an overwhelming majority, the Court's jurisprudence in times of crisis is far more in line with the dictates of *Korematsu* than with the language of *Milligan*.[13] Indeed, the belief that the Court acts to suppress rights and liberties under

---

MENT OF PRESIDENTIAL POWERS (1999).

[9]Indeed, in 1866 the New York Times, in commenting on Ex Parte Milligan, noted that "[t]he experience of our past history showed the wisdom of the framers of the Constitution, in constructing it to be alike efficient in war as in peace." *Washington: Special Dispatches to the New York Times*, N.Y. TIMES, Dec. 18, 1866, at p. 1.

[10]In fact, not once did the Court cite *Milligan* in *Korematsu*.

[11]Grossman, *supra* note 8. While it is true that the U.S. Supreme Court has not overruled itself, a federal district court reversed Korematsu's original conviction in Korematsu v. United States, 584 F. Supp. 1406 (D.C. Cal, 1984). Moreover, Congress apologized for the internment of Japanese Americans and provided for reparations in the Civil Liberties Act of 1988, Pub. L. No. 100-383, 102 Stat. 903 (1988) (codified at 50 U.S.C. app. §1989b-1989b-9 (2003).

[12]See, e.g., Edwards, *supra* note 6; David P. Currie, *The Constitution in the Supreme Court: Civil War and Reconstruction, 1865-1873*, 51 U. CHI. L. REV. 131 (1984); Dean Masaru Hashimoto, *The Legacy of Korematsu v. United States: A Dangerous Narrative Retold*, 4 ASIAN PAC. AM. L. J. 72 (1996).

[13]Some simply dismiss the importance of Milligan altogether, arguing that it cannot be taken to stand for the proposition that justices must become especially vigilant guardians of the Constitution during times of war because (a) the case was decided after the Civil War ended and therefore cannot shed light on how the Supreme Court acts during times of war, and (b) the case, despite its language, begs the question of how the Constitution applies in war and peace. See, e.g., See Donald A. Downs & Erik Kinnunen, *A Response to*

conditions of threat is so widely accepted today in post-September 11th America,[14] and has been so widely accepted since the World War I period,[15] that most observers no longer

Anthony Lewis: Civil Liberties in a New Kind of War, 2003 WIS. L. REV. 385, 394. The empirical analysis we offer in infra Part 3.3 addresses both these concerns by providing an explicit and exogenous framework for distinguishing war and peace cases, as well as a transparent measurement of the causal effect of war on case outcomes.

[14]See, e.g., Floyd Abrams, *The First Amendment and the War Against Terrorism*, 5 UNIV. PENN. J. CON. L. 1 (2002); Gross *supra* note 5; Philip B. Heymann, *Civil Liberties and Human Rights in the Aftermath of September 11*, 25 HARV. J.L. & PUB. POL'Y 440 (2002); Neal K. Katyal & Laurence H. Tribe, *Waging War, Deciding Guilt: Trying the Military Tribunals*, 111 YALE L.J. 1259 (2002); Sanford Levinson, *What is the Constitution's Role in Wartime: Why Free Speech and Other Rights Are Not as Safe as You Might Think* 2001, available at: http://writ.news.findlaw.com/commentary/20011017_levinson.html (last accessed on February 17, 2003); Anthony Lewis, *Civil Liberties in a Time of Terror*, 2003 WIS. L. REV. 257; Anthony Lewis, *Marbury v. Madison v. Ashcroft*, N.Y. TIMES, at A17, Feb. 24, 2003; Richard A. Posner, *Security Versus Civil Liberties*, ATLANTIC MONTHLY; Mark Tushnet, *Defending Korematsu?: Reflections on Civil Liberties in Wartime*, 2003 WIS. L. REV. 273; Thomas W. Yoo, *Presumed Disloyal: Executive Power, Judicial Deference, and the Construction of Race Before and After September 11*, 34 COLUM. HUM. RTS. L. REV. 1 (2002); Judith Resnik, Invading the Courts: We Don't Need Military 'Tribunals' to Sort Out the Guilty, Legal Times, Jan. 14, 2002, at p. 34.

[15]Some trace the idea back to Zechariah Chafee, Freedom of Speech in War Time, 32 HARV. L. REV. 932 (1919). For a sample of supporting work published subsequent to Chafee, see, e.g., RALPH BROWN, LOYALTY AND SECURITY (1958); THOMAS I. EMERSON, THE SYSTEM OF FREEDOM OF EXPRESSION (1970); PAUL L. MURPHY, THE MEANING OF FREEDOM OF SPEECH (1972); REHNQUIST, *supra* note 8; Abrams, *supra* note 14; Thomas Church, *Conspiracy Doctrine and Speech Offenses: A Reexamination of Yates v. U.S. from the Perspective of U.S. v. Spock*, 60 CORNELL L. REV. 569 (1975); Thomas I. Emerson, *Freedom of Expression in Wartime*, 116 U. PA. L. REV. 975 (1968); Thomas I. Emerson, *Freedom of*

debate whether the Court, in fact, behaves in this way; instead, the discussions are over how this came about or whether the Court should embrace a "crisis jurisprudence." As Norman Dorsen puts it:

> national security has been a graveyard for civil liberties for much of our recent history. According to this view, the questions to be answered are not whether this is true—it demonstrably is—but why we have come to this pass and how we might begin to relieve the Bill of Rights of at least some of the burden thus imposed on it.[16]

Association and Freedom of Expression, 74 YALE L. J. 1 (1964); Grossman, *supra* note 8; David Rabban, *The Emergence of Modern First Amendment Doctrine*, 50 U. CHI. L. REV. 1207 (1983); Nanette Dembitz, *Racial Discrimination and the Military Judgment: The Supreme Court's Korematsu and Endo Decisions*, 45 COLUM. L. REV. 175 (1945); Edwards, *supra* note 6; MICHAEL LINFIELD, FREEDOM UNDER FIRE: U.S. CIVIL LIBERTIES IN TIMES OF WAR (1990); Roscoe Pound, *Civil Rights During and After War*, 17 TENN. L. REV. 706 (1943); Rostow, *supra* note 6; Norman Dorsen, Here and There: Foreign Affairs and Civil Liberties, 83 Am. J. Int'l L. 840, 840 (1989); William J. Brennan, Jr. "The Quest to Develop a Jurisprudence of Civil Liberties in Times of Security Crises," (Jerusalem, Israel: Law School of Hebrew University, 1987) (available at: http://www.brennancenter.org/resources/resources_national_security.html [last accessed on February 24, 2003]); Zechariah Chafee, Free Speech in the United States (1941); Ted Finman & Stewart Macaulay, *Freedom to Dissent: The Viet Nam Protests and the Words of Public Officials*, 1966 WIS. L. REV. 632 (1966); Margaret A. Garvin, *Civil Liberties During War*, 16 CONSTIT. COMM. 691 (1999); Harry Kalven, *Ernest Freund and the First Amendment Tradition*, 40 U. CHI. L. REV. 235 (1973); IRVING HOWE & LEWIS COSER, THE AMERICAN COMMUNIST PARTY (1962); HAROLD L. NELSON, ED., FREEDOM OF THE PRESS FROM HAMILTON TO THE WARREN COURT (1967); JAMES G. RANDALL, CONSTITUTIONAL PROBLEMS UNDER LINCOLN (1951); JEROME H. SKOLNICK, THE POLITICS OF PROTEST (1969); CLINTON L. ROSSITER, THE SUPREME COURT AND THE COMMANDER-IN-CHIEF 54 (1951).

[16]Dorsen, *supra* note 15, at 840.

This is a strong claim—and one strongly endorsed by a very large fraction of the analysts who have examined the relationship between Court decisions and threats to the national security. But does this claim, sometimes called the "crisis thesis," accurately capture jurisprudence during threats to the nation's security? Do the justices, in fact, rally around the flag, supporting curtailments of rights and liberties in wartimes that they would not during periods of peace?

We raise these questions because—despite the crisis thesis's resilience—no one ever has rigorously assessed it: virtually all the evidence in its favor come from anecdotes or descriptions of a few selected Court decisions, rather than from systematic analyses of a broad class of cases. Of course, determining whether a piece of conventional wisdom can withstand rigorous scrutiny is almost always a worthwhile undertaking but it is made even more so here for, while the the crisis thesis enjoys widespread support, it continues to provide fodder for debate. A number of judges,[17] along with a handful of commentators,[18]

[17]See, e.g., ABE FORTAS, CONCERNING DISSENT AND CIVIL DISOBEDIENCE 22 (1968) ("[i]t is the courts—the independent judiciary—which have, time and again, rebuked the legislatures and executive authorities when, under stress of war, emergency, or fear of ... revolution, they have sought to suppress the rights of dissenters."); United States v. United States Dist. Court, 444 F.2d 651, 664 (6th Cir. 1971) (noting that "[i]t is the historic role of the Judiciary to see that in periods of crisis, when the challenge to constitutional freedoms is the greatest, the Constitution of the United States remains the supreme law of our land."); Liversidge v. Anderson [1942] A.C. 206, 244 (Lord Atkin, dissenting) (holding that "[i]n this country, amid the clash of arms, the laws are not silent. They may be changed, but they speak the same language in war as in peace. It has always been one of the pillars of freedom, one of the principles of liberty for which on recent authority we are now fighting, that the judges ... stand between the subject and any attempted encroachment on his liberty by the executive ... "); Aharon Barak, *A Judge on Judging: The Role of a Supreme Court in a Democracy*, 116 HARV. L. REV. 16 (2002) 149 (noting that "matters of daily life constantly test judges' ability to protect democracy, but judges meet their supreme test in situations of war and terrorism. The protection of every individual's human rights is a much more formidable duty in times of war and terrorism than in times of peace and security ... As a Justice of the Israeli Supreme Court, how should I view my role in protecting human rights given this situation? I must take human rights seriously during times of both peace and conflict.").

[18]See, e.g., Geoffrey R. Stone, *Civil Liberties in Wartime*, 28 J. SUP. CT. HIST. 215 (2003), (stating that "it is often said that the Supreme Court will not decide a case against the government ... during a period of national emergency .... In fact, however, this does not give the Court its due."); Harold Koh, *The Spirit of the Laws*, 43 HARV. INT'L L.J. 23 (2002), 189 (noting that "[i]n the days since [September 11], I have been struck by how many Americans—and how many lawyers—seem to have concluded that, somehow, the destruction of four planes and three buildings has taken us back to a state of nature in which there are no laws or rules. In fact, over the years, we have developed an elaborate system of domestic and international laws, institutions, regimes, and decision–making procedures precisely so that they will be consulted and obeyed,

challenge the idea in its entirety, suggesting that, in line with *Milligan*, the Court acts as a guardian, not a suppressor, of rights during times of war. Many more, though, question the breadth and depth of thesis, with one group claiming that its reach extends to all cases pertaining to rights and liberties[19] and another asserting that its coverage is limited to particular types of disputes, to certain kinds of crises, or even to specific classes of litigants.[20]

---

not ignored, at a time like this."); Linda Greenhouse, *Judicial Restraint: The Imperial Presidency vs. the Imperial Judiciary*, N.Y. TIMES, Sept. 8, 2002, at 3 (asserting that the judiciary has played a "restraining role" on executive authority in the war on terrorism and that "[e]ven judges whose every instinct is to defer to plausible claims of national security have recoiled"); George P. Fletcher, *War and the Constitution*, AM. PROSPECT, Jan. 1, 2002 (addressing "the fundamental question of whether the Constitution ... is different in wartime versus peacetime" and noting that "[t]he fact of 'wartime' does not change the meaning or scope of due process—either linguistically or historically."). See also Kim Lane Scheppele, *Aspirational and Aversive Constitutionalism: The Case for Studying Cross-Constitutional Influence through Negative Models*, 1 INT'L J. CONST. L. 296 (2003); Mary Dudziak, *The Supreme Court and Racial Equality During World War II*, 1996 J. SUP. CT. HIST. 35.

[19]See, e.g., Emerson, *supra* note 15. See also Resnik, *supra* note 14 (noting that "in times of war, courts often do not protect against incursions on civil liberties").

[20]Along these lines come the writings of Justices Brennan and Rehnquist, both of which place emphasis on the Court's decision in Korematsu, 323 U.S. 214 (1944) and other cases that flow directly from the war or other emergency at hand. See Rehnquist, *supra* note 8; Brennan, *supra* note 15. For commentary suggesting that the thesis is not so much about Court treatment of alleged infringements of rights and liberties made by all types of parties but rather about deference strictly in cases when the U.S. government is a party, see, e.g., Rossiter, *supra* note 15, at 54; Edward S. Greenberg, "Will Things Ever Be the Same? The 'War on Terrorism' and the Transformation of American Government, in AMERICAN GOVERNMENT IN A CHANGED WORLD 23 (Dresang et al. eds., 2003); Lobel, *supra* note 5; John C. Yoo, "The Continuation of Politics by Other Means: The Original Understanding of War Powers," 84 CALIF. L. REV. 167 (1996).

Debate also exists over the duration of the crisis effect. Some suggest that the justices suppress rights only while a war is ongoing, while others argue that the curtailments linger well after the threat has subsided.[21]

We understand why these debates continue, as well as why, to date, no one has attempted a large-scale systematic study aimed at addressing the many questions the crisis thesis raises. It has only been in the last decade or so that scholars have developed the high-quality data and statistical tools requisite to conduct such a study or, more to the point, to conduct it in a sophisticated and convincing fashion. But with those data and tools now in place, the time has now come to put the crisis thesis to the test.[22]

---

[21]Compare Gross, *supra* note 5 and Tushnet, *supra* note 14. See also Eric Posner and Adrian Vermeule, *Accommodating Emergencies*, STANFORD L. REV. (2004), forthcoming (on file with the authors).

[22]Of course, the timing could hardly be more auspicious: In the aftermath of September 11th members of legal, policy, and journalistic communities—in scores of scholarly articles and books, in numerous media reports, and in symposium after symposium—have engaged in a serious debate over the extent role of the federal judiciary in the war on terrorism. See, e.g., Barak, *supra* note 17; James E. Coleman & Barry Sullivan, *Enduring and Empowering: The Bill of Rights in the Third Millennium*, 65 L. & CONTEMP. PROBS. 1 (2002); Viet D. Dinh, *Freedom and Security After September 11*, 25 HARV. J.L. & PUB. POL'Y (2002); Gross, *supra* note 5; Heymann, *supra* note 14; Katyal & Tribe, *supra* note 14; Lobel, *supra* note 5; Yoo, *supra* note 14; THE WAR ON OUR FREEDOMS: CIVIL LIBERTIES IN AN AGE OF TERRORISM (Richard C. Leone & Gregory Anrig Jr. eds., 2003); Greenhouse, *supra* note 18; Lewis, *supra* note 14; Levinson, *supra* note 14; *Symposium, Civil Liberties in a Time of Terror*, 2003 WIS. L. REV. 253 (2003); *Civil Liberties in Times of War*, 28 textscJ. Sup. Ct. Hist. (2003): "Are Constitutions Necessary?," Conference at Columbia University, 2003; "Responding to Terrorism," Conference at the University of Pennsylvania, 2001; "Journalism and Terrorism: How the War on Terrorism Has Affected the Practice of Journalism," Conference at the Missouri School of Journalism, 2002; "Homeland Insecurity: Civil Liberties, Repression, and Citizenship,"

At the very least, this is the task we undertake in this article. Using the best data available on the causes and outcomes of every civil rights and liberties case decided by the Supreme Court since 1941 and employing methods chosen and tuned especially for this problem we explore systematically the Court's decisions during periods when the country is in "emergency and peril" or in relative peace.[23] Our findings, so that there will be no mystery about them, provide the first systematic support for the existence of a crisis jurisprudence: *The justices are, in fact, significantly more likely to curtail rights and liberties during times of war and other international threats.*[24] However, contrary to what every proponent of the crisis thesis has so far suggested, war affects cases entirely unrelated to

Conference at Smith College, 2003; "Infinite Respect, Enduring Dignity: Voices and Visions of the September Attacks," Conference at the University of Iowa, 2002; "Terrorism and Human Rights," Conference at the University of St. Andrews, 2003; "Justice or Just US: A Conference to Address the Erosion of Civil Liberties, American Friends Service Committee, 2003; "Civil Liberties & the War on Terrorism," Conference at Suffolk University Law School, 2003; "At War with Civil Rights and Civil Liberties," Conference at Florida International University, 2003.

[23]This investigation is in line with a strand of international relations scholarship that examines the effects of international relations on domestic politics, often termed "the second-image reversed." See Peter Gourevitch, *The Second Image Reversed: International Sources of Domestic Politics*, 32 INT'L ORG. 881 (1978); INTERNATIONALIZATION AND DOMESTIC POLITICS (Robert O. Keohane & Helen V. Milner eds., 1996) The original three images stem from KENNETH N. WALTZ, MAN, THE STATE, AND WAR (1954).

[24]This finding is robust to a host of other factors that analysts suggest affect Supreme Court decisions, such as long-term changes in legal culture, positions taken by the lower courts, public exposure of the cases, and judicial ideology. Moreover, it does not appear that well-known selection effects of litigation are driving the effect. See, e.g., George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1 (1984).

the war, and there is little evidence that war affects cases related to the war. This is consistent with the notion that war-related cases present the greatest threat of encroachment on the judiciary, and hence, for those cases, the court "rallies 'round the court." Yet for cases unrelated to the war, the court balances security and liberty interests as scholars have commonly conjectured.

3.2 Data

We focus on the outcomes of cases in which parties claimed a deprivation of their rights or liberties and that the Supreme Court resolved on the merits, whether in times of international urgency or not, over the last six decades (1941-2001 terms).

3.2.1 The Cases

The existence of Harold J. Spaeth's U.S. Supreme Court Database makes amassing data on Court decisions straightforward enough.[25] This database, which many scholars have used to study law and judicial politics,[26] contains information on over two hundred

[25] The database and the documentation necessary to use it are available at: http://www.polisci.msu.edu/pljp/supremecourt.html (last accessed on November 10, 2003).

[26] For recent examples, see Frank B. Cross and Blake J. Nelson, *Strategic Institutional Effects on Supreme Court Decisionmaking*, 95 NW. U.L. REV. 1437 (2001); Youngsik Lim, *An Empirical Analysis of Supreme Court Justices' Decision Making,* 29 J. LEGAL STUD. 721 (2000); JEFFREY A. SEGAL & HAROLD J. SPAETH, THE SUPREME COURT AND THE ATTITUDINAL MODEL REVISITED (2002); LEE EPSTEIN, ET AL., THE SUPREME COURT COMPENDIUM (2003) ; Ruth Colker & Kevin M. Scott, Dissing States?:

attributes of Court decisions—including whether the justices ruled in favor of or against individuals claiming a violation of their civil rights or liberties—in all cases decided by the Court with an opinion since the 1953 term.

Spaeth classifies civil rights and liberties cases into one of six broad categories: Criminal procedure, civil rights, First Amendment, due process, privacy, and attorney rights. For example, *Korematsu*[27] is coded as a civil rights case, *Ex Parte Quirin*,[28] as involving criminal procedure, and *Dennis v. United States*,[29] another suit prominent in the literature on the Court's crisis jurisprudence, as a First Amendment dispute. The specification of whether the Court resolved a dispute in favor of the party claiming a deprivation of his or her rights (that is, in the "liberal" direction) "comports with conventional usage."[30] In issues pertaining to criminal procedure, civil rights, First Amendment, due process, privacy, and attorneys, this means that a case is liberal if the outcome favored the person accused or convicted of crime, or denied a jury trial; the civil liberties or civil rights claimant; the indigent; Native American claims; affirmative action; neutrality in religion cases; the choice stance in abortion; the underdog; claims against the government in the context of due pro-

Invalidation of State Action During the Rehnquist Era, 88 VA. L. REV. 1301 (2002); Keith E. Whittington, *Taking What they Give Us: Explaining the Court's Federalism Offensive,* 51 DUKE L.J. 477; Ernest A. Young, *Judicial Activism and Conservative Politics* 73 U. COLO. L. REV. 1139

[27]323 U.S. 214 (1944).

[28]317 U.S. 1 (1942).

[29]341 U.S. 494 (1951).

[30]Harold J. Spaeth, Documentation for the United States Supreme Court Judicial Database, 1953-2000 Terms, at 54. Available at: http://www.polisci.msu.edu/pljp/supremecourt.html (last accessed on November 10, 2003)..

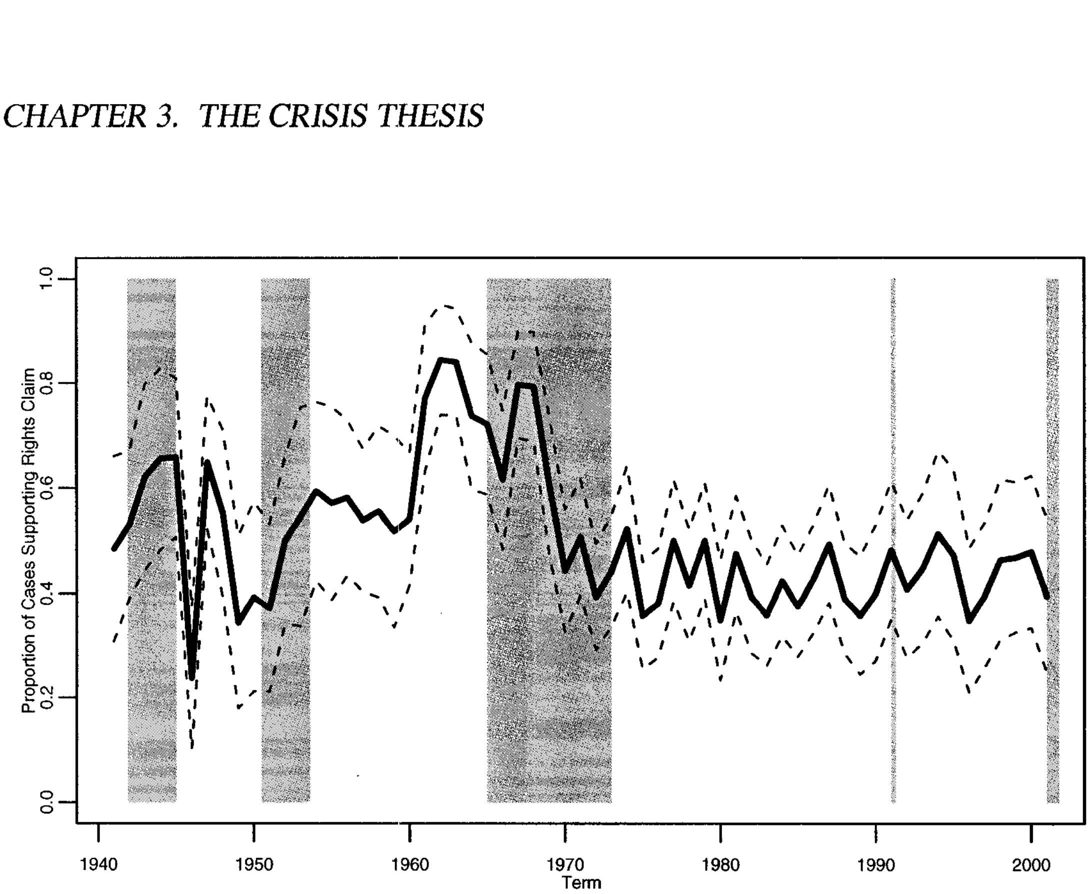


Figure 3.1: The proportion of U.S. Supreme Court decisions supporting rights, liberties, or justice claims, 1953-2001 terms. The line depicts the proportion of support. The gray shading depicts terms during which the Court heard disputes during a war period.[32]

cess; attorney rights; and, disclosure in Freedom of Information Act and related federal statutes, except for employment and student records.[31] Figure 3.1 plots out the proportion of cases decided liberally across our period of interest.

Following Spaeth's coding rules and with his guidance, we backdated the dataset to include the 1941-1952 terms, and updated it to include the 2001-02 term.[33] With these

[31] Id.

[33] Using Spaeth's terms, the analu (the unit of analysis) for this study=0 (case citation) and dec_type (the type of decision)=1 or 7 (cases that were orally argued and decided with a signed opinion). Civil rights, liberties, and justice cases are criminal procedure, civil rights, First Amendment, due process, privacy, and attorneys (values 1-6 of Spaeth's value variable).

additional data we were able to incorporate into our analyses terms coinciding with World War II and the Korean War, as well as the recent military conflict in Afghanistan.

3.2.2 The Crises

Since we are interested in assessing whether a "crisis" affects the Supreme Court and just how expansive that effect (if it exists at all) might be, defining what constitutes a "crisis" in a transparent manner is crucial for our study. In the absence of other large-scale empirical assessments, however, determining the presence or absence of a "crisis" presents a something of a challenge.[34] Is a "crisis" solely a "constitutional war" (that is, a war fought pursuant to a congressional declaration of war),[35] or is that term broad enough to encompass a long-term military effort in the form of a war or even more temporary states of

[34]Even the War Powers Resolution, Pub. L. No. 93-148, 87 Stat. 555 (codified at 50 U.S.C. §§1541-1548 (1976)), provides no definition of what constitutes "hostilities." See Louis Henkin, Foreign Affairs and the Constitution, 1987 Foreign Aff. (1987) (noting that "[a]bove all, the [War Powers R]esolution suffers gravely from a lack of any definition of 'hostilities'").

[35]See U.S. CONST. art. I, §8, cl. 11. Strictly speaking, this definition of constitutional war would exclude the Korean War (1950-53), The Vietnam War (1965-73), the Gulf War (1991), the War in Afghanistan (2001-02), and the Iraq War (2003). Then again, we might consider intermediate expressions of support for war, such as the Gulf of Tonkin Resolution of 1964 used to justify Executive power during the Vietnam War, and the Congressional Resolution expressing support for the deployment of troops in the Gulf War. This approach would entail weighing the impact of the War Powers Resolution, Pub. L. No. 93-148, 87 Stat. 555 (codified at 50 U.S.C. §§1541-1548 (1976)), on the definition of a constitutional war.

conflict? The literature's emphasis on cases such as *Milligan*[36] and *Korematsu*[37] suggests a long-term military effort but an exclusive focus on "war," whether formally declared or otherwise, would eliminate what many specialists in international law view as major U.S. "crises," such as the Berlin Blockade—and, of course, September 11th . Should we include these, as well as several others (e.g., the Cuban Missile Crisis, Iran-Hostage Crisis) in our working definition of a "crisis"?

Because it is difficult, if not impossible, to answer these questions with a high degree of precision,[38] we chose to develop three explicit definitions of crisis.[39] The first two are rather obvious: the absence or presence of war[40] (with wars defined as World War II, the

[36] 71 U.S. 2 (1866).

[37] 323 U.S. 214 (1944).

[38] See, e.g., Gross, *supra* note 5, at 1089-96 (noting that five reasons for a blurring distinction between emergency and normalcy). One strand of classical realist international relations theory views war as merely a continuation of politics by force, so a clear definition of crises may simply not be possible. See, e.g., CARL VON CLAUSEWITZ, ON WAR 78, 87 (Peter Paret ed. trans., 1976) (asserting that "war is never an isolated act" and that war is "a continuation of political intercourse, carried on with other means"). But cf. JOHN KEEGAN, A HISTORY OF WARFARE 3 (1993) (asserting that "war is not the continuation of policy by other means" and proposing a cultural theory of war). Even Keegan's framework, however, does help to clarify a definition of crises for our purpose here.

[39] We do not, however, explicitly consider the notion of a strict constitutional war, since this would exclude, in light of the time frame of our study, all wars save World War II from investigation. Nonetheless, since our analysis estimates causal effects for each individual war case and includes restrictions on terms, we could easily compare the estimated causal effects specific to each war, which would in turn enable us to determine whether the only constitutional war, World War II, exhibited distinct effects on Supreme Court decision making.

[40] Following the vast majority of literature in this area, "war periods do not include the cold war, the war

Korean, Vietnam, and Gulf wars, and the recent war in Afghanistan)[41] and the absence or presence of wars, plus four other international conflicts that specialists have labeled as "major" (the Berlin Blockade, the Cuban Missile and the Iran-Hostage crises, and September 11th). The third measure—the presence or absence of a "rally effect" in the form of 10-point (or greater) surge in presidential popularity caused by an international event—may be less obvious but it well taps a "crisis" as social scientists have employed that term.[42] After all, it is during those periods when the public rallies around a President (typically the catalyst for efforts to suppress rights and liberties) that we might expect the Court to do the same.[43]

---

drugs under the first Bush administration, or the war on poverty under the Johnson administration (but cf. Edwards, *supra* note 6 [considering the Cold War and the war on drugs]). At the same time, it is possible that our third measure of a crisis, a rally effect, may capture periods of particular intensity during the cold war.

[41]Since it remains unclear whether we should code the entire period after September 11th as war or not, we pursue both coding schemes here. See, e.g., Tushnet, *supra* note 14, at 279, 279-80 (arguing that "we ought not think of [the war on terrorism] as a war in the sense that World War II was a war" and that it is "a condition rather than a more traditional war"); but cf. Downs & Kinnunen, *supra* note 13, at 399-402 (arguing that the acts of terrorism "unmistakably bear the characteristics of war").

[42]Beginning with World War II, rally effects have occurred 16 times, 14 of which (not surprisingly) involved international events. (The two that did not centered on the Clinton "scandal" and impeachment.) In each case, the President's popularity jumped at least 10 points, with George W. Bush receiving the biggest boost (35 points); and the surges endured anywhere between 5 and 41 weeks. For this study we include only the 14 international events. For a complete list, navigate to: www.gallup.com/poll/Releases/Pr010918.asp) (last accessed on January 23, 2003)(on file with the authors).

[43]In research seeking to identify the public opinion consequences of crises, the key threat to inference is endogenously defining the explanatory variable on the basis of shifts in the dependent variable (usually presi-

Gathering data on the three measures of crisis requires knowing when each crisis began and each ended. Since that information is readily available,[44] we were left with only one task: determining whether the Supreme Court made its decision during a crisis period (for each measure of crisis) or not. Two possibilities presented themselves: pegging the existence (or lack thereof) of the crisis to the date the Court handed down its decision or to the date it heard oral arguments in the case. We opted for the latter. That is because we are studying the effect of a crisis on the direction of the Court's decision (for or against the rights, liberties, or justice claim)—which typically is determined by an initial vote taken within a few days of oral arguments in a case[45]—rather than the rationale in the opinion—which usually is determined during a bargaining period that occurs between oral argument and the printing of the final version of the opinion.[46] Of course, individual justices do

dential approval). As such, all lists of "rally points" are by definition problematic for that sort of research. For our study, if we take the combination of the crisis and any possible public opinion changes as our explanatory variable, we have no such definitional problems.

[44]The dates for wars are: World War II: 12/7/41-8/14/45; Korea: 6/27/50-7/27/53; Vietnam: 2/7/65-1/27/73; Gulf: 1/16/91-4/11/91; Afghanistan: 10/7/01-3/14/02. With the exception of September 11th, the dates for the major crises are from the International Crisis Behavior Project (ICPSR Study 9286) (available at: http://www.icpsr.umich.edu/); September 11 began on 9/11/01 and continues through the 2001 term, the last in our database. Dates for rally events are available at: www.gallup.com/poll/Releases/Pr010918.asp (last accessed on January 23, 2003)(on file with the authors).

[45]See LEE EPSTEIN & JACK KNIGHT, THE CHOICES JUSTICE MAKE (1998) for information on the Court's internal decision-making procedures.

[46]To see the logic behind our choice, consider September 11 and assume (even though the Court's term does not begin until October) that the Court heard arguments in, say, a First Amendment case on September 1, 2001, took its initial vote on September 2, and handed down its decision on September 13. Had we pegged

change their votes between the conference following oral arguments and publication of the final decision. Yet, those vote shifts rarely produce alterations in the direction (*i.e.,* for or against the claim) of the Court's decision.[47]

To assess whether war impacts only civil rights and liberties cases directly related to the war, we also coded every case as being war-related or not from the fact-pattern. Cases are coded as war-related if the controversy was a direct result of the war, such as draft cases, war protest cases, takings for military purposes, deportation and relocation cases of nationals from war enemies, and court martials for activity occuring in a war zone.

Table 3.1 documents the results of these research decisions, summarizing information on the 3,345 civil rights, liberties, and justice cases in which the Court heard arguments between the 1941 and 2001 terms, along with the three measures of crisis. As we can observe, decision making in times of international urgency is not the norm for the Court. Into none of our measures do a majority of the cases fall; and combining the indicators does not change the picture. Overall, the justices decided 32 percent (n=1,067) of the

---

the existence of a crisis to September 13, rather than to September 2, we would have coded this as a decision made during a crisis despite the fact that September 11 had yet to occur at the time the Court took a vote in the case. And while this is an initial vote (subject to change), alterations in Court disposition (e.g., from an affirm to a reverse), as we note in the text, are quite rare.

[47]Lee Epstein & Jack Knight, "Documenting Strategic Interaction on the U.S. Supreme Court" (paper presented at the annual meeting of the American Political Science Association, Chicago, IL, 1995); Saul Brenner, *Fluidity on the United States Supreme Court: A Reexamination*, in AMERICAN COURT SYSTEMS (Sheldon Goldman & Austin Sarat ed. 1989). For some exceptions to this general rule, see EDWARD LAZARUS, CLOSED CHAMBERS (1998).

Measure of Crisis	% of Cases
War	23
War, Plus Major Conflicts	28
Rally Effect	9
Related to War	4
(Total Cases)	(3,345)

Table 3.1: Crises and U.S. Supreme Court Cases Involving Rights, Liberties, and Justice, 1941-2001 Terms[48]

3,345 cases while a war, major conflict, or rally event was in place. On the other hand, crisis decision making is hardly a rare event. Surely this is true of the first two measures—war and war, plus major conflicts—with roughly a quarter of the cases occurring while they were ongoing. It also holds for rally effects. The percentage of 9 may be small but the number of cases (n=292) is sufficiently large to enable meaningful analysis and, thus, to explore the potential effect of rallies on the Court.

3.2.3 Confounding Factors

The most prominent confounding factors are the *characteristics of cases* that come to the Court during times of war and peace[49] and the *political composition* of the Court

[48]The data on Supreme Court cases come from the U.S. Supreme Court Judicial Database, backdated by the authors, with analu=0, dec_type=1 or 7, and values=1-6 (see *supra* notes 25 and 33). "War" includes World War II, and the Korean, Vietnam, Gulf, and Afghan Wars; (see *supra* note 44 for more details). "War, Plus Major Conflicts" includes Wars and the Berlin Blockade, Cuban Missile, and Iran-Hostage (see *supra* note 44 for more details). Rally Effects are periods during which the President's popularity rises by 10 points or more as a result of an international event (see *supra* notes 42 and 44).

[49]See e.g., Yoo, *supra* note 14; Robert S. Chang, *The Legal, Moral, and Constitutional Issues Involving Diversity,* 66 ALB. L. REV. 349 (2003); David Richman, *Prosecutors and their Agents, Agents and their*

hearing those cases.[50] We also consider several other confounding factors—including case salience, lower court rulings, and time effects—to which, according to the extant literature, any large-scale study of judicial decision making in times of crisis ought be attentive.

The Characteristics of Cases

A primary potential confounding factor, if we take seriously several contemporary essays on the effect of war on judicial decisions[51] relates to the well-known selection effects of litigation,[52]—presenting itself here in the form of distinctions in the characteristics of cases that come to the Court during times of crisis and periods of tranquility. Specifically, several scholars have argued that when the nation is at war, the government becomes increasingly bent on curtailing individual rights. Accordingly, it undertakes prosecutions in which the facts are so "severe" (or "extreme") that even a sympathetic (that is, right-of-center) Court would have difficulty simultaneously ruling in the government's favor and following extant legal principles.[53]

Consider the Fourth Amendment's guarantee against unreasonable searches and seizures—an oft-cited exemplar of constitutional provision that governments attempt to skirt during war times.[54] From various statistical analyses, we know a great deal about the particular

---

Prosecutors, 103 COLUM. L. REV. 749 (2003).

[50]See *supra* note 26.

[51]See *infra* note 49.

[52]See Priest & Klein, *supra* note 24.

[53]For supporting literature, see *supra* note 49.

[54]See, e.g., Dorsen, *supra* note 15; Heymann, *supra* note 14; Michael P. O'Connor & Rumann, *Emergency and Anti-Terrorist Power: Going, Going, Gone: Sealing the Fate of the Fourth Amendment*, 26 FORDHAM

features of cases that lead the Court to interpret this guarantee in a way that favors defendants (typically when it strikes down the challenged search) or the government (typically when it upholds the search);[55] from those same studies, we also have a reasonably good sense of how the various features affect the probability of the Court striking down or upholding a search.[56]

Figure 3.2 lists those features—eleven critical facts pertaining to search and seizure cases, along with the impact empirical investigations have shown them to have on the Court (when compared to a 0.5 baseline probability of a decision favoring the defendant).[57] Consider, for example, the three types of searches that can occur when law enforcement officials make an arrest: searches incident to a valid arrest, searches after lawful arrests,

---

INTERNATIONAL LAW JOURNAL 1234 (2003); William J. Stuntz, *Local Policing After the Terror,* 111 YALE L. J. 2137 (2002).

[55]See, e.g., Jeffrey A. Segal, Predicting Supreme Court Decisions Probabilistically: The Search and Seizure Cases, 78 Am. Pol. Sci. Rev. 891 (1984); Segal & Spaeth, *supra* note 26.

[56]See Segal & Spaeth, *supra* note 26, at 318.

[57]Note that since impact calculations are not additive in a logistic probability model, these impact findings should not be interpreted as constant or additive. The impact depends on which other facts are present, the predicted probability is of course always bounded between 0 and 1. See, e.g., GARY KING, UNIFIYING POLITICAL METHODOLOGY: THE LIKELIHOOD THEORY OF STATISTICAL INFERENCE 97-110 (1989); Gary King, Michael Tomz & Jason Wittenberg, *Making the Most of Statistical Analyses: Improving Interpretation and Presentation*, 44 AM. J. POL. SCI. 341, 355 (2000).

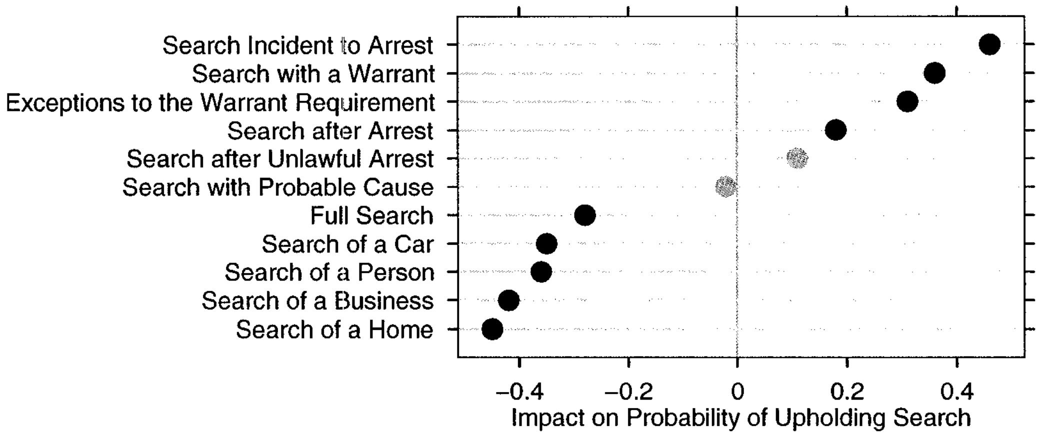


Figure 3.2: Facts relevant to the Supreme Court's adjudication of Fourth Amendment Cases, and the impact of the facts on the predicted probability of the Court upholding a search. Grey circles indicate statistically insignificant impacts.[60]

and searches after unlawful arrests.[58] In light of existing legal principles,[59] it is hardly a surprise to find, as the figure shows, that those searches incident to arrest are the most likely to receive Court validation (such a search has a 0.46 greater predicted probability of being upheld than a search that was not incident to an arrest); searches after lawful arrests receive less favorable treatment from the justices; and searches after unlawful arrests are the least likely of all "arrest" searches to be upheld.

[58]In determining whether an arrest is valid or unlawful, scholars typically rely the holding of the lower courts. See Segal & Spaeth, *supra* note 26.

[59]For a review of those principles, see LLOYD WEINREB, WEINREB'S CRIMINAL PROCESS, 2D: PART ONE—INVESTIGATION (1998).

[60]This is a graphical depiction of data in Segal & Spaeth, *supra* note 26, at 318, which contains the statistical model from which they derived these predictions. The probabilities for Person, Home, Car, and Business are all compared to a baseline where the defendant does not have a property interest. The probability for Full Search is compared to the baseline of a limited intrusion such as a stop-and-frisk.

Now suppose it is in fact the case that the government acts in a more repressive fashion during times of war. Such "overzealous" prosecution might in turn generate more "extreme" cases—for example, a disproportionate number of cases that involve searches incidents to unlawful arrests, which the justices, on average, are less likely to uphold than searches incident to a lawful arrest (see Figure 3.2). As the Court begins to adjudicate these "unlawful arrest" cases,[61] rather than, say, "incident to arrest" cases,[62] we might expect to find more and more holdings in favor of the defendant, even if the justices did not alter existing legal doctrine whatsoever. To put it more generally, the facts, in response to overzealous government efforts, may move sufficiently far to the right during times of war as to compel the Court—in face of extant legal principles—to articulate a position that favors the defendants.

To investigate this possibility, we computed the on average degree of severity (or extremeness) in the facts of Fourth Amendment cases for each term in our database,[63] hypothesizing that this average should increase during times of war if, in fact, the government is overzealous in its prosecution efforts.[64] We then ran a series of time-series models to assess the hypothesis, with Table 3.2 depicting one specification of an ordinary least squares

[61] This holds for any suits in which the facts are such that the justices typically strike down the search.

[62] This holds for any others in which the facts are such that the justices typically uphold the search.

[63] Figure 3.2 illustrates the facts we included.

[64] We also estimated a series of models that assessed the effect of international crises on the severity of case facts at the case (rather term) level. After controlling for the political ideology of the Court—a crucial variable, we cannot uncover statistically significant differences between cases decided during times of war and peace. In other words, the results (available from the authors) confirm the analyses presented in Table 3.2.

Explanatory Variables	Average Per Year Severity of Search & Seizure Cases Dynamic Specification— Partial Adjustment		
War	-.330 (.340)	—	—
War, & Major Conflicts	—	-.146 (.333)	—
Rally Effect	—	—	.460 (.359)
Average Severity$_{t-1}$	.360 (1.88)	.415 (.188)*	.476 (.156)*
Constant	-1.26 (.393)*	-1.18 (.395)*	-1.17 (.367)*
Adj. R^2	.20	.17	.22
N	32	32	32

Table 3.2: Time series assessments of the effect of international crises on the severity of case characteristics. * indicates $p \leq .01$.[65]

with a variable representing the lag of "case severity" on the right-hand side.

We can draw several conclusions from this table[66] but only one is relevant here. All the models depicted in the table, along with every plausible alternative time-series specification we tested, returned the same substantive result: a lack of any detectable impact of the crises variables on the extremity of cases on the Court's docket.[67] In other words, the facts

[65]The database used to conduct this analysis is available on our web site. For the crisis measures, War includes World War II, and the Korean, Vietnam, Gulf, and Afghan Wars; War, Plus Major Conflicts includes War and the Berlin Blockade, Cuban Missile, and Iran-Hostage; Rally Effects are periods during which the President's popularity rises by 10 points or more as a result of an international event ((see *supra* notes 42 and 44 for more details). We calculated the severity levels of searches from the same coefficients used to calculate the probabilities in Figure 3.2. Because negative values are associated with more extreme searches, the case-severity account predicts significantly negative coefficients on our crisis measures.

[66]Another conclusion is that the case facts heard by the Court are generally not random, as judged by the lagged severity coefficients.

[67]We did obtain a set of significant coefficients by running OLS and excluding lagged severity. But the

presented in these cases simply did not vary; they became no more (or less) "severe" (or "extreme") in times of war, suggesting, in turn, that it was no more (or less) difficult for the Court to decide the suits for (or against) the government.

The Political Composition of the Court

While we find little support for the existence of a systematic correlation between any of the facts and any of our crisis measures, the political composition of the Supreme Court appears to differ substantially in war and peace. Specifically, owing to the confluence of several historical phenomenon—the dominance of the Democratic party during long periods of the 20th century and the resulting appointment of relatively liberal justices,[68] coupled with a greater frequency of wars during those periods—Courts deciding cases during times of crisis were composed of considerably more left-of-center ("liberal") justices than those deciding cases during times of peace.

While social scientists and legal academics have proposed several operational approaches

---

Durbin-Watson statistics show clear signs of autocorrelation ($DW_{2,35}$ = 1.12, 1.22, and 0.89 respectively), which means that the reported standard errors are biased downward. Every analysis we conducted that controls for autocorrelation finds no significant relationship. Our web site houses the complete set of findings.

[68]See *infra* Figure 3.3.

to measure ideology,[69] we rely here on a particularly powerful and pervasive one:[70] the ide-

[69]For a review of many of these measures, see Lee Epstein & Carol Mershon, *Measuring Political Preferences*, 40 AM. J. POL. SCI. 260 (1996). The simplest one is the use of party affiliation, but this fails to capture nuances along the ideological spectrum, which are so vital to controlling for policy preferences. See, e.g., Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, at 74-75 (2002) (discussing the relation between theory and operationalization of policy preferences). Accordingly, political scientists have proposed much more sophisticated models using item response theory. See, e.g., Simon Jackman, *Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference and Model Checking*, 9 POL. ANAL. 227 (2002); Simon Jackman, *Estimation and Inference are 'Missing Data' Problems: Unifying Social Science Statistics via Bayesian Simulation*, 8 POL. ANAL. 307 (2000). For a particularly sophisticated estimation of Supreme Court ideal points using Markov Chain Monte Carlo methods, see Andrew D. Martin & Kevin M. Quinn, *Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999*, 10 POL. ANAL. 134 (2002); and Andrew D. Martin & Kevin M. Quinn, Patterns of Supreme Court Decision-Making, 1937-2000, Working Paper No. 26, Center for Statistics and the Social Sciences (2002) (on file with authors). This measure is not well suited for our purposes, however, since Martin & Quinn estimate the ideal points by using actual votes, thus endogenously defining them with respect to our main causal variable of interest. But an alternative Martin & Quinn measure—one that used votes from all Supreme Court cases *except* those involving rights civil rights and liberties could provide an alternative to the Segal & Cover scores, assuming that (a) war does not affect non-civil rights and liberties cases and (b) ideal points are non-separable and therefore correlated among issue areas. We conducted analyses using these alternative measures (graciously provided by Martin & Quinn) to analyze the robustness of our results.

[70]Indeed, these score figure prominently into many studies of judicial decisions. See, e.g., Lee Epstein, Jack Knight, & Andrew D. Martin, *The Supreme Court as a Strategic National Policy Maker*, 50 EMORY L.J. 101 (2001); William N. Eskridge, Jr., *Reneging on History? Playing the Court/Congress/President Civil Rights Game,* 79 CALIF. L. REV. 613 (1991); Barry Friedman and Anna L. Harvey, *Electing the Supreme*

ology scores that Jeffrey A. Segal and Albert Cover have assigned to each justice serving since the 1930s.[71]

Accounting for the prominence of the Segal & Cover scores is not altogether difficult: They have proven to be highly accurate predictors of judicial votes, especially in the areas (civil liberties and rights)[72] and in the terms (1941-2001)[73] under analysis here. They also are exogenous to the vote, since Segal and Cover developed them not from examining the decisions reached by justices but rather by analyzing newspaper editorials written between the time of justices' nominations to the Court and their confirmations.[74]

Court, 78 IND. L.J. 123 (2003). One commentator, though, has criticized the Segal & Cover scores as "reflect[ing] rather general opinions about the political orientation of a justice" and not being a "good guide[] to the views of justices in specific areas of constitutional controversy." STEPHEN M. GRIFFIN, AMERICAN CONSTITUTIONALISM, 32-33 (1996). Yet capturing the general "political orientation" is precisely the purpose of the Segal & Cover scores. As we employ them, they are not meant to provide predictive power for one particular case, but rather to control for long-term changes in judicial ideology of the court. Moreover, we use them to study decisions in the areas of civil rights and liberties, where they work well as general predictors of trends in decision making. See, e.g., Epstein & Mershon, *supra* note 69.

[71] Jeffrey A. Segal & Albert D. Cover, *Ideological Values and the Votes of U.S. Supreme Court Justices,* 83 AM. POL. SCI. REV. 557 (1989).

[72] Epstein & Mershon, *supra* note 69.

[73] Jeffrey A. Segal et al., *Ideological Values and the Votes of U.S. Supreme Court Justices Revisited,* 57 J. POL. 812 (1995).

[74] Specifically, as Segal and Cover tell it:

> We trained three students to code each paragraph [in the editorial] for political ideology. Paragraphs were coded as liberal, moderate, conservative, or not applicable. Liberal statements include (but are not limited to) those ascribing support for the rights of defendants in criminal cases, women and racial minorities in equality cases, and the individual against the government in privacy and First Amendment cases. Conservative statements are those with an opposite direction. Moderate statements include those explicitly ascribe moderation to the nominees or those that ascribe both liberal and conservative values.

From these analyses, Segal & Cover devised a scale of judicial political ideology, which ranges from -1 (the most conservative) to 0 (moderate) to +1 (the most liberal),[75] such that each justice obtains a score within this range. For example, William J. Brennan received a score +1.0; Scalia's is -1.0, while O'Connor's is a more moderate -.17).[76] (For subsequent analysis, we rescaled the score to take a minimum value of 0.) But, because we are interested in examining the political composition of the Court as a whole, rather than the policy preferences of particular justices, we used the Segal & Cover scores to calculate the ideology of the median justice serving on each Court,[77] for each year in our analysis. So doing is consistent with public choice and jurisprudential theories emphasizing the importance of the swing vote[78]—not to mention with contemporary commentary, which often stresses the

Segal & Cover, *supra* note 71, at 559. They arrived at their measure by subtracting the fraction of paragraphs coded conservative from the fraction of paragraphs coded liberal and dividing by the total number of paragraphs coded liberal, conservative, and moderate. Id.

The fact that this measure is exogenous to the judicial vote is crucial to the analysis, as noted in *supra* note 69. Since the entire question of this study is the causal effect of war on the outcome of opinions, we cannot very well use outcomes to derive a control variable.

[75]In subsequent analyses in which we employ propensity score matching (see Part 3.3), we rescale this so that the ordering of the Segal & Cover score is invariant to a square transformation.

[76]For a list of the Segal and Cover scores for all justices, see Epstein et al., *supra* note 26, Table 6-1.

[77]The median is the middle justice, such that four justices are more liberal and four are more conservative.

[78]See, e.g., Jeffrey A. Segal, *Separation-of-Powers Games in the Positive Theory of Congress and Courts*, 91 AM. POL. SCI. REV. 28 (1997); Epstein and Mershon, *supra* note 69; R. Randall Kelso and Charles D. Kelso, *Swing Votes on the Current Supreme Court: The Joint Opinion in Casey and Its Progeny*, 29 PEPP. L. REV. 637 (2002); Mario Bergara, Barak Richman & Pablo T. Spiller, *Modeling Supreme Court Preferences in a Strategic Context*, 28 LEG. STUD. Q. 247 (2003). The role of the pivotal justice is deeply rooted in the

critical role Justice O'Connor (and, to a lesser extent, Kennedy) has played on the Court by casting key votes in many consequential cases.[79]

Figure 3.3 depicts these Court "swings": the median's political ideology, computed on the basis of the Segal & Cover scores, for the 1941-2003 terms. As we can observe, the data are well in line with commonly held intuitions about particular Court eras.[80] Note, for example, the increase in liberalism during the Warren Court years (1953-1968 terms) and the decrease that occurs thereafter as more and more justices appointed by Republican Presidents Richard M. Nixon, Ronald Reagan, and George H.W. Bush ascended to the bench.

Note too the gray shading in Figure 3.3, indicating terms during which wars were ongoing. For the vast majority, as we can see, the median was quite liberal, thereby providing the first glimpse of support for our claim that the political composition of the Court has

theory of the median voter. See generally ANTHONY DOWNS, AN ECONOMIC THEORY OF DEMOCRACY (1957); Harold Hotelling, *Stability in Competition*, 39 ECON. J. 41 (1929).

[79]See, e.g., Editorial, *A Moderate Term on the Court,* N.Y. Times, at 4:12, June 29, 2003 (noting Justice O'Connor's status as the "court's critical swing vote"); Associated Press, *Affirmative Action Case Puts Judges in Spotlight,* April 1, 2003 (describing Justices O'Connor and Kennedy as the "perennial swing voters"); Charles Lane, *Supreme Court: On the Sidelines, for Now,* Wash. Post, at A5, Sep. 30, 2001 (describing Justice O'Connor as the "perennial swing voter").

[80]For ideological characterizations of particular Courts, see, e.g., Thomas W. Merrill, *Childress Lecture: The Making of the Second Rehnquist Court: A Preliminary Analysis*, 47 ST. LOUIS U. L.J. 569 (2003); HOWARD GILLMAN, THE VOTES THAT COUNTED (2001); William N. Eskridge, Jr., *Overriding Supreme Court Statutory Interpretation Decisions*, 101 YALE L. J. 331 (1991).

[81]The Segal & Cover scores are available in Epstein et al., *supra* note 26, Table 3-12.

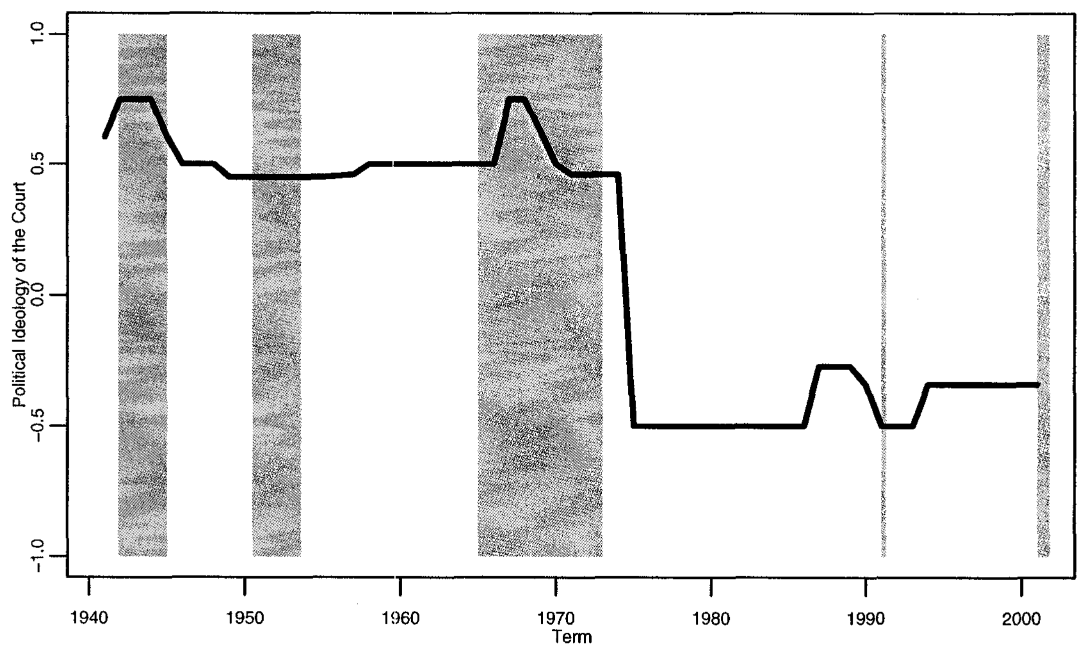


Figure 3.3: The political ideology of the median justice on the U.S. Supreme Court, 1941-2003. The line depicts the Segal & Cover score of the median justice for each term. The scores range from 1.00 (most liberal) to -1.00 (most conservative).[81] The gray shading depicts terms during which the Court heard disputes during a war period.

varied during times of war and peace. Figure 3.4 plots density estimates representing the distribution of cases by political ideology during times of war and the other, during times of peace. During periods of peace ("no war"), relatively left-of-center Courts (those toward the left end of the figure) and relatively right-of-center Court (toward the right end) decided roughly the same number of cases. Yet almost exclusively liberal courts decided cases during war.

Figure 3.5 plots the distribution of the cases aggregated by term such that the higher the circles (which represent the terms included in our dataset) the greater the proportion of decisions supporting rights; and the further left the circles, the more "liberal" the Court

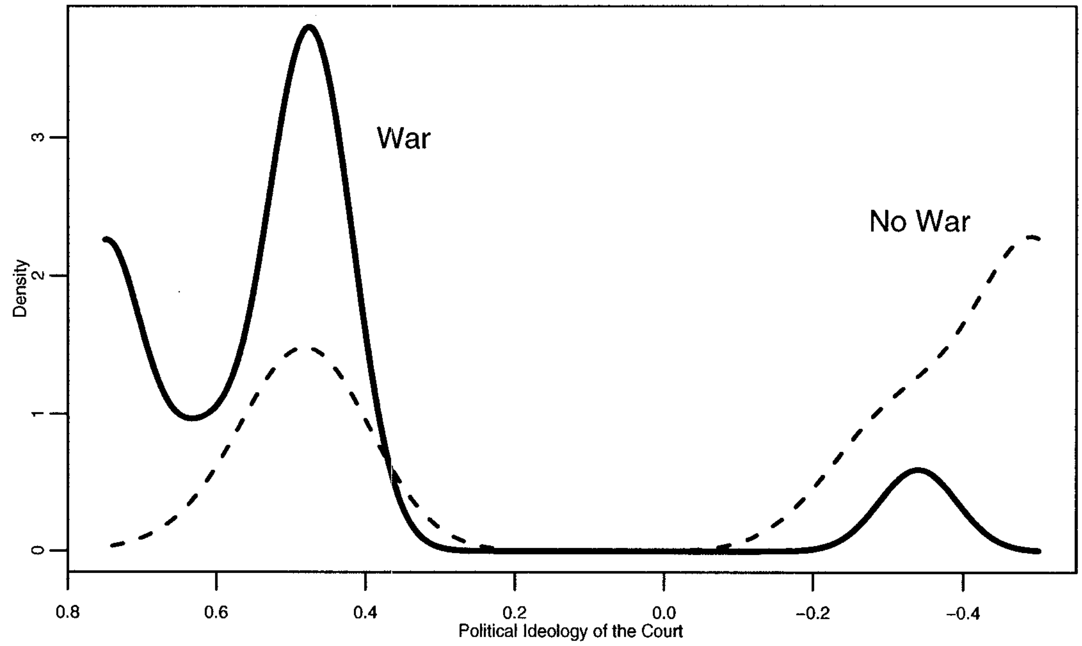


Figure 3.4: Kernel Density Estimate of Ideology Conditional on War, 1941-2002[82]

(again, on the Segal and Cover scale). Note that while right-of-center justices dominated during many terms as indicated by the gray circles, during only two was a crisis in effect (the Gulf and Afghan War) as indicated by the white circles; rather liberal justices controlled during all others.

Case Salience, Lower Courts, and Time Effects

We also consider other confounding variables of case salience, decisions of lower courts, and changing dynamics across time. We measure case salience by whether a case received front-page coverage in the *New York Times* on the day after the Court announced its

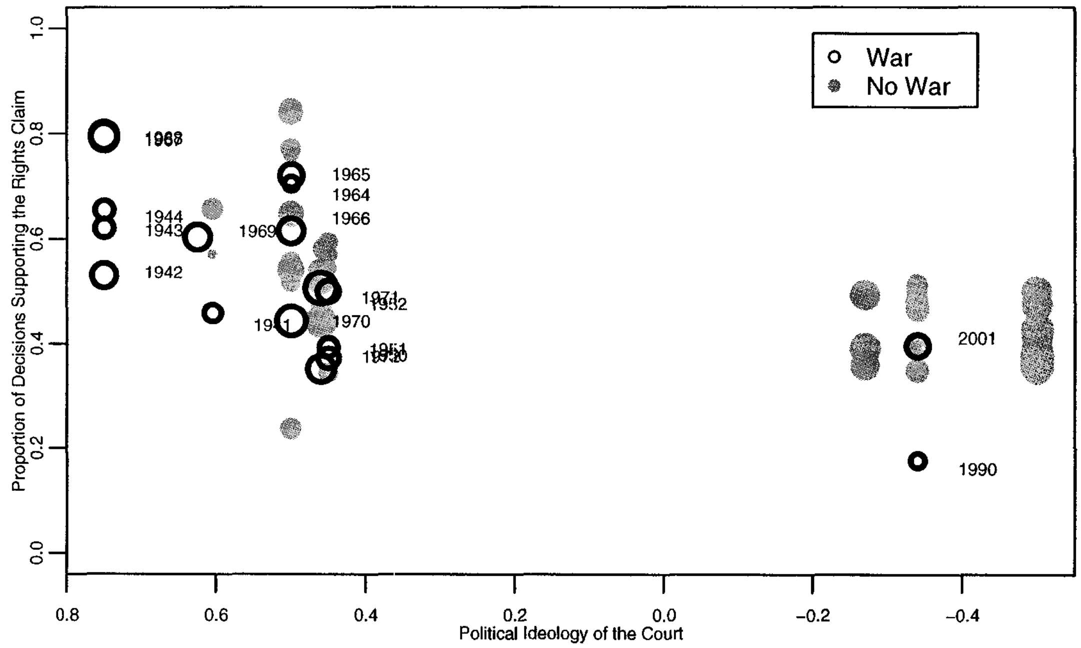


Figure 3.5: The Preponderance of Left-of-Center Courts During Times of War: The Empirical Distribution of the Data. N=3,345. Each circle is weighted by the number of cases decided in the term, with an average of roughly 55 cases decided each term.[83]

decision, as shown in Figure 3.6.[84] Overall, 711 of the 3,345 civil liberties cases (roughly 21 percent) were covered on the front page of the *Times*, with 199 of the 711 (roughly 28 percent) occurring in the midst of war and the remaining 512 (72 percent), reported in times of peace.

To account for the "reversal" tendency of the Court, we examined the outcome (either favorable or not to the litigant claiming a rights infringement) reached in the lower court. Lastly, to account for unobservable changes across time, we consider the term in which a case is decided. Many reasons exist for this decision, not the least of which is Goldsmith

[84]Lee Epstein & Jeffrey A. Segal, Measuring Issue Salience, 44 Am. J. Pol. Sci. 66 (2000). A list of the salient cases is available at: http://artsci.wustl.edu/~polisci/epstein/ajps/.

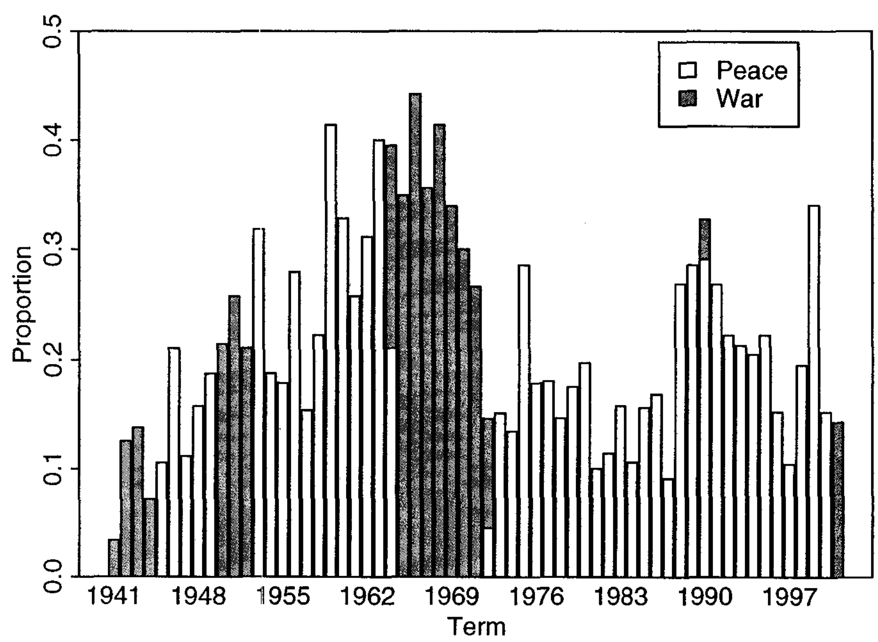


Figure 3.6: Proportion of U.S. Supreme Court decisions in the areas of rights, liberties, or justice that are salient, 1941-2001 terms. 711 of the 3,345 cases of rights, liberties, and justice decided during the period under analysis are salient. The white bars indicate the absence of a crisis (e.g., the country was not at war) at the time the Court heard oral arguments in the cases; darker bars indicate the presence of a crisis.[85]

and Sunstein's theory that a fundamental shift has occurred in legal culture over time.[86]

3.3 Empirical Results on the Causal Effect of War

Table 3.3 presents balance statistics of the full data. In line with our earlier analyses, we can observe that cases decided during war are far more likely to have been resolved by a left-leaning Court. They are also much less likely to have been decided in the liberal direction in the lower courts and are obviously likely to have been decided in earlier terms.

[86]Jack Goldsmith & Cass R. Sunstein, *Comment: Military Tribunals and Legal Culture: What a Difference Sixty Years Makes*, 19 CONST. COMMENT. 261 (2002).

	Mean under War	Mean under Peace	Standard Deviation	t-stat
Lower Court	0.30	0.46	0.49	8.79
Politics	1.01	0.37	0.49	−48.84
Term	1964.21	1976.94	15.71	21.48
Salience	0.26	0.20	0.41	−3.54
Pre-1975	0.92	0.33	0.50	−44.55

Table 3.3: Summary statistics of overall sample.

Given these substantial differences, regression-based techniques are likely to fare poorly.[87] We place our analysis in the framework of causal inference known as the Rubin Causal Model.[88] Let the decision of the Court for case $i = 1, \ldots, n$ be denoted by a binary variable Y, such that $Y_i = 1$ if the Court decided case i liberally and 0 otherwise. Let $T_i = 1$ if case i was decided during war, and 0 otherwise, and X_i denote observed pre-treatment covariates for i. $Y_i(1)$ and $Y_i(0)$ signify the *potential outcomes* for case i under war $T = 1$ and $T = 0$, respectively. We make the stable-unit-treatment-value assumption that potential outcomes of some unit i are independent of the assignment of treatment of some unit $j \neq i$. The crucial identification assumption, where war is not randomized, is that of ignorable treatment

[87]See, e.g., Gary King & Langche Zeng, When Can History Be Our Guide? The Pitfalls of Counterfactual Inference (on file with authors).

[88]See generally Paul W. Holland, *Statistics and Causal Inference*, 81 J. AM. STAT. ASS'N 945 (1986); Joshua D. Angrist & Alan B. Krueger, *Empirical Strategies in Labor Economics*, in 3A HANDBOOK OF LABOR ECONOMICS 1277 (Orley Ashenfelter & David Card eds., 1999); PAUL R. ROSENBAUM, OBSERVATIONAL STUDIES (2d ed., 2002); and JEFFREY M. WOOLDRIDGE, ECONOMETRIC ANALYSIS OF CROSS SECTION AND PANEL DATA 603-44 (2002).

	Mean under War	Mean under Peace	Standard Deviation	t-stat	Bias Reduction
Lower Court	0.32	0.30	0.46	0.76	91.4%
Politics	0.86	0.87	0.29	−0.28	99.4%
Term	1967.64	1967.66	15.10	−0.02	99.9%
Salience	0.25	0.28	0.44	−0.96	72.9%
Pre-1975	0.85	0.86	0.35	−0.10	99.8%

Table 3.4: Summary statistics of matched cases for Model 4 of Table 3.5 for non-warcases.[89]

assignment:

$$\{Y(1),Y(0)\} \perp\!\!\!\perp T|X$$

$$0 < Pr(T=1|X) < 1$$

which enables us to estimate missing potential outcomes after conditioning on covariates:

$$E(Y(0)|T=1,X) = E(Y(0)|T=0,X)$$

Matching on the propensity score $e(X) = Pr(T=1|X)$ simply provides a solution to matching when X is high-dimensional:

$$E(Y(0)|T=1,e(X)) = E(Y(0)|T=0,e(X)).$$

Concordantly, Table 3.4 shows that matching on the propensity score leads to better balance across all all pre-treatment covariates, reducing balance bias statistics from 73 to 99 percent. (A 100 percent bias reduction would indicate that we have matched exactly on all covariates.)

[89]In both cases, the estimated bias reduction R is calculated by $R = \frac{t_a - t_m}{t_a}$, where t_a represents the absolute value of the means test statistic from the overall sample and t_m represents the absolute value of the means test statistic for the matched sample.

	Non-Warcases			Warcases		
Method	ATE	SE	N	ATE	SE	N
Logistic Model	−0.07	0.03	3210	0.10	0.09	135
Exact Matching Except for Term	−0.07	0.03	1549	0.09	0.11	89
Exact Matching	−0.16	0.07	209			
Propensity Score Matching	−0.11	0.04	814	−0.02	0.13	62
Propensity Score Matching (Logistic Adjustment)	−0.10	0.03	814	−0.01	0.12	62

Table 3.5: Estimated causal effects of war. ATE is the average treatment effect, the causal effect of war on the probability of a liberal decision in civil liberties and rights cases. Model 1 presents estimated ATE from logistic regression of direction on judicial ideology, judicial ideology^2, lower court direction, war, and an indicator variable for pre-1975 term, where the Segal-Cover score is rescaled such that the minimum score is 0. Model 2 matches cases exactly on all covariates except for term. Model 3 matches cases exactly on all covariates including term. Model 4 matches on the propensity score, where the assignment model is estimated by a logistic regression of war on the lower court decision, judicial ideology, term, salience, term, salience and pre-1975 indicator. Model 5 additionally adjusts matched cases with a logistic regression.[91]

Table 3.5 presents matching estimates of the causal effect of war. Matching exactly on all covariates except for term, thereby dropping a substantial number of cases (primarily from the World War II and the Vietnam War periods), the estimated average treatment effect ("ATE") is -9 percent.[90] Matching on all covariates exactly, including term—meaning that we only consider cases decided during terms in which a war either ended or began (i.e., 1941, 1964, 1972 and 1991)—results in a rather large ATE of -16 percent. Since this estimate has a close relation to a difference-in-difference estimator, it may be the case that the short-term effects of war outstrip the long-term effects.

Results from the other matching models presented in Table 3.5 all roughly suggest that

[90]Of course this fails to take into account the smoothness of the Segal & Cover score and the relevance of time dynamics.

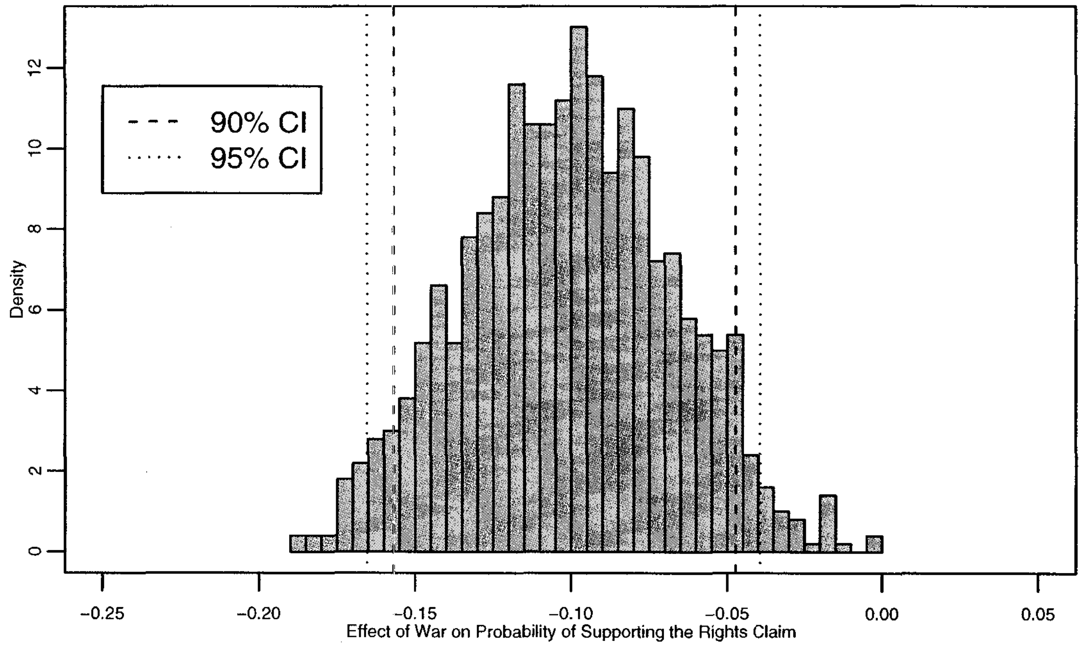


Figure 3.7: Histogram of the effect of wars civil rights and liberties. The vertical lines represent 90% and 95% confidence intervals.[92]

war decreases the probability of a liberal decision by 10% for cases not related to the war on the left side. Yet contrary to expectations, war-related cases exhibit no effect of war.

Figure 3.7 plots the histogram (with confidence intervals) of the simulated average treatment effect of war on non-warcases using the propensity score matching model in the final line of Table 3.5. Note that the bulk of the posterior mass of the causal effect of war is clearly below 0.

Figure 3.8 presents boxplots of the effect of war on cases unrelated to the war by issue area. Note the uniform decrease in the probability of a liberal decision in all areas, with boxes squarely to the left of the vertical line. The effect is least robust for civil rights cases,

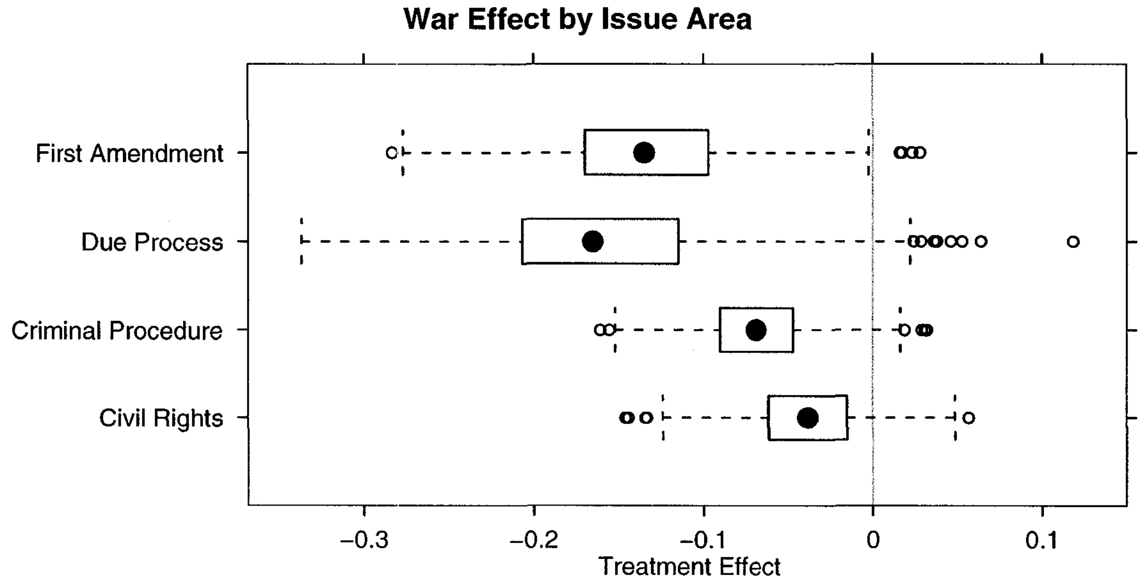


Figure 3.8: The effect of wars on the outcomes of Supreme Court Cases in the four areas of rights, liberties, and justice for non-warcases. The large circle represents the median effect, and the box represents 25^{th} and 75^{th} percentiles of the treatment effect, where box "whiskers" represent coverage 1.5 times the length of the box, and dots represent rare outliers.[94]

giving some credence to theories that desegregation can be sparked by wartime efforts. [93]

Figure 3.9 illustrates the estimated causal effect of war for each term in which a war

[93]See PHILIP A. KLINKER & ROGERS M. SMITH THE UNSTEADY MARCH: THE RISE AND DECLINE OF RACIAL EQUALITY IN AMERICA 4-5, 353 n4 (1999); Sheppele, *supra* note 18; and Dudziak, *supra* note 18.

[94]Boxplot of posterior distribution of 1000 simulated ATEs in each issue area, using a one-to-one matching model, discarding common support, with a logistic adjustment for non-warcases only. Simulated ATEs for each issue area are:

occurred.[95] The circles represent matched pairs of cases (weighted by the number of cases), with those in white indicating all the "war" cases in the pair and those in gray, all the "non-war" matched cases; the arrows specify the direction of the outcomes of the cases, whether they were more (an up, dashed arrow) or less (a down, solid arrow) favorable toward rights and liberties.

The most curious finding, unanticipated by extant theory, is that war does not exhibit a robust effect on war-related cases. While the number of war-related cases is small and potentially suspect due to endogenous coding, we hypothesize that such cases present particular challenges of executive encroachment on the court. As a result, the court protects its independence – rather than deferring to the executive, the justices rally 'round the court. If this hypothesis is true, we should find evidence that traditional determinants of judicial

Issue	**ATE (%)**	**SE**	**P-value**	**N**
Criminal Procedure	−0.07	0.03	0.03	354
Civil Rights	−0.04	0.03	0.12	256
First Amendment	−0.13	0.05	0.00	154
Due Process	−0.16	0.07	0.02	68

On simulation techniques for presenting results, see King, Tomz & Wittenberg, *supra* note 57. On boxplot techniques, JOHN W. TUKEY, EXPLORATORY DATA ANALYSIS (1977); and Robert McGill, John W. Tukey & Wayne A. Larsen, *Variations of Box Plots*, 32 AM. STATISTICIAN 12 (1978).

[95]So as to illustrate effects for without discarding warcases, this graph is based on a matching algorithm with replacement, where the propensity score is estimated using the logistic model, regressing war on lower court, politics, term, politics^2, salience, pre-1975 indicator, pre-1975*salience, term*salience, term*politics^2, term^2, lower court*salience, politics*salience, politics*lower court, lower court*term^2, and salience*term^2.

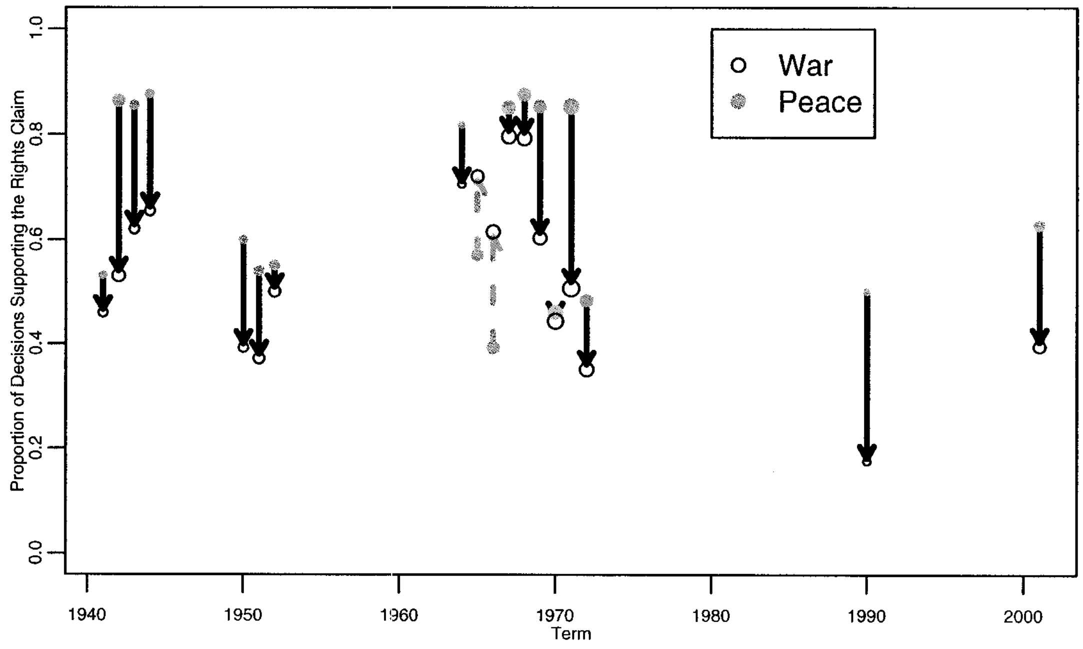


Figure 3.9: The effect of wars on the outcomes of supreme court cases in the areas of rights, liberties and justice: a comparison of matched pairs[96]

behavior, such as ideology, case salience, and lower court reversal, do not provide leverage over the subset of warcases. Figure 3.6 provides rough estimates consistent with this story: while the court decides cases that are salient and unrelated to the war more liberally, salience has no effect on warcases. Similarly, while the most liberal court exhibits an increase of roughly 15% in the probability of a liberal decision, this effect is not as robust for warcases. This evidence suggests that warcases may be of a different breed.

	Direct Effect	SE	N_1	N_0
Lower court, non-warcase	−0.23	0.02	1377	1793
Lower court, warcase	−0.17	0.12	23	51
Salient, non-warcase	0.09	0.02	679	2474
Salient, warcase	−0.11	0.13	25	45
Ideology, non-warcase	0.15	0.03	202	3008
Ideology, warcase	0.06	0.12	22	91

Table 3.6: Direct effects of other binary case variables for warcases and non-warcases, matching exactly on all other covariates. Ideology here is dichotomized by whether it takes on the highest observed value of the Segal-Cover score or not. N_1 and N_0 represent the number of observations matched.

3.4 Conclusion

Our results challenge extant the crisis thesis at its core. The justices of the U.S. Supreme Court seems to feel little responsibility to "rebuke the legislative and executive authorities when, under the stress of war [those authorities] have sought to suppress the rights of dissenters," as Justice Abe Fortas once wrote; nor have they acted in accordance with the maxim of *inter arma silent leges*. Instead, the justices appear to resist challenges to judicial independence posed by war-related cases, while balancing security and liberty interests only in cases unrelated to the war.

Part II

Experimental Approaches

Chapter 4

Shaken, Not Stirred: Evidence on Ballot Order Effects from the California Alphabet Lottery, 1978 – 2002

4.1 Introduction

For decades, scholars have attempted to assess the effects of ballot forms on elections, an effort that has intensified since the election debacles of *Bush v. Gore*. Ballot reform bears significant policy implications, with the Help America Vote Act of 2002 authorizing almost 4 billion dollars to reform efforts. One particular research agenda, spanning five decades and dozens of books and articles, examines the causal effect of name order on ballots. Scholars worry that particular rules of election administration may have major unintended, or possibly intended, consequences on election outcomes. Although some have claimed that candidates listed earlier on the ballot gain more votes solely because of ballot position, previous studies have yielded conflicting results about whether ballot order effects even exist.

The source of the disagreement may well be methodological. While scholars who assert large ballot order effects rely on observational data, where name order is not randomized and possibly confounded, studies finding no effect have often used laboratory experiments that may lack external validity. To overcome these difficulties, we analyze a *natural experiment*: statewide elections in California from 1978 to 2002. Since 1975, California elections law has mandated that the ballot order for statewide offices be physically randomized – after being "shaken vigorously," alphabet letters would be drawn from a lottery container to determine the order of candidates (Cal. Elec. Code 13112(c), 2003). The California alphabet lottery therefore offers a series of ideal natural experiments that allow us to test ballot order effects for varying types of candidates and offices in actual elections.

Examining a total of 473 candidates in 80 races from 13 general elections and 8 primary elections, we find that in general elections, ballot order substantially impacts minor party candidates, while having inconclusive effects on major party candidates. In primaries, on the other hand, being listed first significantly increases the vote share for any candidate. Major party candidates generally gain two percentage points of the total party vote, while minor party candidates may increase their vote shares by fifty percent of their baseline vote. In fact, ballot order might have changed the winner in as many as twelve percent of all primary races examined. In general elections, we find the largest effect for nonpartisan races where candidates in first position gain two percentage points on average. In contrast, we observe little difference in estimated causal effects of ballot order between types of offices for general elections, although effects appear to be somewhat larger for major offices in primaries. Our results are largely consistent with a theory of partisan cuing, where party labels convey information to uninformed voters (e.g., Schaffner and Streb, 2002; Snyder and Ting, 2002). When party labels are not available, as in nonpartisan races, or not informative, as in party primaries, voter decisions are most likely to be influenced by the ballot order.

As general methodological contributions, we demonstrate widely applicable statistical techniques to draw causal inferences in randomized experiments. While political scientists have begun to recognize the benefits of randomized experiments in the laboratory (e.g., Kinder and Palfrey, 1993) and in the field (e.g., Gerber and Green, 2000), many of such works do not place their analysis within a formal statistical framework of causal inference. We illustrate how this framework enables researchers to formally test treatment randomization, to identify substantively meaningful assumptions needed to estimate treatment effects, and to relax assumptions for crucial sensitivity analyses. The statistical methods we introduce in this paper can be applied to other experimental studies, allowing researchers to draw more robust causal inferences from their data.

Randomized *natural experiments* such as the California alphabet lottery provide an ideal opportunity for political scientists to draw causal inferences. Natural experiments take place in real settings (unlike laboratory experiments) and are not planned and implemented by researchers themselves (unlike field experiments). Compared to observational studies where treatments are not randomized, the analysis of natural experiments relies on assumptions that are actually verifiable from the data. Since external validity can be maximized and researchers are unconstrained by practical, financial, and ethical constraints, natural experiments can be more desirable than laboratory or field experiments. Although rare, when natural experiments such as the California alphabet lottery exist, they offer a special opportunity to test causal relationships.[1]

The rest of the paper proceeds as follows. Section 4.2 provides some background on ballot order effects and extant findings. Section 4.3 describes the California alphabet lottery and examines the crucial identification assumption that the resulting alphabets are indeed randomized. In Section 4.4, we discuss methodological issues of estimating ballot order effects and present the results of our analysis. Section 4.5 conducts alternative parametric and nonparametric robustness tests. Section 4.6 spells out the policy implication that election officials in all states should randomize the ballot order to minimize ballot effects, and provides evidence that randomization may be substantially more cost-effective at reducing voting bias than currently proposed electoral reforms. Section 4.7 concludes.

4.2 Elections and Ballot Order

Political scientists have rediscovered the importance of ballots since the days of counting chads in Florida (Niemi and Herrnson, 2003). Recent studies have ranged from examining the causal effects of the butterfly ballot (Brady et al., 2001; Wand et al., 2001), forms of voting equipment (Tomz and Van Houweling, 2003), partisan labels (Ansolabehere et al.,

[1]For example, in economics the Vietnam draft lottery has been used for a major study on the income returns of education (Angrist, 1990).

2003), and the ballot order of candidates (Kimball and Kropf, 2003; Krosnick, Miller and Tichy, 2003; Koppell and Steen, 2004). Current interest in ballot order is rooted in a half century of research investigating the causal effect of the order in which candidates appear on ballots (e.g., Bain and Hecock, 1957; Darcy, 1986; Darcy and McAllister, 1990; Gold, 1952; Miller and Krosnick, 1998; Scott, 1972). This research has spanned even beyond the United States, with studies in Australia (MacKerras, 1970), Britain (Bagley, 1966), Spain (Lijphart and Pintor, 1988), and Ireland (Robson and Walsh, 1973).

Beyond the academic literature, practical implications abound. Dozens of U.S. court decisions (e.g., *Bradley v. Perrodin*, 106 Cal. App. 4th 1153 (2003), *Gould v. Grubb*, 14 Cal. 3d 661 (1975); *Mann v. Powell*, 333 F. Supp. 1261 (1969)) and the drafting of electoral statutes in all fifty states (e.g., Ohio Rev. Code Ann. 3505.03 (Anderson 2003); N.M. Stat. Ann. 1-10-8.1 (2003)) rely on a version of the claim that vote shares will accrue to a candidate solely for being listed first on the ballot. And, electoral jurisdictions, from town to city, and from province to country, have proposed "remedying" ballot order effects by instituting some form of rotation or randomization.[2] At heart in these reform efforts lies the empirical claim of ballot order effects.

Scholars have also developed theoretical propositions about ballot order effects. Most broadly, psychological theory offers a hypothesis of "primacy effects," whereby the cog-

[2]For example, see Michael White, D'Hondted House of British PR, Guardian, at 23, Sep. 24, 1998 (proposing randomization in Britain); CA Assembly Bill AB 718 (Feb. 19, 2003) (mandating randomized alphabets for California cities); Bill Would Use Lottery to Place Candidate Names on Ballot, Union Leader, at A13, April 16, 1999 (proposing randomization for New Hampshire legislature).

nitive costs to processing alternatives bias individuals toward earlier choices (Miller and Krosnick, 1998, pp.293–295). Stemming from this theory, information salience, such as party labels, the prominence of the office or candidates, and media coverage, is hypothesized to affect the magnitude of ballot order effects. In a similar, though not necessarily consistent, vein, scholars have also proposed that candidates listed last should benefit from a "recency effect" (Bain and Hecock, 1957), or that candidates toward the middle of the ballot should be advantaged (Bagley, 1966). Ballot order effects may also emerge due to the fact that in most states, ballot order is actually informative, as major party candidates are generally listed earlier on the ballot.

Yet previous empirical studies disagree sharply over the existence of ballot order effects. Adherents claim that ballot order systematically affects the outcomes of many electoral contests, "[m]ost strikingly...in the highly-publicized and hotly contested presidential race [of 2002]" (Krosnick, Miller and Tichy, 2003, p.52). Detractors assert that "there is no evidence that there is a ballot position advantage in general elections" (Darcy, 1986, p.649).

The reason for this lack of consensus in the existing literature may well be methodological. The bulk of previous studies has relied on *observational data*, in which the name order is not physically randomized. Such analyses necessarily rest on assumptions that are difficult to verify, and their validity may be questioned if any confounding effects, and thereby omitted variable bias, exist. As outlined in Section 4.7, the majority of U.S. states arrange the order of candidates on ballots by some partisan or alphabetical rule, making the identification of ballot order effects difficult. In addition, Darcy and McAllister (1990, pp.8–10) finds numerous cases of abuse by elections officials and candidates, who, believing ballot order effects to exist, manipulate the placement of candidates to maximize expected vote share.

Even worse, some studies may not have any evidence on the quantity of interest asserted. In 1975, the California Supreme Court ruled to prohibit listing incumbents first. This decision was largely based on a study that analyzes only non-incumbent elections for the reason that these were the only ones for which ballot order was rotated (Scott, 1972). Inferences from that study may thereby result in severe extrapolation beyond the bounds of the data if non-incumbent elections systematically differ from elections with incumbents (King and Zeng, 2002). The best observational research to date examines elections in which candidate names were rotated. Even those studies, however, assume complete randomization without testing it.

Studies that actually analyze some form of randomized data have found little evidence for ballot order effects (Darcy, 1986; Gold, 1952). Miller and Krosnick (1998, p.297) conclude, after reviewing over 30 articles and books on the subject, that the only studies without design flaws (with randomization) detect no ballot order effects.[3] Yet inferences from such randomized experiments might be limited to unrepresentative samples, such as university students and county fairs (Darcy, 1986), elections for the anthropological association (Gold, 1952), and unrealistic lab settings, such as an election with candidates for whom no other information but the name is known (Bagley, 1966). In response to

[3]One exception to this general experimental finding is Forsythe et al. (1993).

these problems, Darcy and McAllister (1990, p.5) concludes that "much of the literature is methodologically flawed," while one expert opined that "there is virtually nothing at all [that has] been done on the subject much less anything shown" (Miller and Krosnick, 1998, p.318).

To our knowledge, the only study that analyzes randomized name order in real elections is Krosnick, Miller and Tichy (2003) ("KMT"). We extend and improve that study in two principal ways. First, KMT only examines two randomized races from California and one randomized race from North Dakota for the US President and Senate races for the general election in 2000. In contrast, we analyze 80 races in 13 general and 8 primary elections, spanning 10 offices from 1978 to 2002.[4] As a result, we are able to analyze different types of offices, candidates, and elections to address key substantive propositions in the literature. Our dataset also allows us to treat races as repeated natural experiments, yielding inferences with greater precision. This is an important advantage given that the analysis of each election consists only the sample size of 80 assembly districts. Second, we improve the methodology used in KMT's analysis. In particular, we do not impose parametric assumptions, except for as a sensitivity analysis in Section 4.5.1, to identify the treatment effect.[5] Our conclusions contradict a main claim of KMT that ballot order significantly affects major candidates in general elections, most notably in the presidential election of 2000.

[4]The bulk of the analysis in KMT relies on observational data from Ohio elections, for which randomization is more difficult to verify.

[5]KMT used linear regression in its analysis.

4.3 The California Alphabet Lottery

In this section, we first describe the procedure of the California alphabet lottery as defined in state election law. Second, we conduct statistical tests to show that the alphabets used for the elections in the past twenty years are indeed randomly ordered, a crucial assumption of our subsequent analysis.

4.3.1 Lottery Procedure

California election ballots are printed in column-vertical format, depicting the name, party, and occupation of all candidates. Until 1975, California elections law mandated that incumbents appear first on the ballot in the majority of statewide elections (Scott, 1972, p.365). In 1975, however, the California Supreme Court, in a startling act of judicial activism, struck down the provision that reserved the first ballot position to incumbents, and held as unconstitutional, on equal protection grounds, ballot forms that present candidate names in alphabetical order (*Gould v. Grubb*, 14 Cal. 3d 661, 1975). The decision relied prominently on studies and testimonies by Bain and Hecock (1957) and Scott (1972). Scott (1972, p.376) investigated the effect of ballot order using ballot rotations in ten non-incumbent California races. While providing only point estimates of the ballot order effect, the study concluded that "one can attribute at least a five percent increase in the first listed candidate's vote total to a positional bias," a figure that has often been quoted by the Secretary of State since.[6]

In response to that decision, the California legislature passed an alphabet randomization procedure to determine the ballot order of candidates.[7] The randomization applies to the national offices of the U.S. Presidency and U.S. Senators, as well as the statewide offices of the Governor, Lieutenant Governor, Secretary of State, Controller, Treasurer, Attorney General, Insurance Commissioner, and the Superintendent of Public Instruction. The law spells out in precise detail the procedure for drawing a "randomized alphabet":

> Each letter of the alphabet shall be written on a separate slip of paper, each of which shall be folded and inserted into a capsule. Each capsule shall be opaque and of uniform weight, color, size, shape, and texture. The capsules shall be placed in a container, which shall be shaken vigorously in order to mix the capsules thoroughly. The container then shall be opened and the capsules removed at random one at a time. As each is removed, it shall be opened and the letter on the slip of paper read aloud and written down. The resulting random order of letters constitutes the randomized alphabet, which is to be used in the same manner as the conventional alphabet in determining the order of all candidates in all elections. For example, if two candidates with the surnames Campbell and Carlson are running for the same office, their order on the ballot will depend on the order in which the letters M and R were drawn in the randomized alphabet drawing (Cal. Elec. Code 13112(a), 2003).

The container used in the drawing is in the same style as that used in one of official state lotteries. The code further mandates that the drawing be open to public inspection and advance notice made to the media, the representative of local election officials, and party chairmen (Cal. Elec. Code 13112(c), 2003). These explicit procedures defined in the

[6]Bain and Hecock (1957, p.85) similarly found that "[t]he first position in a vertical list was universally the preferred when paper ballots were used" in several Michigan cities.

[7]The provision was added under Assembly Bill 1961, 1975–76 Regular Session of the California Assembly, as Stats 1975, ch. 1211, Sections 16 & 17.

law are designed to ensure accurate implementation of randomization. California election officials appear to have taken this duty seriously. The Secretary of State, in charge of the randomization procedure, maintains two designated "random alpha persons" who draw the letters from a lottery bin. When queried about the process, officials insist that "it's the law" to randomize.[8]

Equally important to our estimation strategy, California elections law mandates that the randomized ballot order is rotated through the 80 assembly districts for all statewide candidates,

> the Secretary of State shall arrange the names of the candidates for the office in accordance with the randomized alphabet ... for the First Assembly District. Thereafter, for each succeeding Assembly district, the name appearing first in the last preceding Assembly district shall be placed last, the order of the other names remaining unchanged (Cal. Elec. Code 13111(c), 2003).

The rotation itself is not implemented randomly, an issue which we take into account in our statistical analysis. The procedure nonetheless provides substantial variation of the ballot order, which enables the estimation of candidate-specific ballot order effects.

Figure 4.1 depicts the 80 Assembly Districts that have been in effect from 1992 until 2002. Note that the ordering of Assembly Districts is *not random*, a property that we explicitly address in our analysis. To the contrary, the California Constitution mandates, *inter alia*, that (a) districts be numbered from north to south, (b) the population be "reasonably equal" across districts, (c) all districts be contiguous, and (d) geographical subregions be respected to the extent possible (Cal. Const., Art XXI, 1, 2003). Every ten years following

[8]Telephone interview with Melissa Warren, Elections officer at Office of Secretary of State, Aug. 15, 2003.

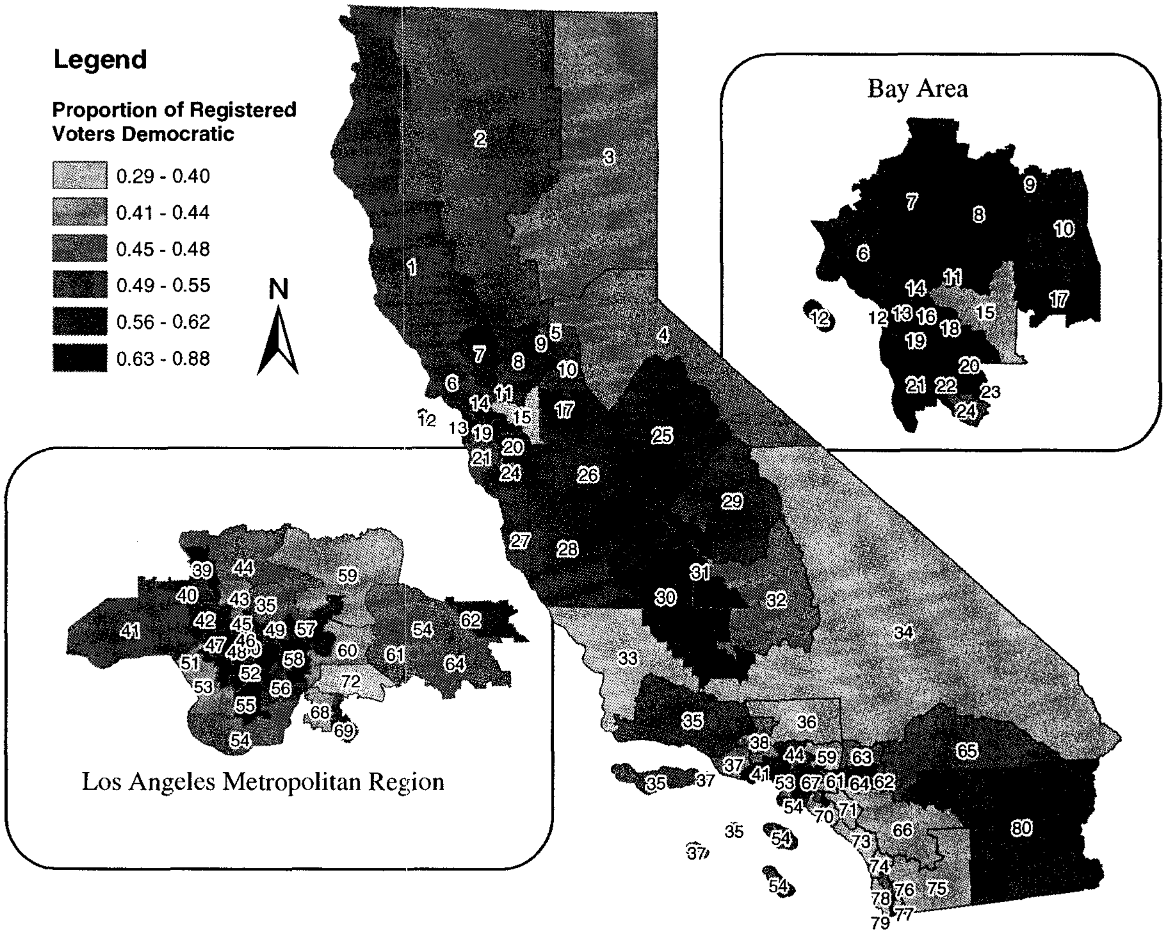


Figure 4.1: 1992 California Assembly Districts. The districts with darker color are those with a higher proportion of Democrats among registered voters.
Source: The California Spatial Information Library. Map created using Arcmap.

the national census, the districts are adjusted accordingly in state legislative reapportionment. For the time period of our interest, redistricting occurred in 1982, 1992, and 2002.

The randomized rotation procedure has remained virtually unchanged since 1975. A review of the legislative history reveals that original Assembly Bill 1961 passed in 1975 has remained identical with respect to the randomization procedure.[9] This stability of the

[9]Assembly Bill 1961, 1975 – 76 Regular Session of the California Assembly, as Stats 1975, ch. 1211, Sections 16 and 17.

election process allows us to examine the causal effects of ballot order for a host of elections from the past twenty five years.

One concern about the California alphabet lottery is that the randomized alphabet may induce behavioral changes of candidates, making it difficult to isolate the direct effects of ballot order on voters. For example, candidates listed last on the ballot in a particular assembly district might campaign more intensely in that district, in fear of some ballot order effect. Or, candidates might be chosen to assure a higher ballot order in favorable districts (Masterman, 1964). However, such a scenario seems unlikely given that the randomized alphabet is drawn very late in the game.

All but write-in candidates must have declared candidacy and been certified by the time that the drawing of a randomized alphabet takes place, and even sample (non-randomized) ballots are printed before the drawing. Only minor adjustments, such as removal of a candidate from the ballot in the case of a death, occur after the drawing.[10]

4.3.2 Are Alphabets Really Random?

Election officials seem to have taken seriously their legal obligation of conducting the alphabet lottery. Given the evidence of manipulation of ballot order in other states (e.g., Darcy and McAllister, 1990), however, we conduct statistical tests to ensure empirically the accurate implementation of the randomization. Such tests often help discover unexpected implementation errors of randomization (Imai, 2004). As shown in Table 4.1, we collected

[10]Even if there are candidate behavioral changes resulting from the drawing, this "intention-to-treat" effect may still be of important policy significance.

Year	Election	Randomized Alphabet
1982	Primary	S C X D Q G W R V Y U A N H L P B K J I E T O M F Z
	General	L S N D X A M W V T O F I B K Y U P E Q C J Z H R G
1983	Consolidated	L C P K I A U G Z O N B X D W H E M F V R S T Y Q J
1984	Primary	W M F B Q Y T D J U O V I K R H S N P C A E L Z G X
	General	V W I H R Q G J O M T S Y C A F U X K B P E Z N D L
1986	General	Q N H U B J E G M V L W X C K O F D Z R Y I T S P A
1988	Primary	W O K N Q A V T H J F Z L B U D Y M I R G C E S X P
	General	S W F M K J U Y A T V G O N Q B D E P L Z C I X R H
1990	Primary	E J B Y Q F K M O V X L N Z C W A P R D G T H I S U
	General	W F C L D I N J H V K O S A R E Q B T M Y U G Z X P
1992	Primary	U R F A J C D N M K P Z Y X G W O H E B I S V L Q T
	General	F Y U A J S B Z G O E Q R L I M H V N T P D K X C W
1994	Primary	K J H G A M I Q U N C Z S W V R P Y B L O T D F E X
	General	V I A E M S O K L B G N W Y D P U F Z Q J X C R H T
1996	Primary	G E F C Y P D B Z I V A U S M L H K N T O J Q R X W
	General	J Y E P A U S Q B H T R K N L X F D O G M W I Z C V
1998	Primary	L W U J X K C N D O Q A P T Z R Y F E V B H G I M S
	General	W K D N V A G P Y C Z I S T L J X Q O F H R B U M E
2000	Primary	O P C Y I H X Z V R S Q E K L G D W J U T M B F A N
	General	I T F G J S W R N M K U Y L D C Q A H X O E B V P Z
2002	Primary	W I Z C O M A Q U K X E B Y N P T R L V S J H D F G
	General	H M V P E B Q U G N D K X Z J A W Y C O S F I T R L
2003	Recall	R W Q O J M V A H B S G Z X N T C I E K U P D Y F L

Table 4.1: Randomized Alphabets Used for the California Statewide Elections Since 1982.

the randomized alphabets used for 23 California statewide elections since 1982. We use this list to test whether the randomization procedure described above has in practice produced completely randomized alphabets not favoring any particular letters, and hence particular candidates.

We conduct a rank test under the null hypothesis that the alphabet is completely randomized. In particular, we compare the relative positions of all possible pairs of letters by calculating the mean absolute rank differences of paired letters across elections,

$$\frac{1}{325}\sum_{i=1}^{26}\sum_{j\neq i}^{26}\left|\frac{1}{23}\sum_{k=1}^{23}\{R(L_{ik})-R(L_{jk})\}\right|, \tag{4.1}$$

where $R(L_{ik})$ denotes a rank or position of the ith letter of the alphabet on the randomized list of the kth election. This statistic averages the relative positions of two distinct letters over 23 elections and all possible such pairs. The resulting sample statistic for the 23 observed alphabets in Table 4.1 is 2.07, representing the average absolute difference in the relative positions of all possible pairs of distinct letters. Under the null hypothesis of complete randomization, the distribution of this statistic can be calculated *exactly* by considering all possible lists of alphabet which are equally likely. However, since there are 26! such lists for each election, we approximate this statistic by simulation. We draw $10,000$ lists of 23 randomized alphabets with equal probability, and then calculate the statistic for each list. Finally, we compute the one-tailed p-value by comparing the observed value of the statistic with its simulated values. The resulting one-tailed p-value is 0.15, indicating that we fail to reject the null hypothesis of complete randomization.[11]

4.4 Causal Effects of Ballot Order

With the aid of the California State Archives and the Statewide Data Base at the University of California, Berkeley, we coded election returns data by assembly districts for a

[11] We also conducted similar randomization tests based on the rank differences between even and odd letters, and letters in the top and bottom half of the true alphabet. And again we cannot reject the null hypothesis of complete randomization with one-tailed p-values of 0.27 and 0.30, respectively.

Election	President	Senate	Governor	Lt. Gov.	Atty Genl	Controller	Ins. Comm.	Sec. State	Treasurer	Supt Educ
1978 General	–		5							
1980 General	7	5	–	–	–	–	–	–	–	–
1982 General	–	5	5							
Primary	–	19	20							
1984 General	5	–	–	–	–	–	–	–	–	–
1986 General	–	5	5							
Primary	–	20	9							
1988 General	5	5	–	–	–	–	–	–	–	–
Primary		6	–	–	–	–	–	–	–	–
1990 General	–	–	5							
Primary	–	–	19							
1992 General	6	5,5[a]	–	–	–	–	–	–	–	–
1994 General	–	6	5							
Primary	–		12							
1996 General	8	–	–	–	–	–	–	–	–	–
1998 General	–	7	7	7	5	7	6	7	6	2
Primary	–	13	17	13	10	7	8	8	9	5
2000 General	7	7	–	–	–	–	–	–	–	–
Primary	23	15	–	–	–	–	–	–	–	–
2002 General	–	–	6	7	5	5	6	7	6	2
Primary	–	–	11	8	6	10	11	13	7	4

Table 4.2: Number of Candidates Running in All Races Examined. "–" indicates that no election was held for that office in a particular year. Blank cells represent races where election returns data were not available by assembly districts. The number of candidates in this table differs slightly from total number of candidates analyzed because of several uncontested party primaries.

[a]There were two senatorial elections in 1992 both of which had five candidates running.

total of 80 statewide races (44 primary races and 36 general races), going back to 1978. Table 4.2 lists all the races examined in this paper. These include 13 general elections and 8 primaries for 10 statewide offices, yielding a total of 473 candidates analyzed. We also collected the candidate names from which we reconstructed the ballot order for each of these races in each district using the official randomized alphabets.[12]

In what follows, we describe our analysis of the California alphabet lottery and present estimates of our quantities of interest, i.e., ballot order effects, and such effects conditional on parties, offices, and elections. We first place our analysis in the formal statistical framework of causal inference. Second, we describe our estimation strategies and interpret the identification assumptions. Finally, we present our estimates and compare them to the margins of victory observed in the races in order to compute the potential substantive impact on election outcomes if the candidate names were ordered differently.

4.4.1 Causal Inference and Treatment Assignment

We consider the California alphabet lottery as a series of repeated randomized experiments. In particular, we have a total of $i = 1, \ldots, N$ races in each of which the candidates' name order on the ballot is independently randomized.[13] In race i, we have $j = 1, \ldots, J_i$ candidates running for the office, and we observe the randomized (and rotated) ballot order

[12]When the official randomized alphabet was not available, we gathered available Assembly District ballots to recover the ballot order.

[13]Strictly speaking, the randomization is conducted in every election not in every race. However, since candidates differ in every race of the same election, we consider different races in the same election as independent experiments.

in each of $k = 1, \ldots, K$ districts ($K = 80$). That is, we observe the ballot order, the value of the multi-valued treatment variable, $T_{ijk} = t$, for all k where $t \in \mathcal{T} = \{1, \ldots, J_i\}$. Since we analyze each race separately, for the rest of this section, we suppress the is for notational simplicity. Finally, we also observe each candidate's vote share for every district with the corresponding ballot position, denoted by $Y_{jk}(t)$ for $t = T_{jk}$.

We adopt the formal statistical framework for causal inference, frequently referred to as the Rubin causal model (Rubin, 1974; Holland, 1986). In this framework, $\mathcal{Y} = \{Y_{jk}(t);\, t \in \mathcal{T}\}$ is regarded as a set of potential outcomes, and $\mathcal{T}$ is a set of potential treatment values and $Y_{jk}(t)$ is a random variable that maps a particular potential treatment, t, to a potential outcome. The fundamental problem of causal inference is that only one realization of potential outcomes for each unit is observed. This means that we do not observe the counterfactual vote shares in a district if the candidates' names on the ballot were ordered differently. Causal inference hence requires estimating these missing potential outcomes.

In the majority of experimental studies, researchers assign treatment to units that are randomly selected with equal probability. We call this common procedure "simple random treatment assignment" with the following definition,

DEFINITION 1 (SIMPLE RANDOM TREATMENT ASSIGNMENT) *From a list of K units, assign a treatment to n units that are randomly selected with an equal probability (without replacement).*

In the California alphabet lottery, the randomization procedure is somewhat different since randomization is conducted only for the first Assembly District and the treatment assignments for the other districts are systematically determined thereafter. That is, the randomized ballot order in the first district will be rotated such that in the next district, the candidate in the jth position ($j \geq 2$) will be in $(j-1)$th position and the candidate in the first position will be placed in the last position, and so on. We call this procedure "systematic random treatment assignment" and formally define it as follows,

DEFINITION 2 (SYSTEMATIC RANDOM TREATMENT ASSIGNMENT) *From a list of K units, assign a treatment to the rth unit, and every Jth unit thereafter. For simplicity, assume that $K = nJ$, where n is the desired size of a treatment group, J is a positive integer no less than 2, and r is an integer variable randomly drawn with an equal probability from* $\{r : 1 \leq r \leq J\}$.

The names, *systematic* and *simple*, come from the fact that these two randomization schemes resemble simple random sampling and systematic sampling in the survey sampling literature (e.g., Cochran, 1977, ch.8). This connection enables us to apply the results of this literature to our analysis of the California alphabet lottery.

4.4.2 Identifying Causal Effects of Ballot Order

Identification of Average Treatment Effects

Within the Rubin causal model, we make two additional assumptions that have a clear substantive interpretation and allow us to quantify ballot order effects for each candidate (hence, we suppress subscripts j's for further notational simplicity).

ASSUMPTION 1 (NO INTERFERENCE AMONG UNITS, RUBIN (1980)) $Y_k(t) \perp\!\!\!\perp T_{k'}$ *for all* $t \in \mathcal{T}$ *and* $k' \neq k$,

where $\perp\!\!\!\perp$ denotes independence. This assumption is also referred to as stable unit treatment value, and implies that the potential outcome for one unit does not depend on the treatment assignment of another unit (Cox, 1958).[14] In the case of the California alphabet lottery, this assumption is reasonable when considering candidate-specific effects. However, the assumption is violated in an analysis that pools candidates: since candidate vote shares in one district must sum to 1, a ballot order effect on one candidate necessarily affects the remaining candidates. One potential solution to this interference would be to explicitly model all candidates' vote shares at the same time as a function of their ballot positions by, for example, a multinomial logit model. The advantages and disadvantages of such an approach are explored in Section 4.5.1. Given the small sample size of 80 districts and the large and different number of candidates in many of these races, we first estimate the causal effects of ballot order separately for each candidate. This not only meets Assumption 1, but also relaxes implicit pooling assumptions in extant studies, permitting us to estimate differential effects for different party candidates.

The second assumption is essential for unbiased estimation of treatment effects and is satisfied by simple random assignment of Definition 1.

ASSUMPTION 2 (RANDOM ASSIGNMENT) $Y_k(t) \perp\!\!\!\perp I(T_k = t)$ *for all* k *and* $t \in \mathcal{T}$,

where $I(\cdot)$ represents an indicator function. It is straightforward to show that systematic random assignment of the California alphabet lottery (Definition 2) also satisfies this assumption since the ballot order is independent of potential outcomes. In Appendix 4.8, we empirically test this assumption by examining the balance of the covariates from the

[14] In addition, the stable unit treatment value assumption asserts that there are no differing version of the treatment. This means that changes in ballot order are the same across Assembly Districts, and could be violated if ballots differed dramatically across districts.

Census and registration data.

Assumptions 1 and 2 suffice to identify the average ballot order effect for each candidate from the observed data. Specifically, the average treatment effect for candidate j, $\tau \equiv \overline{Y}(t') - \overline{Y}(t)$ where $\overline{Y}(s) \equiv \sum_{k=1}^{K} Y_k(s)/K$ for $s = t, t'$ with $t \neq t'$, can be estimated without bias,

$$E(\hat{\tau}) \;=\; E\left[\widehat{Y}(t') - \widehat{Y}(t)\right] \;=\; \tau, \tag{4.2}$$

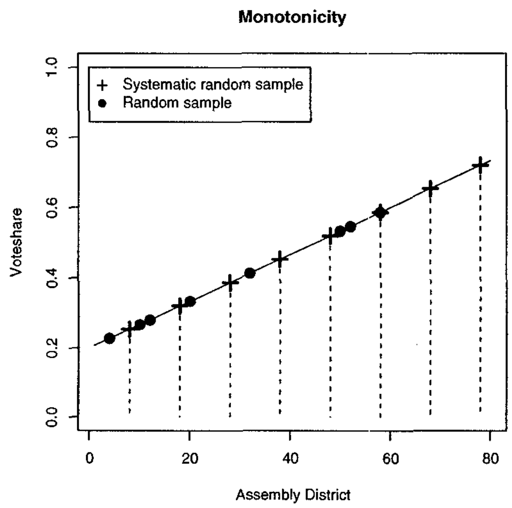
where $\widehat{Y}(s) \equiv \sum_{k \in \{k:T_k = s\}} Y_k(s)/n_s$ and n_s is the number of assembly districts where the candidate is assigned to the sth ballot position.

Identification of Variance

Although an unbiased estimate of the average ballot effect is readily available, the variance calculation of this estimator is not straightforward. This is because systematic random assignment, unlike simple random assignment, involves only one randomization. The population variance of the estimated average ballot order effect $\hat{\tau}$ in equation 4.2 is the sum of the variances for the two potential outcomes, i.e., $V(\hat{\tau}) = V\{\widehat{Y}(t')\} + V\{\widehat{Y}(t)\}$.[15] Using the result from the systematic sampling literature (e.g., Madow and Madow, 1944), for $s = t, t'$ with $t \neq t'$ each of the two variances is

$$V\{\widehat{Y}(s)\} \;=\; \frac{\sigma_s^2 (K-1)}{n_s K}\{1 + (n_s - 1)\rho_s\}, \tag{4.3}$$

[15]Here we consider the population to consist of all potential outcomes for each candidate.



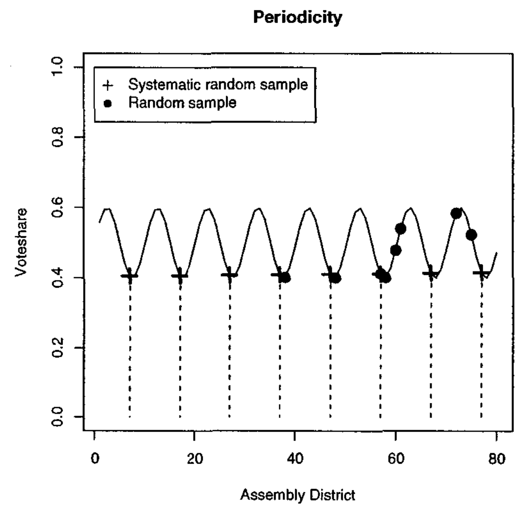


Figure 4.2: Simple Random and Systematic Random Assignment under the Populations with Monotonic and Periodic Trends. The figure shows how the order of the population affects variance estimation under the a given assignment mechanism.

where σ_s^2 is the population variance of $Y_k(s)$. ρ_s is the intraclass correlation coefficient between pairs of potential outcomes within the same systematic sample and is given by

$$\rho_s = \frac{2}{(n_s-1)(K-1)\sigma_s^2}\sum_{l=1}^{J}\sum_{m<m'}\{Y_{lm}^*(s)-\overline{Y}(s)\}\{Y_{lm'}^*(s)-\overline{Y}(s)\}, \quad (4.4)$$

where $Y_{lm}^*(s)$ denotes the potential vote share in the mth district in the lth systematic sample (for the candidate and under the sth ballot position). ρ_s represents a measure of the homogeneity of each potential outcome within a sample averaging over the J possible treatment assignment combinations. Unfortunately, $V(\hat{\tau})$ cannot be consistently estimated without making some assumptions about the population since we only observe one systematic random sample of the treatment assignment combination.

Nevertheless, the expression of $V(\hat{\tau})$ from equation 4.3 has a useful interpretation. If $\rho_s = 0$, the variance is the same as that for simple random assignment. When $\rho_s < 0$, we have a heterogeneous sample that is more representative of the population and the variance is lower than that of simple random assignment. For example, suppose that $Y_k(s)$ is monotonically increasing in k as in the left panel in Figure 4.2. Systematic random assignment ensures that we obtain units across the whole range of k, whereas simple random assignment does not. In the figure, the circles representing simple random assignment are centered toward the lower end of the vote share, whereas systematic random assignment is evenly distributed across the assembly districts. On the other hand, when $\rho_s > 0$, we have a homogeneous sample, and thereby the variance of the estimator is greater than that of simple random assignment. The most pathological case is one of periodicity that coincides with J, as shown in the right panel of Figure 4.2. In that case, simple random assignment is more efficient, since it ensures sampling units that are along any part of the wave-like pattern of the population. Systematic random assignment, however, samples only those assembly districts with low vote shares, since the periodicity coincides almost exactly with J.

Given this nature of systematic random assignment, we estimate the variance based on different assumptions about the population. In particular, we consider the following four types of variance estimators for $V\{\widehat{Y}(s)\}$ developed in the literature (e.g., Wolter, 1984). They are based on the population models with <u>rand</u>om order, <u>line</u>ar trend, <u>stra</u>tification, and autocorrelation.

$$\widehat{V}_{rand} = (1-f)\frac{\sum_{k\in\{k:T_k=s\}}\{Y_k(s)-\widehat{Y}(s)\}^2}{n_s(n_s-1)}, \tag{4.5}$$

$$\widehat{V}_{line} = (1-f)\frac{\sum_{k\in\{k:T_k=s\}}\{Y_k(s)-2Y_{k-1}(s)+Y_{k-2}(s)\}^2}{6n_s(n_s-2)}, \tag{4.6}$$

$$\widehat{V}_{stra} = (1-f)\frac{\sum_{k\in\{k:T_k=s\}}\{Y_k(s)-Y_{k-1}(s)\}^2}{2n_s(n_s-1)}, \tag{4.7}$$

$$\widehat{V}_{auto} = \begin{cases} \widehat{V}_{rand}\left[1+2/\log\hat{\rho}_s+2\hat{\rho}_s/(1-\hat{\rho}_s)\right] & \text{if} \quad \hat{\rho}_s > 0, \\ \widehat{V}_{rand} & \text{if} \quad \hat{\rho}_s \leq 0, \end{cases} \tag{4.8}$$

where $f = n_s/K$ is the finite population correction and $\hat{\rho}_s = \sum_{k\in\{k:T_k=s\}}\{Y_k(s)-\widehat{Y}(s)\}\{Y_{k-1}(s)-\widehat{Y}(s)\}/\sum_{k\in\{k:T_k=s\}}^{n_s}\{Y_k(s)-\widehat{Y}(s)\}^2$. A few remarks about each estimator are worthwhile. First, $\widehat{V}_{rand}$ assumes that assembly districts are randomly ordered. While $\widehat{V}_{line}$ is designed to eliminate a linear trend by taking successive differences, $\widehat{V}_{strat}$ assumes that the mean of the potential vote shares is constant within each stratum of J districts. Finally, $\widehat{V}_{auto}$ is based on the autocorrelated population model where the correlation of two potential vote shares depends only on the difference in their assembly district number.

Given that we do not know which of these candidate estimators best approximates the true variance of the potential vote shares, we employ an auxiliary variable approach advocated in the systematic sampling literature to select the estimator. Since party registration is known to be one of the best predictors for a candidate's actual vote share in an election, it provides an ideal auxiliary variable. We evaluate the performance of the four estimators using party registration data for each election.[16] For any party and number of candidates

[16]If official registration data was unavailable for a particular election, we used registration data from the closest election.

running in a particular race, we can then calculate how the estimators perform across all possible systematic samples compared to known true variance of party registration.[17]

Given this auxiliary variable, we select the variance estimator that performed best in terms of mean squared error (MSE) criteria to estimate the variance of ballot order effects. For the 1998 and 2000 general elections, for example, among 66 candidates considered, 47% of the time the minimum MSE is the random list estimator and 33% of the time it is the autocorrelation estimator. The median variance bias among the selected estimators is 0.4%, and the variance bias ranges from −25% (5 percentile) to 35% (95 percentile). Interestingly, assuming a random list is generally *conservative* for California, since the intra-class correlation coefficient for all parties is negative at observed J. This is consistent with the registration patterns across Assembly Districts in California as seen in Figure 4.1, with more liberal urban districts clustered in the North and in Los Angeles, but generally more conservative districts in the South.

4.4.3 Estimated Causal Effects of Ballot Order

We estimate two primary quantities of interest: (a) the average *absolute gain* for each candidate due to being in first position, $\widehat{Y}(t=1) - \widehat{Y}(t \neq 1)$, and (b) the average *relative gain* for each candidate due to being in first position, $[\widehat{Y}(t=1) - \widehat{Y}(t \neq 1)]/\widehat{Y}(t \neq 1)$. While we additionally investigated effects of all other positions, we generally found that the primary robust effect was that of being in first position. We first illustrate our analysis

[17]For closed primary races, this approach may not be appropriate since party registrants are the only eligible voters. Thus, we conducted sensitivity analyses using both the random list and minimum MSE estimators.

with the results for the 1998 and 2000 elections, and then summarize the estimated ballot order effects for all elections by considering each race as a repeated experiment.

The top panel of Figure 4.3 presents estimates for the average *relative* percentage gain of all candidates in the 1998 and 2000 general elections, with vertical bars indicating estimated 95% confidence intervals, using the minimum MSE variance estimator. For 28 out of the 68 candidates there are significant effects for which the confidence intervals do not intersect zero. The median gain was roughly 10% of the baseline vote share. On the other hand, almost all of these estimates stem from minor party candidates, as seen by the fact that major party candidate estimates for Democrats and Republicans, signified by the dark thin bars, are concentrated in the bottom half of the ordering. Indeed, third party candidates have a median gain score of roughly 17%, whereas major party candidates had a relative gain of roughly 1%. In terms of absolute gains, however, the estimates are relatively small for general elections, with a median gain of roughly 0.2% of the total vote.

The bottom panel of Figure 4.3 presents the estimated average *relative* gains for the 1998 and 2000 primary elections.[18] The magnitude of the effects is substantially larger than

[18]The analysis of primary races is complicated slightly by the fact that California primary rules and reporting changed substantially over the years. In 1998 California changed from a closed party primary to an open primary, and reversed partially again in 2002 to a "modified closed" primary, under which registered voters could vote only on their affiliated party's ballot, but unaffiliated voters could still request party ballots or receive nonpartisan ballots by default. To accurately capture how winners are determined from primary races, and to facilitate comparisons across elections, we calculate candidate vote shares as a proportion of the party vote in primaries when multiple candidates are running, and vote share proportions of the total vote for uncontested (usually minor party) candidates. Lastly, since party registration for closed primaries

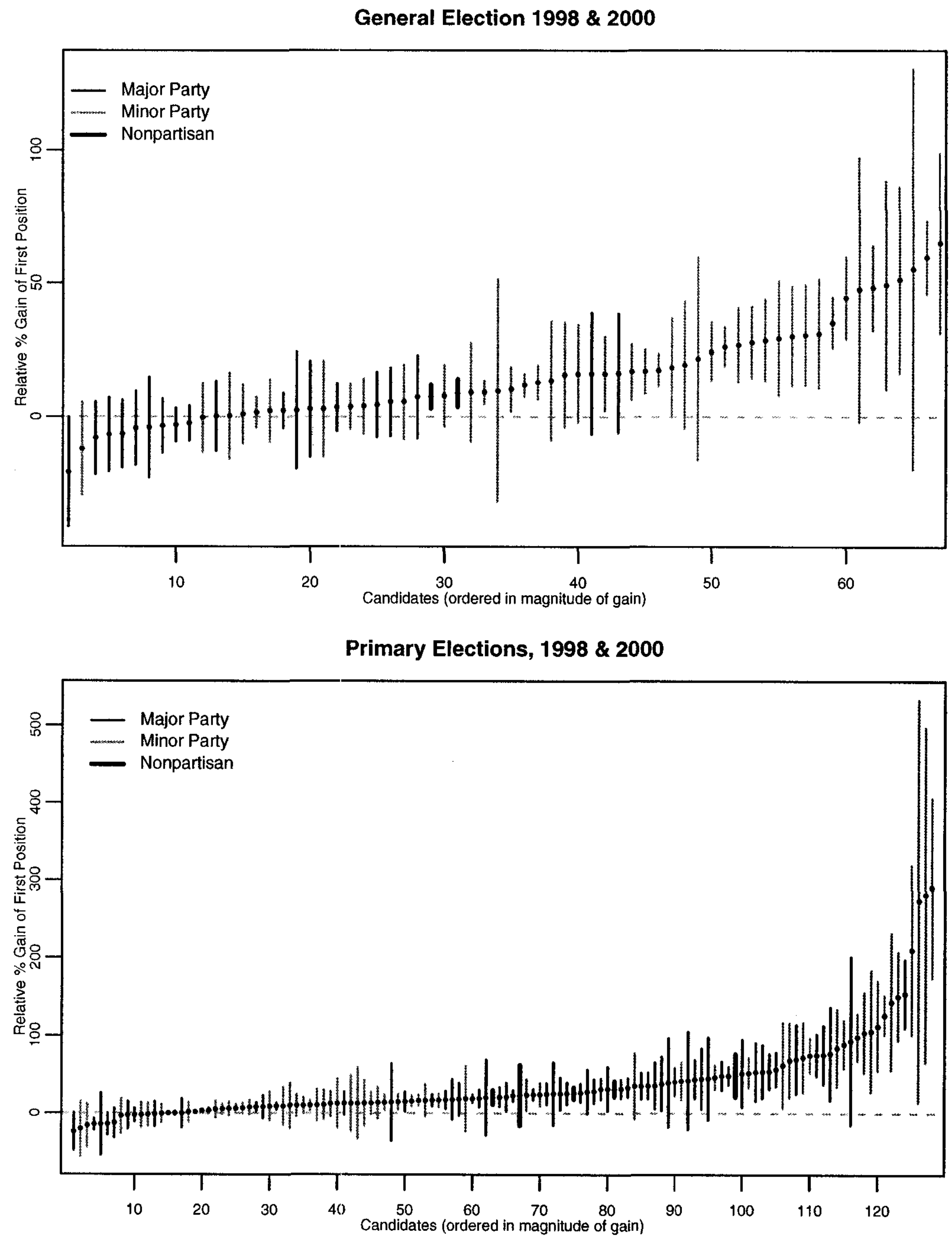


Figure 4.3: Candidate-Specific Average Relative Gain due to Being Listed in First Position on Ballots for 1998 and 2000 Elections. The top panel shows the results for general elections, and the bottom panel displays those for primary elections. Circles indicate point estimates for each candidate, and vertical bars represent estimated 95% confidence intervals. In general elections, only minor party and nonpartisan candidates are affected by the ballot order. In primary elections, however, major party candidates are also affected.

	General				Primary			
	Absolute		Relative		Absolute		Relative	
	ATE	SE	ATE	SE	ATE	SE	ATE	SE
Democratic	0.05	0.46	0.25	0.90	1.89	0.32	43.58	5.53
Republican	-0.06	0.53	-0.43	1.29	2.16	0.46	33.62	5.91
American Independent	0.16	0.02	20.83	1.39	2.33	0.15	26.76	3.55
Green	0.56	0.17	21.18	5.82	3.15	1.16	6.24	3.54
Libertarian	0.23	0.02	14.56	1.03	6.59	1.42	71.92	13.55
Natural Law	0.31	0.06	26.13	2.85	0.40	0.08	44.78	5.45
Peace and Freedom	0.28	0.03	25.49	2.15	6.31	0.53	14.75	1.43
Reform	0.26	0.07	19.57	2.23	4.11	1.56	48.45	9.66
Nonpartisan	1.95	0.30	9.21	3.31	3.44	0.78	19.42	4.05

Table 4.3: Party-Specific Average Causal Effects of Being Listed in First Position on Ballots Using All Races from 1978 to 2002. ATE and SE represent the average causal effects and their standard errors, respectively. For general and primary elections, the left two columns present the estimates of average absolute gains in terms of the total or party vote, respectively, while the right two columns show those of average relative gains. Each candidate-specific effect is averaged over different races to obtain the overall average effect for each party. In general elections, only minor party and nonpartisan candidates are affected by the ballot order. In primaries, however, the candidates of all parties are affected. The largest effects are found for nonpartisan candidates.

in general elections. For 74 of the 128 candidates, the confidence intervals do not include zero, but more importantly the ballot order affects major and minor party candidates alike, with a median relative ballot effect of roughly 21%, and a striking range of gains across candidates. The median absolute gain is roughly 1.6% of the party vote. Given that primary races have a much larger number of candidates, it is notable that the absolute gain is larger than for general elections (see also Section 4.4.4).

Averaging over all the races from 1978 to 1992, Table 4.3 summarizes the estimated ballot effects.[19] The rough patterns of the 1998 and 2000 elections hold across all elections

is largely uninformative, we conduct inferences with both the minimum MSE variance estimator and the random estimator, with no substantive difference in results.

[19]In cases where multiple candidates from the same party or multiple nonpartisan candidates contested

studied. In general elections, major party candidates exhibit no discernible ballot order effect, while the effect on minor party candidates is substantial. Minor party candidates typically gain from 15 to 30% of their baseline vote share in general elections. Given that minor party candidates generally receive only a small proportion of the vote, however, this amounts to an average absolute gain of roughly 0.2 to 0.6% of the total vote cast.

Testable propositions deriving from partisan cue theory would predict that cognitive biases such as ballot effects should be most prominent for nonpartisan races, independent candidates, and primary races, since party labels are least informative in such races. These predictions bear out consistently in our results. Independent and nonpartisan candidates gain 2.4% of the absolute vote share when listed first, and when the office itself is nonpartisan, candidates gain roughly 3.3% of the total vote share when in first. This magnitude difference is consistent with the notion that more information about candidate policy preferences is conveyed in races where at least some candidates are partisans. On the other hand, since the only nonpartisan office in our dataset is the Superintendent of Education, we cannot determine whether larger cognitive biases might stem from lack of partisan labels, lower prominence of the office, or both.

In primaries, where the least information is conveyed by party affiliation, ballot order affects all candidates. Both Democratic and Republican candidates gain roughly one to two percent of the party vote when in first position, which constitutes a relative gain of roughly

---

the election, such as in primaries or nonpartisan elections, the simple average of those candidate-specific point estimates and standard errors are used to obtain an estimate for each race, and these estimates are then averaged across elections with the number of candidates in each race as weights.

30%. Since the number of candidates is generally much larger in primaries, with, for example, five Republican and six Democratic candidates running for the gubernatorial party nomination in 1998, this does not mean that the effect is confined to minor candidates in the major parties. To the contrary, many of major Democratic and Republican candidates are affected by ballot order. From Michael Huffington in 1994 (ATE=4.5, SE=1.4), to Barbara Boxer in 1998 (ATE=2.7, SE=0.7), to Dianne Feinstein in 2000 (ATE=1.5, SE=0.6), to Gary Mendoza in 2002 (ATE=2.7, SE=0.8), we observe a robust effect of ballot order on all types of candidates.

To provide another example, in the race for the Republican nomination for Lieutenant Governor in 1998, the absolute effect for Tim Leslie, who won the nomination by 10% of the vote, was borderline significant but substantial at 11% (SE=6.8), and the effect on the runner-up, Richard Mountjoy, was 9% (SE=2.2). In the Democratic Controller primary in 2002, both candidates for Controller, Steve Westly and Johan Klehs, gained roughly 6% (SE=1.0) of the vote when listed first. In Section 4.4.4, we further analyze the potential effects that non-randomized ballot order could have on who wins a race.

Minor party candidates in primaries receive anywhere from 6 to 70% of their baseline vote share, with Libertarian and Reform party candidates exhibiting the largest relative gains. Nonpartisan candidates gain roughly two to six percent of the total vote when listed first, which does not differ appreciably from nonpartisan gains in general elections or gains by other candidates in primaries. Given that partisan labels are relatively uninformative in primaries, where there are often multiple party candidates running, this result is not surprising in light of partisan cue theory.

Tables 4.4 and 4.5 present estimated average absolute and relative gains broken down by office and party, respectively. In both general and primary elections, no discernible patterns emerge with respect to the prominence of the office, or to the order in which the office appears on the ballot. The only exception is the Superintendent of Education, which is a nonpartisan race.

Appendix 4.9 presents a host of other conditional effects and yields further insight into various behavioral models of ballot order effects. First, one might expect ballot order effects to be smaller in non-incumbent races, since incumbency may act as an informational cue to voters. Incumbents are denoted on California ballots, which provide current employment descriptions for all candidates. While we find few differences for incumbent and open races in general elections, in primaries open seat races appear to be associated with larger ballot order effects (see Table 4.11). Second, we test the degree to which ballot order effects are driven by small uninformed groups of voters who turn out only for the prominent races. We do this by examining on-year versus off-year (or midterm) elections. Since contested offices differ in on-year and off-year elections with the exception of US Senate elections, we examine Senate results. The finding shows that the ballot order effect for on-year elections is generally larger (see Table 4.12). In particular, Democratic candidates in on-year general elections gain roughly two percentage points when listed first, while exhibiting no gains at all in off-year elections.

Lastly, we investigate the magnitude of ballot order effects conditional on the number

	General Elections									
Party	President	Senate	Governor	Lt. Gov.	Atty Genl	Controller	Ins. Comm.	Sec. State	Treasurer	Supt Educ
Democrat	1.1	0.7	0.2	−1.1	−0.7	−1.9	0.2	−3.0	0.4	
	(1.0)	(0.7)	(1.0)	(3.0)	(1.4)	(2.0)	(1.5)	(2.8)	(1.7)	
Republican	−0.8	−0.6	1.5	2.2	−0.7	−5.0	1.5	2.6	−2.0	
	(1.2)	(0.9)	(1.1)	(2.7)	(1.6)	(2.1)	(2.3)	(3.0)	(2.3)	
Amer. Indep.	0.1	0.2	0.1	0.1	0.3	0.1	0.1	0.4	0.2	
	(0.0)	(0.0)	(0.0)	(0.1)	(0.1)	(0.1)	(0.0)	(0.1)	(0.0)	
Green	0.1	0.9	0.4	1.0	0.8	−0.5	0.2	1.3	0.5	
	(0.4)	(0.5)	(0.4)	(0.7)	(0.4)	(0.4)	(0.3)	(0.8)	(0.3)	
Libertarian	0.0	0.2	0.4	0.2	0.3	0.0	0.5	0.6	0.2	
	(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(0.1)	(0.1)	(0.1)	(0.1)	
Natural Law	0.0	0.1	0.1	0.0		0.2	0.6	0.7	0.5	
	(0.0)	(0.0)	(0.0)	(0.0)		(0.2)	(0.3)	(0.1)	(0.3)	
Peace & Frdm	0.1	0.4	0.2	1.1	0.1	0.3	0.2	0.5	0.0	
	(0.0)	(0.1)	(0.0)	(0.2)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	
Reform	0.3	0.1		0.3		0.1		0.3		
	(0.3)	(0.0)		(0.0)		(0.1)		(0.0)		
Nonpartisan	0.4		0.0							4.0
	(0.4)		(0.4)							(0.5)
	Primary Elections									
Party	President	Senate	Governor	Lt. Gov.	Atty Genl	Controller	Ins. Comm.	Sec. State	Treasurer	Supt Educ
Democrat	1.6	1.5	0.6	5.6	4.6	3.3	3.6	2.4	7.1	
	(2.5)	(0.5)	(0.5)	(2.8)	(2.0)	(1.0)	(1.4)	(1.1)	(1.8)	
Republican	−0.9	2.8	0.6	5.5	4.8	2.1	3.2	2.8	3.1	
	(1.6)	(1.0)	(0.4)	(2.7)	(1.8)	(1.0)	(1.0)	(1.3)	(1.5)	
Amer. Indep.	0.0	0.1	8.6	0.4	0.4	0.2	0.0	0.8	0.1	
	(0.0)	(0.0)	(0.6)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)	(0.1)	
Green	0.9	4.6	−0.2	−0.7		6.2				
	(0.8)	(2.8)	(0.2)	(0.3)		(0.9)				
Libertarian	17.9	0.5	0.2	0.2	0.4	0.2	0.2	0.7	−0.1	
	(4.0)	(0.1)	(0.1)	(0.3)	(0.1)	(0.1)	(0.3)	(0.2)	(0.2)	
Natural Law	0.1	0.2	0.1			0.1	0.5	1.1	1.0	
	(0.0)	(0.0)	(0.0)			(0.1)	(0.2)	(0.2)	(0.6)	
Peace & Frdm		3.2	8.2	11.5	9.8	0.1	8.2	5.4	0.2	
		(0.7)	(0.8)	(3.3)	(2.0)	(0.2)	(3.3)	(1.1)	(0.2)	
Reform	5.2	5.8		0.5		0.5		0.6		
	(3.3)	(1.6)		(0.2)		(0.1)		(0.1)		
Nonpartisan										3.4
										(0.8)

Table 4.4: Average Absolute Gain due to Being Listed in First Position on Ballots using All Races from 1978 to 2002. Standard errors are in parentheses. As in Table 4.3, all candidate-specific effects are averaged over different elections to obtain the overall average effect for each office and party. In general elections, no discernible patterns emerge with respect to the prominence of the office, or to the order in which the office appears on the ballot. In primary elections, ballot order effects are sometimes larger for major offices. In both cases, nonpartisan candidates for the Superintendent of Education are significantly affected by ballot order.

	General Elections									
Party	President	Senate	Governor	Lt. Gov.	Atty Genl	Controller	Ins. Comm.	Sec. State	Treasurer	Supt Educ
Democrat	2.7	1.5	0.3	−2.2	−1.3	−3.4	0.4	−6.2	0.9	
	(2.1)	(1.4)	(2.0)	(5.6)	(2.7)	(3.6)	(3.1)	(5.8)	(3.2)	
Republican	−2.3	−1.6	3.5	6.7	−1.8	−14.5	3.1	6.4	−5.2	
	(2.8)	(2.2)	(2.7)	(7.5)	(4.2)	(5.9)	(5.1)	(7.3)	(6.0)	
Amer. Indep.	55.8	12.9	11.9	9.3	11.8	10.3	6.7	29.8	19.1	
	(5.5)	(2.2)	(2.4)	(7.3)	(3.2)	(5.0)	(2.1)	(4.2)	(2.7)	
Green	10.4	37.9	26.7	29.0	21.8	−8.3	4.7	38.6	10.5	
	(17.0)	(19.2)	(14.5)	(19.1)	(10.7)	(6.7)	(9.4)	(22.3)	(6.9)	
Libertarian	9.8	9.4	29.1	11.1	16.4	−0.2	17.6	20.7	8.1	
	(3.1)	(1.7)	(2.5)	(4.7)	(2.0)	(6.6)	(3.2)	(2.5)	(4.0)	
Natural Law	45.3	10.4	27.7	5.7		25.2	25.2	36.1	23.2	
	(9.7)	(3.4)	(4.4)	(5.7)		(7.3)	(11.0)	(4.9)	(11.1)	
Peace & Frdm	46.1	18.1	18.1	74.2	4.0	30.7	13.6	49.4	2.3	
	(7.8)	(2.7)	(3.3)	(12.6)	(5.2)	(9.1)	(11.2)	(19.7)	(6.0)	
Reform	4.8	14.1		28.7		10.4		35.2		
	(5.7)	(4.4)		(4.5)		(4.2)		(4.9)		
Nonpartisan	12.0		1.0							8.5
	(7.1)		(7.4)							(1.1)
	Primary Elections									
Party	President	Senate	Governor	Lt. Gov.	Atty Genl	Controller	Ins. Comm.	Sec. State	Treasurer	Supt Educ
Democrat	98.1	31.0	62.9	30.0	28.1	7.5	12.2	23.0	35.6	
	(49.4)	(9.5)	(11.0)	(17.5)	(9.4)	(1.7)	(4.6)	(6.9)	(7.2)	
Republican	22.5	52.0	33.8	27.1	11.2	14.1	8.9	9.4	5.7	
	(17.5)	(15.0)	(7.2)	(11.5)	(3.7)	(3.8)	(2.5)	(3.4)	(3.2)	
Amer. Indep.	14.2	30.0	29.0	54.0	28.7	13.8	1.5	48.6	13.1	
	(25.4)	(9.1)	(5.0)	(17.6)	(8.0)	(10.3)	(6.6)	(6.0)	(7.2)	
Green	12.9	7.9	−15.9	−24.0		22.4				
	(4.2)	(6.9)	(14.6)	(12.7)		(3.1)				
Libertarian	166.2	34.1	31.6	12.8	25.1	10.2	9.5	20.6	−2.1	
	(37.4)	(6.6)	(14.7)	(16.1)	(7.1)	(5.3)	(12.6)	(4.8)	(8.3)	
Natural Law	75.8	51.0	53.5			5.2	24.4	52.4	44.9	
	(19.3)	(7.6)	(19.1)			(5.4)	(7.9)	(9.9)	(27.3)	
Peace & Frdm		6.4	19.6	25.0	19.8	10.1	16.8	11.2	10.9	
		(2.2)	(1.9)	(6.9)	(4.1)	(14.9)	(7.3)	(2.3)	(9.7)	
Reform	63.2	24.3		62.3		33.5		48.5		
	(20.1)	(5.4)		(28.3)		(6.2)		(9.5)		
Nonpartisan										19.4
										(4.0)

Table 4.5: Average Relative Gain due to Being Listed in First Position on Ballots using All Races from 1978 to 2002. Standard errors are in parentheses. As in Table 4.3, all candidate-specific effects are averaged over different elections to obtain the overall average effect for each office and party. In general elections, no discernible patterns emerge with respect to the prominence of the office, or to the order in which the office appears on the ballot. In primary elections, ballot order effects are sometimes larger for major offices. In both cases, nonpartisan candidates for the Superintendent of Education are significantly affected by ballot order.

of candidates. This addresses two competing behavioral models of ballot order effects, one positing that evaluating each additional candidate entails some cognitive cost, and the other positing that the first positions solves a coordination problem between voters (e.g., Forsythe et al., 1993; Mebane, 2000). The cognitive cost model implies monotonically increasing ballot effects in the number of candidates, while the latter provides a unclear prediction when the number of candidates is greater than two. We find that ballot order effects roughly increase monotonically in the number of candidates, lending credence to the cognitive cost model (see Table 4.13).

4.4.4 Margin of Victory and Ballot Order Effect

To get a sense of the substantive size of these estimated effects, Figure 4.4 plots the estimated ballot order effect of the second-highest vote-getter of each race against the margin of victory. The margin of victory is defined as the difference in vote shares between the winner and the second-highest vote-getter in a race.[20] Thick confidence intervals indicate that they include or exceed the margin of victory. The figure underscores the fact that the substantive effect of ballot order on election outcomes hinges largely on how close the races are. In general elections, as suggested by our previous results, we find no conclusive evidence of ballot order effects on major candidates. In contrast, ballot order effects were significantly positive and possibly greater than the margin of victory in 7 of 59 primary races. This indicates, for example, that ballot order might potentially have changed the

[20]Note that for primaries, we define a race here as a competition for the nomination for the party nomination.

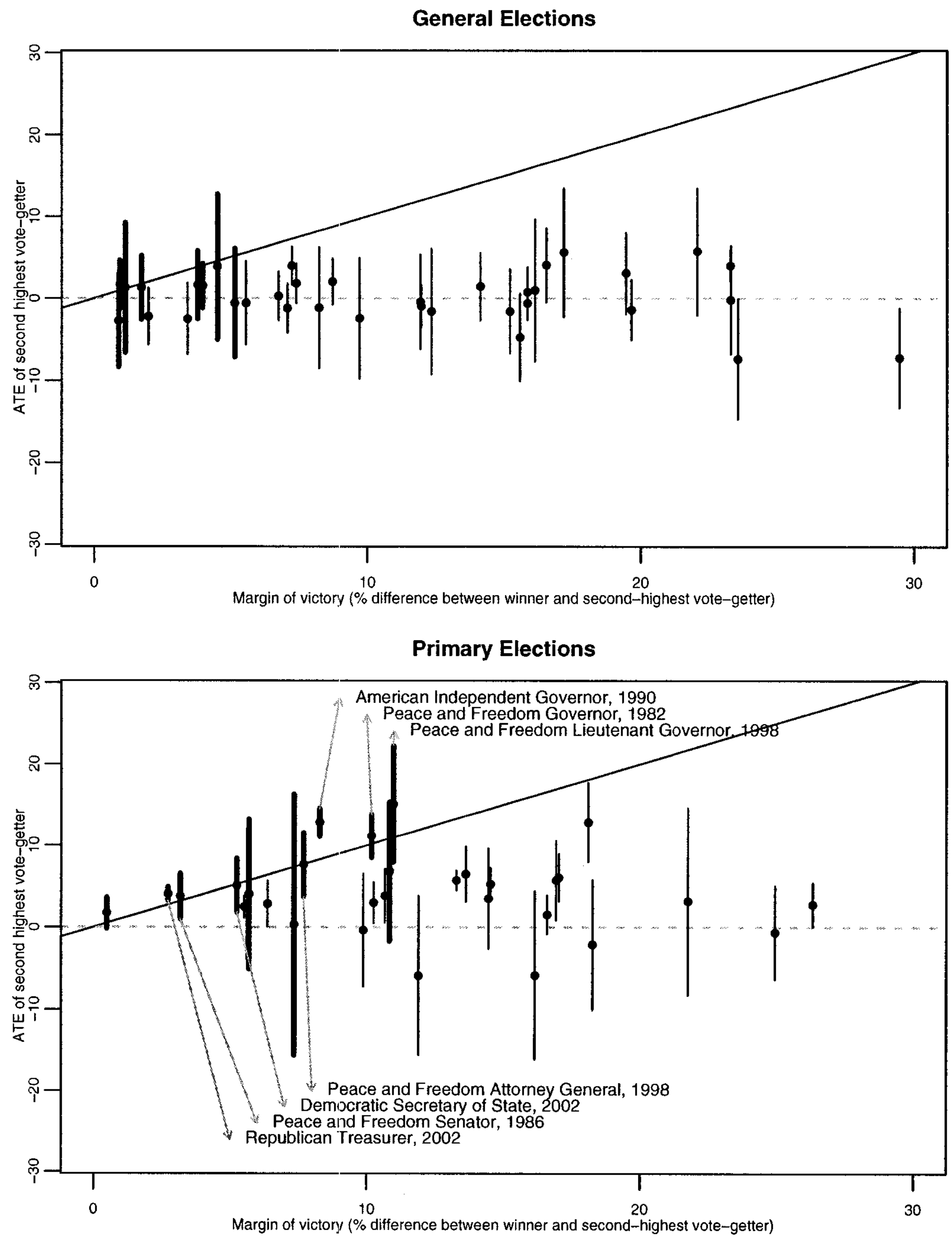


Figure 4.4: Comparison of Estimated Average Ballot Order Effect for Second-highest Vote-getter and Margins of Victory from 1978 to 2002. The top panel shows general elections, and the bottom panel represents the primary elections. Circles indicate the point estimate for the (absolute) average ballot order effect whereas vertical bars represent 95% confidence intervals. The 45 ° lines represent the instances where the ballot order effect equals the margin of victory. Thicker intervals indicate the races where the margin of victory is included in or below the 95% confidence interval. The figure implies that the outcomes of four primaries might have been different if the candidates were listed differently on ballots.

winner of the Democratic primary for the office of Secretary of State or the Republican primary for the office of Treasurer in 2002 *if ballot order were determined differently.*

4.5 Alternative Approaches

In this section, we present two alternative approaches to the estimation of ballot order effects by relaxing some of the assumptions made in the analysis of the previous section. Fewer assumptions mean that these alternative methods provide more limited inferences. In particular, it is difficult to obtain efficient estimates of candidate-specific ballot order effects. Nonetheless, the alternative approaches test the sensitivity of our main results to the assumptions that are made to identify the causal effect of ballot order.

4.5.1 A Multinomial Model

The analysis of this paper so far does not explicitly model how the ballot order effect of one candidates affects the vote share of the remaining candidates. Although such a multinomial modeling approach is optimal, as the number of candidates increases it becomes more difficult to efficiently estimate the ballot order effects for all candidates at once.

Nevertheless, it is possible to apply a multinomial approach to the California alphabet lottery by combining the results from multiple races in several elections. This allows us to obtain party-specific estimates of the ballot order effect for general elections and make aggregate inferences. We analyze all races from 13 general elections using the overdispersed multinomial logit model (McCullagh and Nelder, 1989, pp.174–175). The response variable is a vector of vote counts for each candidate in an assembly district, and we allow for overdispersion that takes into account the clustering of votes within each district (Mebane and Sekhon, 2004).

We consider both constrained and saturated multinomial logit models using a binary treatment variable for each candidate indicating whether that candidate was listed first on the ballot in a particular district. In the constrained model, we assume that the ballot order effect of a candidate does not affect the relative vote shares among the other candidates.[21] In the saturated model, we relax this assumption such that the ballot order effect of one candidate can affect other candidates disproportionally.[22] Although the saturated model is more general, it requires $(J-1) \times (J-2)$ more parameters to be estimated, yielding very inefficient estimates when the number of candidates, J, is large. After fitting the two models, we compute the estimated average (relative) ballot order effects by simulating model parameters from their asymptotic multivariate normal distributions.[23] Then, we average these results across different races and elections for each party to obtain the estimated average party-specific ballot order effects.

Table 4.6 presents the estimates from the two models. Though standard errors are larger, the results are in a substantial agreement with our main estimates of Table 4.3. In particular,

[21] This is commonly termed as the independence of irrelevant alternatives (IIA) assumption (McFadden, 1973).

[22] Although the IIA assumption is relaxed in terms of ballot order effects, the saturated model unlike the multinomial probit model still maintains the IIA assumption for the baseline vote share (Imai and van Dyk, 2004).

[23] These models are fit via VGAM package (version 0.5-17) written by Thomas Yee (Yee and Wild, 1996).

	Constrained Model				Saturated Model				
	Absolute		Relative		Absolute		Relative		
	ATE	SE	ATE	SE	ATE	SE	ATE	SE	Races
Democratic	-0.07	0.59	0.43	1.27	-0.07	0.64	0.46	1.30	33
Republican	0.06	0.64	0.35	1.58	-0.27	0.66	-0.09	1.72	33
American Independent	0.13	0.09	30.42	19.49	0.16	0.11	64.30	68.33	33
Green	0.49	0.30	17.00	8.39	0.70	0.34	35.93	14.55	13
Libertarian	0.19	0.11	16.36	8.18	0.20	0.13	26.39	12.60	31
Natural Law	0.17	0.11	32.86	37.78	0.29	0.18	87.54	179.60	15
Peace and Freedom	0.24	0.09	30.55	13.64	0.24	0.13	54.76	33.76	23
Reform	0.08	0.15	4.28	8.98	0.25	0.24	36.88	27.76	9

Table 4.6: Party-Specific Average Causal Effects of Being Listed in First Position on Ballots in 13 General Elections using the Multinomial Logit Model. ATE and SE represent the average causal effects and their standard errors, respectively. For both constrained and saturated models, the left two columns present the estimates of average absolute gains, while the right two columns show those of average relative gains. Each candidate-specific effect is averaged over different races (different offices and elections) to obtain the overall average effect for each party. The last column indicates the number of such races.

the point estimates from the saturated model, though very inefficient, are close to those of the main results, supporting the validity of our main analysis. Both models also confirm the finding that in general elections, only minor party candidates benefit from being listed first on ballots.

4.5.2 Distribution-Free Randomization Inference

The multinomial approach in the previous section was based on a particular parametric model, i.e., the multinomial logit model. Here, we further relax this assumption and conduct distribution-free (or non-parametric) randomization inference, by extending Fisher's exact test to systematic treatment assignment (Fisher, 1935). The advantage of Fisher's exact test is that given the choice of a particular test statistic, it only requires knowledge of the treatment assignment mechanism. The test is distribution-free in the sense that it does not specify a particular distribution for potential vote shares. On the other hand, the disadvantage of applying the Fisher test to our data is that the number of candidates and the number of races largely determine its power. Therefore, one cannot analyze candidate-specific effects in our data, limiting our inferences by party and office to major offices on which data from multiple elections are available.

Since the treatment assignment is known in the California alphabet lottery, we derive the exact distribution of any test statistic under the null hypothesis of no ballot effects. In particular, our test statistics consist of average absolute and relative gains of being listed first on ballots, defined in Section 4.4.3, averaged across different elections for each party and office. Since there are J_i ways in which a candidate could have been placed first in each race i (where J_i denotes the number of candidates in that race), the power of the test is then determined by the permutation of treatment assignments pooling across elections, i.e., $\prod_i J_i$.

Fisher's exact test consists of three steps. First, we assert the null hypothesis of no ballot effect. Under this hypothesis, all candidates would have received the same vote share in each district *irrespective* of their ballot position. For example, in the 1998 gubernatorial general election, Gray Davis was listed second in the first Assembly District. Under the null hypothesis, Davis would have received the same vote share in this district even if he were listed in any other position (e.g., third). Second, we calculate the distribution of the test statistic under this null hypothesis. Under systematic random assignment of

Definition 2, there exist J_i possible combinations for listing each candidate first across Assembly Districts, stemming from the randomized order of the first Assembly District. In the 1998 gubernatorial election, there were seven possible ways in which Gray Davis could have been listed in the first Assembly District and then rotated through the rest of the districts. Considering all Democratic candidates for governor in general elections as a series of repeated experiments, the $\prod_i J_i$ possible combinations of treatment assignment yield the exact distribution of the test statistic.

The final step is to compare the observed values of the test statistics with these reference distributions and compute the one-tailed p-value. The p-value represents how rare such outcomes are under the null hypothesis of *zero* ballot order effects. Table 4.7 presents the observed test statistics, associated p-values, and permutations for major offices in general elections. In 13 of the 17 tests for minor party candidates contesting for the US Presidency, Senate or Governor's Office, the p-value was less than 0.05, with large relative gains ranging from roughly 10 to 50%. Little evidence suggests that major parties are affected, although surprisingly Republican candidates in gubernatorial races may have gained roughly 2% of the absolute vote share when in first place. These results are largely consistent with our main results of Section 4.4.[24]

[24] We also conducted Fisher tests pooling candidates within each race, using the rank-sum $\sum_{k=1}^{K} R(Y_{\cdot k}(1))$ and log-ratio $\frac{1}{K(J-1)} \sum_{k=1}^{K} \sum_{t=2}^{J} \left| \log \frac{Y_{\cdot k}(1)}{Y_{\cdot k}(t)} \right|$ test statistics, where $Y_{\cdot k}(t)$ is the vote share in the kth assembly district for the candidate whose name appears the tth on the ballot. In this case, we have $J_i!$ possible treatment combinations. The tests yielded p-values < 0.05 for 95% and 80% of elections analyzed, respectively.

		Absolute Effects		Relative Effects		
Office	Party	ATE	p-value	ATE	p-value	Permutations
US President	Democrat	0.40	0.37	1.00	0.35	8400
	Republican	−0.75	0.73	−2.41	0.77	8400
	Amer. Indep.	0.11	0.00	52.46	0.00	8400
	Green	0.14	0.39	10.37	0.34	56
	Libertarian	0.07	0.00	12.00	0.00	8400
	Natural Law	0.05	0.02	45.32	0.02	56
	Peace & Frdm	0.11	0.00	49.29	0.00	240
	Reform	0.33	0.09	4.82	0.20	56
US Senate	Democrat	0.70	0.31	1.41	0.29	918750
	Republican	−0.95	0.75	−2.78	0.78	918750
	Amer. Indep.	0.23	0.01	15.85	0.00	918750
	Green	0.91	0.02	37.93	0.02	42
	Libertarian	0.13	0.03	7.95	0.02	918750
	Natural Law	0.06	0.10	10.38	0.10	49
	Peace & Frdm	0.37	0.04	18.66	0.00	131250
	Reform	0.14	0.04	14.07	0.04	49
Governor	Democrat	0.22	0.41	0.27	0.46	26250
	Republican	1.96	0.04	4.69	0.04	26250
	Amer. Indep.	0.11	0.00	9.63	0.01	26250
	Green	0.42	0.19	26.68	0.10	42
	Libertarian	0.38	0.00	30.62	0.00	26250
	Natural Law	0.12	0.02	27.69	0.02	42
	Peace & Frdm	0.19	0.00	18.84	0.01	4375

Table 4.7: Fisher's Exact Test by Party and Offices. When the number of permutations exceeded 30,000, we simulated the distribution under the null hypothesis of no ballot order effects to obtain the p-values. In general elections, only minor party candidates benefit from ballot order effects.

4.6 Policy Implications for Ballot Reform

If one of the primary goals of election law is to provide equal opportunity to candidates, our findings suggest that election officials may want to randomize the name order of candidates on ballots and minimize ballot order effects. In this section, we first outline the costs and benefits of randomization. Second, we conduct a cost-effectiveness analysis and show that randomization of ballot order is more cost-effective at reducing election day bias than currently proposed voting reforms by more than a factor of 100. The basic intuition is that since randomization requires no substantial financing of new voting equipment, it is by far the most cost-effective way to reduce voting bias. Lastly, we examine the range of statutes governing ballot order across the fifty states to suggest potential avenues of reform.

4.6.1 Cost and Benefits of Randomization

The primary benefit of randomization lies in improving the fairness of elections.[25] Randomization would improve the fairness of elections for major and minor parties alike. Our study shows that non-random ballot order disproportionately benefits one candidate in virtually all primaries. While in general elections randomization is unlikely to change outcomes, in twelve percent of primary races examined, ballot order might have changed the winner. Randomization would therefore help all parties, including the Democratic and

[25] Without purporting to answer larger and vitally important philosophical questions of fairness, we concur that "[u]nder any reasonable standard of fairness, ballot format should not determine the outcome of an election." Jonathan N. Wand, Michael C. Herron and Henry E. Brady, "Ballot Cost Gore Thousands of Votes," San Diego Union-Tribune, Nov. 19, 2000, at G-3.

Republican parties, nominate the most preferred candidates for the general election.

That said, various randomization methods differ in the effectiveness of reducing ballot order effects. Statistical theory clearly predicts that the bias of ballot order effects will decrease in the number of units across which the ballot is randomized. To truly reduce ballot order effects to zero, states may want to conduct randomization across smaller units, such as counties or precincts. In the future, the advancement of electronic voting technology might even allow the name order of candidates to be randomized separately for each individual voter. On the other hand, in practice, randomizing across every precinct, or every voter, may not be feasible. We therefore suggest principal cost criteria by which states might decide how to implement cost-effective randomization.

The costs of randomization are fourfold. First, election officials may incur the *administrative costs* in conducting a randomized drawing of the ballot order. This includes concerns over added complexity of randomization and risking mistakes in electoral administration. With the aid of modern computer technology, however, the cost of drawing a random alphabet itself is minor. To make this process transparent, administrative costs may additionally entail publicizing the event, as in California (see Section 4.3.1). Second, election officials may incur marginal *printing costs* for randomized ballots. This presents a tradeoff: as the number of randomizations increases, ballot order effects decrease but printing costs increase. On the other hand, precincts and counties already print out specific versions of ballots due to local offices and issues. California, for example, currently prints roughly 18,000 styles of ballots for a general election and over 25,000 styles for primaries.[26] As a result, the marginal printing costs of randomization are relatively small. In fact, as we show in Section 4.6.2, these printing costs are miniscule compared to all existing voting reform efforts.

The third cost is in *voter confusion.* Some argue that alphabetical or partisan ballots permit voters to locate their preferred candidate more efficiently than randomized ballots. Randomization might also disproportionately harm some parties over others. Candidates, for example, could no longer campaign on specific ballot positions (e.g., "Vote No. 3 on the Ballot") and voting a straight party tickets may be more difficult. Yet we find that the magnitude of ballot order effects for Democratic and Republican candidates in primaries is substantially the same, providing little evidence for disproportionate effects among the major parties. Rather than listing candidates in some defined order to ease locating of candidates, officials may explore alternatives to make the identification of candidates easier. As the California Supreme Court suggested, listing incumbent status and party, for example, allows voters to identify candidates even when randomizing the ballot order.

Lastly, randomization may entail costs in *governmental stability.* This is premised on the argument that a government objective of promoting stability may reasonably justify listing majoritarian parties and incumbents first. This rationale is most directly at odds with the fairness benefit of randomization. Furthermore, the argument against partisan orders does not apply directly to primaries, where major parties are often nominating candidates within their own party. And the argument for incumbency was explicitly refuted by the California

[26]We thank Melissa Warren and Joanna Southard at the Elections Division, Secretary of State, for this insight.

Supreme Court, which "emphatically reject[ed] the notion that the government may consciously choose to favor the election of incumbents over non-incumbents ... distort[ing] the preferences of participating voters" (*Gould v. Grubb*, at 673). Based on the above tradeoffs, states will have to decide whether to adopt and how to implement randomization. Next, we show that current reform efforts that focus largely on voting equipment appear inconsistent with a principle of cost-effective reduction of voting bias.

4.6.2 The Cost-Effectiveness of Randomization

A primary goal of voting reform is to reduce voting bias due to arbitrary factors such as voting machines and ballot formats. Following the 2000 presidential election, the most significant reform effort has focused on updating old voting equipment to optical scan or electronic voting machines in order to reduce the number of undervotes. In this section, we assess the cost-effectiveness of randomizing ballot order relative to various reforms by calculating the cost of reducing voting bias per vote. Cost-effectiveness analysis "is designed to compare a set of regulatory actions with the same primary outcome" (US Office of Management and Budget, 2003, p.128) (here, one biased vote), to assess the efficiency of policy options. Accordingly, we estimate the dollar amount required for each reform measure to reduce voting bias by one vote.[27]

Table 4.8 presents such estimated costs (per biased vote) of extant ballot reforms and

[27] Note, of course, that this is inherently different than comparing the total costs or the cost per voter of reform measures. Only looking at such costs ignores the benefit side of reform, thereby failing to assess cost-effectiveness entirely.

Source of Voting Bias	Magnitude (%)	Proposed Reform	Dollar Amount Spent to Eliminate One Biased Vote
Ballot Order	2.03	Randomized Rotation	
		by Assembly District	0.19 – 0.38[a]
		by City and Town	7.30[b]
Voting Equipment[c]			
Lever Machines	1.88	Replacement with	
		Optical Scans	105.89
		Electronic Voting Machines	473.56
Paper Ballots	1.89	Replacement with	
		Optical Scans	103.91
		Electronic Voting Machines	456.90
Punch Cards	2.67	Replacement with	
		Optical Scans	44.37
		Electronic Voting Machines	129.47

Table 4.8: Cost Effectiveness of Selected Ballot Reforms. The magnitude of potential bias for ballot order represents the estimated ballot order effect from our analysis of California primary elections. For voting equipment, the figures are based on the estimated undervote rate obtained from the United States General Accounting Office. The estimated cost per ballot represents dollar amount one needs to spend in order to reduce bias by one ballot.

[a]The figure is based on California's primary system on the estimated cost obtained through the authors' interview with an election official of Sonoma County Registar of Voters for the 2004 Presidential Primary.

[b]The figure is based on New Hampshire's primary system and the estimated cost given by the Secretary of State Office for the 2004 Presidential Primary. See Dan Tuohy, "Alphabetical Ballot Puts Dean at Top," Eagle Tribune, Feb. 12, 2003.

[c]The magnitude of bias is based on the estimated undervote rate reported by US General Accounting Office (2001). The undervote rates for optical scans and electronic voting machines are 1.32 and 1.59%, respectively. The cost calculation is based on the estimates reported by Caltech and MIT Voting Technology Project (2001).

compares them with the cost of randomization. Although existing estimated costs of voting reform and randomization are somewhat rough, the table illustrates the relative cost-effectiveness of randomization. We take 2.03% as the estimate of the ballot order effect, which is the average of point estimates for Democratic and Republican candidates in primary elections. For the estimated marginal cost of randomization, we conducted a series of interviews. While on several occasions, it was asserted to us that the costs are effectively zero, here we use the marginal cost of printing different ballots to provide conservative estimates. Specifically, we use an estimated cost of randomizing across Assembly Districts that is given by a California County Registrar of Voters as well as the estimated marginal cost of randomization across cities and towns that is given by New Hampshire Secretary of State Office.[28] To obtain the dollar amount necessary to reduce voting bias by one vote, we divide the marginal costs over the estimated number of voters who voted for the first candidate solely due to ballot order.

We estimate that California spends roughly 20 to 40 cents to eliminate a biased vote due to ballot order.[29] When the units of randomization are small, the cost is higher: New Hampshire will spend about $7 to eliminate one biased vote if they decide to randomize ballot order by town and cities. The second column of Table 4.8 gives the estimates of undervote rate for different voting machines in use, which are based on the report by US General Accounting Office (2001, Table 1). Again, as in the calculation of the cost-effectiveness of randomization, we use these figures to calculate the dollar amount spent to eliminate one biased vote, where a biased vote is defined for voting equipment as a vote that remains uncounted solely due to voting equipment. For example, switching from punch cards to optical scan would reduce the undervote rate by about 1.35 percentage points. Given the cost of obtaining new equipment, this amounts to approximately $44 for optical scans and

[28]Both figures are for 2004 primary election. New Hampshire's estimate stems from 2003 state law that repudiated randomized rotation.

[29]Note that "elimination" here should be interpreted as uniformly distributing votes that are determined by ballot order across all candidates.

$130 for electronic voting machines.[30]

Table 4.8 clearly demonstrates the relative cost-effectiveness in reducing voting bias of randomization. While states may spend anywhere from $40 to $470 to reduce the bias of one hanging chad, less than a dollar could be invested to reduce the bias of ballot order by randomizing across units that are size of California Assembly Districts. Although our estimates do not account for other dimensions of benefits and costs of reforms and are based only on rough estimates, the basic reason for the relative cost-effectives of randomization is transparent: equipment reform entails large acquisition costs of expensive machines, whereas randomization does not. Our analysis strongly suggests that the focus of current reform efforts can be altered to achieve more cost-effective voting reform.

4.6.3 Existing Policies and Possible Reforms

Depending on how states choose to address the tradeoff in costs and benefits of randomization, many potential areas and intermediate steps for reform may exist. Table 4.9 summarizes ballot order statutes for gubernatorial general elections in all the fifty states, from statutory research and interviews with election officials.[31] We classify these statutes

[30]The Caltech and MIT Voting Technology Project (2001, p.52) estimated the costs of updating the old equipments to optical scan and electronic voting machine to be $0.60 and $1.40 respectively *for every voter*, such that the marginal costs of reducing one biased vote with optical scans, for example, equals 0.60/0.0135. These figures are marginal cost estimates, assuming equipment acquisition costs are borne over the equipment's lifespan.

[31] The sources are as follows. Ala. Code 17-8-4 (2003); Interview with Vicki Balogh, Director of Elections, Office of Secretary of State, Alabama, Aug. 19, 2003; Alaska Stat. 15.15.030 (2003); Ariz. Rev. Stat. 16-502 (2003); Ark. Code Ann. 7-5-208 (2003); Cal. Elec. Code 13112 (2003); Colo. Rev. Stat. 1-5-404 (2002); Conn. Gen. Stat. 9-249a (2003); Del. Code Ann. tit. 15 4502 (2003); Fla. Stat. 101.151

Type	Specific Rule	Adopting states
Partisan	Incumbent first, then alphabetical	MA
	Partisan vote share, then alphabetical	AZ, CT, GA, MI, MN[a], NH, NC, PA
	Partisan vote share, then order of certification	FL, IN, MO
	Partisan vote share, then discretionary	NY, WI
	Partisan vote share, then randomized	WY, KY, RI
	Main parties first, then alphabetical	DE, MD, TN, UT, VA[b]
	Main parties first, then rotational	NE
	Main parties first, then randomized	TX
	Partisan rotational	OK, SC
	Discretionary	ID[c], MS[d]
Alphabetical	Alphabetical only	AL[e], CO, HI, LA, ME, NV, VT
	Alphabetical rotational	OH, IA, KS, MT
Randomized	Randomized non-rotational	AK, AR, NJ[f], NM, OR, SD, WA, WV
	Randomized rotational	CA, ND[g]
Other	Order of filing, ties resolved randomly	IL

Table 4.9: Types of Ballot Order Rules for State Gubernatorial Races in General Elections as of 2003. Only two states, California and North Dakota, randomize the ballot order and have some variation across electoral districts. In all the other states, the ballot order is likely to favor certain candidates.

[a] Candidates with *lowest* partisan vote share listed first

[b] Parties randomized

[c] Secretary of State

[d] Officer charged with printing ballot

[e] Parties alphabetized first, then independents alphabetical

[f] Randomized by party within county

[g] Randomized separately in each of 53 counties, and then rotated through precinct ordered by total votes cast for governor in the last election

by the primary determinant of the ballot order, namely partisan, alphabetical, and randomized types.

Twenty-eight states adopt partisan rules that may introduce the largest bias due to ballot order. Partisan rules hold that incumbents, major party candidates, or candidates with the largest vote shares in previous elections be listed first.[32] Connecticut is representative of these states, providing that the "names of the parties shall be arranged" by the order of "[t]he party whose candidate for Governor polled the highest number of votes in the last–preceding election" (Conn. Gen. Stat. 9-249a(1), 2003). While substantial variation in the specific implementation exists, particularly in how independents are listed, in all of these jurisdictions incumbents and majoritarian candidates benefit from ballot order effects.

(2002) Ga. Code Ann. 21-2-285 (2002); Haw. Rev. Stat 11-115 (2003); Idaho Code 34-903 (2003); 10 Ill. Comp. Stat. 5/7-12 (2003); Ind. Code Ann. 3-11-2-6 (2003); Iowa Code 49.31 (2003); Kan. Stat. Ann. 25-610 (2002); Ky. Rev. Stat. Ann. 118.225 & 118.215 (2002); Interview with Patsy Casey, Executive Staff Adviser for Election Division, Kentucky, Aug. 20, 2003; La. Admi. Code tit. 18 551 (2003); Code Me. R. tit. 21 601 (2003); Md. Code Ann., Elec. Law 9-210 (2002); Mass. Gen. Laws ch. 54 41A (2003); Mich. Comp. Laws 168.703 (2003); Interview with Brad Wittman, Director of Information Services, Secretary of State, Michigan, Aug. 19, 2003; Minn. Stat. 204D.13 (2002); Miss. Code Ann. 23-15-367 (2003); Mo. Code Regs. Ann. tit. 9, 115.239 (2003); Mont. Code Ann. 13-12-205 (2002); Neb. Rev. Stat. Ann. 32-814 (2002); Nev. Rev. Stat. Ann. 293.267 (2003); N.H. Code Admin. R. Ann. 656:5 (2002); N.J. Admin. Code tit. 19 19:14-12 (2003); N.M. Stat. Ann. 1-10-8.1 (2003); N.Y. Elec. 7-116 (McKinney 2003); N.C. Gen. Stat. 163-165.6 (2003); N.D. Cent. Code, 16.1-11-27 & 16.1-06-05 (2003); Ohio Rev. Code Ann. 3505.03 (Anderson 2003); Okla. Stat. tit. 26 6-106 (2002); Or. Rev. Stat. 254.155 (2001); 25 Pa. Stat. Ann. 3010 (2003); R.I. Gen. Laws 17-19-9.1 (2002); Interview with Mike Narducci, Clerk, Elections Division, Secretary of State, Rhode Island, Aug. 19, 2003; S.C. Code Ann. 7-13-330 & 7-13-335 (2002); Interview with Liz Simmons, Technician, Elections Commission, South Carolina, Aug. 19, 2003; S.D. Codified Laws 12-16-3.1 (2003); Tenn. Code Ann. 2-5-206 (2003); Tex. Elec. Code Ann. 52.094 (2003); Utah Code Ann. 20A-6-301 (2003); Vt. Stat. Ann. tit. 17 2472 (2003); Va. Code Ann. 24.2-613 (Michie 2003); Wash. Rev. Code 29.30.025 (2003); W. Va. Code 3-5-13a (2003); Wis. Stat. 5.64 (2002); Wyo. Stat. Ann. 22-6-121 (2003).

[32]We also include discretionary statutes of Idaho and Mississippi that allow the Secretary of State to determine the ballot order in this category, since it vests power in incumbents to determine the order. Other states that are classified as alphabetical types, such as Alabama, also alphabetize party names first, so the typology distinguishing between primarily partisan and alphabetical types may not be entirely clearcut.

Some 11 states follow alphabetical rules, differing primarily in whether the alphabetical order is uniform across the state or rotated across some sub-state unit.[33] Naturally, uniform alphabetical orders favor candidates with surnames earlier in the alphabet. Alphabetical order may, for example, disproportionately affect particular ethnic groups that have names clustered in the Roman alphabet, such as Chinese names like Zhang, Wang, or Yi. Nonetheless, alphabetical rotation may provide one solution to states. It enables voters to easily locate candidates on ballots, while roughly ensuring that all candidates are listed first an equal number of times. Of course, even then the problem remains that the relative position of candidates stays the same for most districts (and across elections), thereby still potentially favoring one candidate.

Roughly 10 states already employ some form of randomization to determine the ballot order. While this substantially reduces bias *across elections*, the majority of these states, after one randomization, maintain the same ballot order *in any particular election*. Hence, for each election, the candidate randomly selected to be listed first still reaps substantial benefits of ballot order. To reduce bias further, states may want to consider randomizations across smaller units, such as counties or precincts, and/or randomization of candidates, rather than the alphabet.

Only two states, California and North Dakota, employ randomization with different ballot orders across subunits of the states for general elections, thereby providing the smallest

[33]Ohio, for example, prints ballots in alphabetical order in the first precinct and for "each succeeding precinct, the name in each group that is listed first in the preceding precinct shall be listed last, and the name of each candidate shall be moved up one place" Ohio Rev. Code Ann. 3505.03 (Anderson 2003).

bias due to the ballot order effect among existing election practices. Nonetheless, even in these two states, there is room for improvement in reducing ballot effects. Just like the alphabetical rotational scheme, California's rotational rule may produce a relative advantage of one candidate over another in any given race.[34] Randomization across many subunits without rotation therefore might further reduce bias resulting from ballot order.

In sum, if states choose to minimize ballot bias, they have much potential to improve. The good news is that even small steps towards complete randomization, such as rotation or one-shot randomization, are likely to drastically reduce ballot order effects, at a cost that appears to be substantially lower and more effective than many other areas of ballot reform.

4.7 Concluding Remarks

Our analysis of the California alphabet lottery from 1978 to 2002 provides solid empirical ground to the study of ballot order effects, avoiding the external validity problems of laboratory experiments and the potential biases of observational studies. Free from financial, ethical and other practical constraints of field experiments, natural experiments when available provide a promising way to draw causal inferences. Our paper is the first systematic study that takes seriously the assumption of randomization and offers methods that are applicable generally to the analysis of experimental data in political science. The results of

[34]Even in North Dakota, where randomization is conducted separately within each of the 53 counties, the precincts across which ballots are rotated are arranged in the order of total votes cast for governor in the last election.

this study are largely consistent with theories emphasizing the importance of informational cues. We detect the largest ballot effects, when voters lack partisan labels on ballots, as in nonpartisan races, or when those labels cannot distinguish between multiple candidates, as in primary races. Lastly, our study provides a clear scientific estimate that may inform reform efforts to improve the fairness of electoral systems. With respect to ballot order, elections officials may want to do like California and implement a cost-effective form of randomization: to shake, not stir.

4.8 Appendix: Assessing Balance of Covariates

While randomization balances covariates in expectation across repeated experiments, in any given sample covariates might still remain imbalanced. In particular, systematic treatment assignment is susceptible to trends such as periodicity in the population, since randomization occurs only once per race. As a result, checking the balance of Assembly Districts in any one particular race remains a crucial test for the validity of inferences. Table 4.10 reports simple means tests for the covariates, taken from 1990 Census and registration data, that are widely known to be good predictors for voting behavior. In particular, we calculate mean differences and their t-statistics between districts in which Gray Davis was listed first in the 1998 gubernatorial general election and the remaining districts. In this case, the covariates appear to be balanced well, with only negligible differences in all dimensions.

To check balance of covariates for all potential treatment assignments, we calculate

	Treated	Control	T-stat
Registered Democratic	0.45	0.47	0.75
Proportion Male	0.53	0.54	1.30
Proportion Black	0.16	0.16	0.05
Proportion Latino	0.05	0.09	1.71
High school graduate	0.75	0.74	−0.13
Age (≥45 years)	0.28	0.28	−0.24
Income (≥ $40,000)	0.54	0.56	0.53
Poverty level	0.11	0.13	0.57
Urban Housing	0.91	0.93	0.33

Table 4.10: Means Tests of Selected Covariates. The table compares the 11 districts in which Gray Davis was listed first (Treated) with the remaining districts (Control) in the 1998 gubernatorial election. The covariates are selected from the 1990 Census data.

means tests for 34 covariates from the 1990 Census data and party registration reports $j = \{3, \ldots, 8\}$ corresponding to the range of observed number of candidates running.[35] The covariates include major potential confounding variables such as gender, race, education, income, urbanization, unemployment, industry, poverty levels, and party registration for the major seven parties recognized in California. Since for each j, there are j possi-

[35]The following covariates are taken from the 1990 Census and Voter Registration reports: proportion male, proportion latino / white / black (others as base), proportion of adult population graduated highschool / have associate or college degrees (less than high school as base), proportion ages 18-24 / 25-34 / 35-44 / 45-54 / 55-64 / 65 and over (12-17 as base), proportion with household income $0-9,999 / $10,000-19,999 / $20,000-29,999 / $ 30,000-39,999 / $ 40,000-49,999 / $ 50,000-99,999 / $100,000-149,999 ($150,000 or more as base), proportion of housing in urban environment, proportion of male / female labor force unemployed, proportion of industry in agriculture / manufacturing, proportion of population in poverty status, and proportion of voters registered with the Democratic / Republican / American Independent / Green / Libertarian / Natural Law / Reform parties and proportion registered with other parties or declining to state registration.

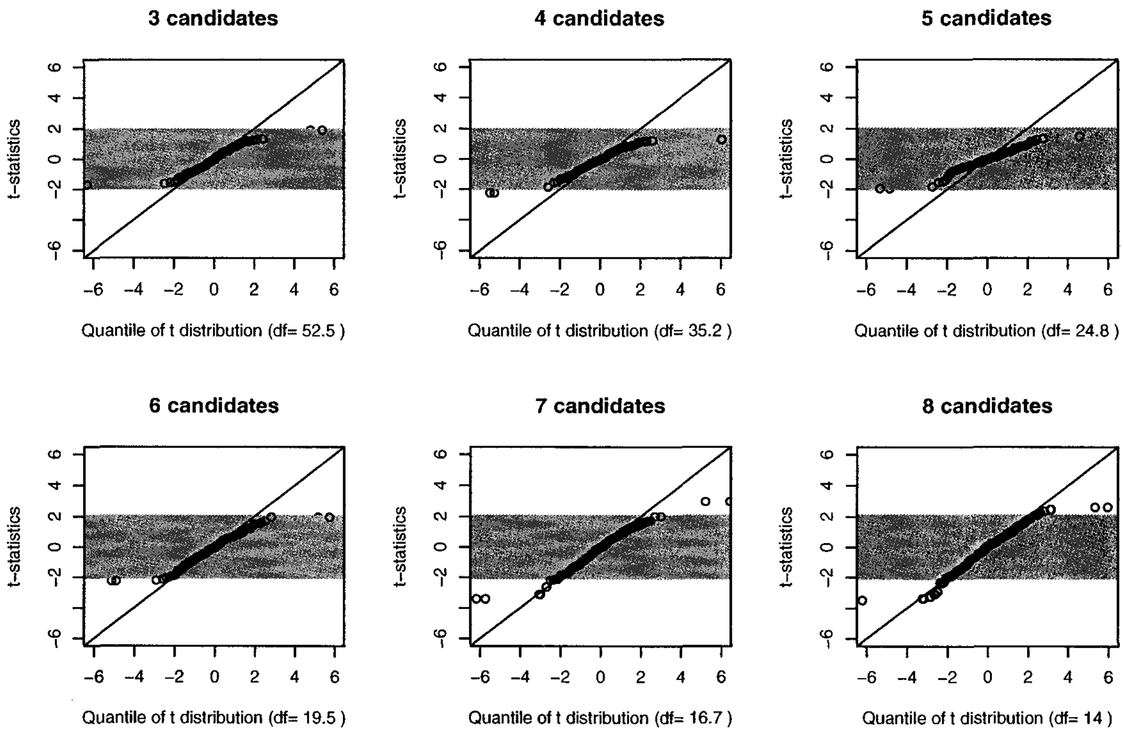


Figure 4.5: Distribution of T-statistics from Covariate Means Tests. The t-statistics are calculated using every systematic treatment assignment combination for 34 district level covariates. The statistics are plotted against the quantiles of a t-distribution with a corresponding degrees of freedom. The solid lines represent 45% degree line, and the gray bar indicates the area where p-values are less than 0.95.

ble treatment assignments, this yields $1,122 = 34 \times \sum j$ means tests. Figure 4.5 compares these t-statistics with the quantiles of the t-distribution with appropriate degrees of freedom. Under simple random assignment, we expect the distribution of the test statistics to approximate a t-distribution. The statistics are overwhelmingly bounded by the shaded 95% intervals, indicating that there is relative balance across all observed treatment assignments. In fact, in the case of small number of candidates, the covariate balance appears to be even better than under simple random assignment. As the number of candidates increases, balance decreases, as indicated by the few outliers for the 8 candidate panel, although the distribution still generally follows a t-distribution. All together, the analysis of this section shows no evidence of incomplete randomization.

4.9 Appendix: Conditional Effects

	General Election				Primary Election			
	Incumbent Race		Open Race		Incumbent Race		Open Race	
Party	ATE	SE	ATE	SE	ATE	SE	ATE	SE
Democratic	0.27	0.59	-0.30	0.73	1.79	0.32	1.92	0.22
Republican	-0.66	0.64	0.91	0.90	3.78	0.65	1.74	0.27
American Independent	0.15	0.02	0.18	0.02	0.26	0.05	3.81	0.18
Green	0.83	0.28	0.24	0.20	3.34	3.96	3.10	0.38
Libertarian	0.24	0.02	0.21	0.03	0.43	0.08	10.01	1.05
Natural Law	0.18	0.03	0.51	0.15	0.41	0.06	0.38	0.20
Peace and Freedom	0.23	0.04	0.39	0.04	4.57	0.86	6.86	0.43
Reform	0.31	0.10	0.15	0.03	3.73	0.69	4.42	1.37
Nonpartisan	1.59	0.39	2.67	0.47	3.08	0.54	3.89	0.65

Table 4.11: Absolute Ballot Order Effects Conditional Whether Incumbents are Running.

	General Election				Primary Election			
	On-Year		Off-Year		On-Year		Off-Year	
Party	ATE	SE	ATE	SE	ATE	SE	ATE	SE
Democratic	1.94	0.90	-0.84	1.23	0.88	0.29	1.55	0.30
Republican	-0.08	1.23	-1.16	1.40			2.81	0.51
American Independent	0.18	0.06	0.12	0.03			0.15	0.05
Green	1.50	1.04	0.33	0.05			4.58	2.07
Libertarian	0.20	0.05	0.13	0.02			0.49	0.10
Natural Law	0.02	0.02	0.10	0.03			0.19	0.03
Peace and Freedom	0.51	0.07	0.25	0.13	6.06	1.05	1.99	0.55
Reform	0.26	0.07	0.01	0.06			5.83	1.14
Nonpartisan								

Table 4.12: Absolute Ballot Order Effects for On or Off-Year Elections for Senate Elections.

General Elections

No. Cand.	Democrat		Republican		Am. Indep.		Green		Libertarian		Natl Law		Peace & Frdm		Reform		Nonpartisan	
	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE
2																	8.5	1.1
5	1.2	1.0	1.0	1.4	17.1	1.4	6.7	6.3	11.8	1.4	5.8	12.4	17.7	1.6			10.3	5.4
6	0.1	1.6	−1.8	2.6	18.1	2.2	10.3	4.7	19.3	1.8	20.8	6.3	14.0	6.4			2.8	6.5
7	−1.4	2.2	−0.4	3.2	26.3	3.5	31.3	10.7	13.4	2.0	32.1	3.1	41.6	5.7	20.9	2.3	12.7	13.3
8	3.2	6.3	−13.2	7.8	45.0	15.5	32.8	32.8	31.6	11.9	25.4	9.1	76.7	19.7	9.2	8.0		

Primary Elections

No. Cand.	Democrat		Republican		Am. Indep.		Green		Libertarian		Natl Law		Peace & Frdm		Reform		Nonpartisan	
	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE	ATE	SE
2	11.8	1.0	7.4	0.7	29.5	1.2	18.4	1.9					13.9	1.0				
3	14.1	3.2	8.7	1.1														
4	12.3	1.7	14.9	1.6													18.9	3.3
5	29.1	2.4	20.8	3.5													19.9	2.5
7	−2.4	4.5	34.3	6.4	13.8	10.3			10.2	5.3	5.2	5.4	10.1	14.9	33.5	6.2		
8	5.4	4.1	10.6	6.3	25.1	4.5			15.0	6.7	38.4	6.3	14.0	3.0	48.5	9.5		
9	35.6	4.5	3.0	4.6	13.1	7.2			−2.1	8.3	44.9	27.3	10.9	9.7				
10	36.7	7.8	11.2	2.6	28.7	8.0			25.1	7.1			19.8	2.9				
11	63.7	5.7																
13	60.3	5.5	52.3	5.9	36.0	11.0	−24.0	12.7	18.4	8.2	26.6	6.2	15.5	5.1	43.0	15.1		
15	11.0	12.5	68.4	9.5	42.1	12.3	1.4	10.7	44.0	12.7	75.5	13.8			24.5	4.9		
17	73.2	19.0	65.1	11.3	28.0	14.7	−15.9	14.6	31.6	14.7	53.5	19.1	19.1	3.4				
23	98.1	44.4	22.5	10.1	14.2	25.4	12.9	3.7	166.2	19.1	75.8	19.3			63.2	9.6		

Table 4.13: Relative Ballot Order Effects Conditional on the Number of Candidates.

Chapter 5

Assessing Effects of Racial Perceptions on Political Knowledge: A Bayesian Approach to Instrumental Variables

5.1 Introduction

Causal inference in observational studies is notoriously difficult without physical randomization of treatment. Yet even when the treatment of interest cannot be randomized, randomized instruments may enable researchers to draw causal inferences of theoretical interest (e.g., Howell and Peterson, 2002; Gerber and Green, 2000). As a result, they bear much promise in political science, where randomization of the treatment itself is often infeasible or unethical. This paper illustrates a Bayesian framework of analyzing, making explicit, and relaxing assumptions that are implicit in virtually all instrumental variables ("IV") analyses. Imbens and Rubin (1997a) and Imbens and Rubin (1997b) developed this framework, which improves IV estimation by permitting researchers to test the sensitivity of inferences to crucial exclusion restrictions commonly imposed in studies (see also, Hirano et al., 2000; Frangakis and Rubin, 2002; Frangakis, Rubin and Zhou, 2002). Angrist, Imbens and Rubin (1996) showed that such IV analyses may be interpreted as causal effects under a potential outcomes framework of causal inference, commonly termed as the Rubin Causal Model (Rubin, 1974; Holland, 1986).

We illustrate the usefulness of this framework with data from a telephone political knowledge survey, to assess how racial perceptions affect answers to the survey. Davis and Silver (2003) ("DS") first analyzes this dataset examining primarily the causal effect of racial perceptions on answers to factual political knowledge questions. Based on a theory of stereotype threats, DS proposes two hypotheses. First, black respondents should perform worse on survey responses when *perceiving* a white interviewer: "the 'threat' [of negative intellectual capacity] is likely to be induced by the *perception* that the interviewers are white, not directly by whether the interviewers are *actually* white" (Davis and Silver, 2003, p. 39). Second, white respondents should not be affected by the race, or perception of race, of interviewers. Accordingly, DS finds that the perception of race, but not race itself, negatively affects answers by black respondents, and that race does not affect white respondents. These propositions are controversial. Few studies have investigated the differences between race and racial perception as causal factors. If indeed only racial perception matters, irrespective of other traits associated with race such as language and communication, the finding may substantially help in isolating the complex dynamics of race, which many have eschewed as being difficult to study from a causal perspective (Holland, 2003). Even more controversial, and at odds with a burgeoning social psychology literature on the effects of race on majority groups (Danso and Esses, 2001; Aronson et al., 1999), is the finding that the race and racial perception of interviewers does not affect white respondents.

Our analysis shows that the DS findings are likely to be artifacts of selection bias. Employing the randomized instrument of the race of the interviewer to overcome this bias, we find that both white and black respondents are at times affected by race of the interviewer and that the effect is unlikely to be solely due to racial perceptions. Even respondents who incorrectly assessed the race of their interviewer are affected by the mere assignment, not perception, of race.

This paper proceeds as follows. We first describe the political knowledge data, and problems inherent in assessing causal effects of a possibly endogenous variable, such as racial perception. Second, we describe the framework of causal inference that capitalizes on the random assignment of an instrument (here, interviewer race) to draw causal inferences about the treatment of interest (here, respondent racial perception). Third, we outline the Bayesian pattern-mixture model used to implement the framework for the political knowledge data. We then discuss results and conclude with implications.

5.2 The Political Knowledge Data

We observe $i = 1,\ldots,N$ telephone survey respondents to the Michigan State of the State survey conducted in 2001. Each respondent was randomly assigned the race of a phone interviewer, denoted by $Z_i = 1$ if the interviewer identified herself as black, and 0 if white, which we will refer to as *assignment*. The treatment of interest is respondent i's perception of the race of the interviewer, denoted as $T_i = 1$ if the respondent perceived the interviewer to be black, and 0 if white. We also observe three pre-treatment covariates of gender, equal to 1 if the respondent is male and 0 if female, college degree, equal to 1 if the respondent graduated college and 0 otherwise, and college attendance, equal to 1 if the respondent attended some college but did not receive a degree and 0 otherwise. Let X_i denote the row vector of these pre-treatment covariates for each respondent i. The outcomes of interests are answers to seven political knowledge questions denoted by Y_{ij} for questions $j = 1, \ldots, 7$, which equals 1 if respondent i answered the jth question correctly, and 0 otherwise. These questions are presented in Table 5.1 and summary statistics for the variables of interest are presented in Table 5.2.

While Table 5.2 suggests that there is relative balance for observed covariates along the perception of race, any inference using a non-randomized regressor such as racial perception relies on the assumption that conditional on observed covariates, the assignment of racial perceptions is *random*. In the econometric literature, this is termed as exogeneity conditional on covariates or selection on observables (Heckman and Robb, 1985), and in the statistical literature it is known as ignorability of treatment (Rosenbaum and Rubin, 1983a). This may be violated for a variety of reasons, such as the nature of coding racial perception. Interviewers asked respondents at the end of the interview: "what do you think is my racial background?" Accurate responses are likely to be confounded with a host of other variables, such as social politeness, embarassment, socioeconomic background

Political Knowledge Question	Answer	Proportion Correct White	Black
Minimum voting age	18 years	0.85	0.87
Presidential term limits	2 terms	0.90	0.85
Majority party in state legislature	Republican	0.62	0.50
Vote required to override presidential veto	$\frac{2}{3}$ vote	0.42	0.21
Number of US Supreme Court Justices	9 Justices	0.32	0.22
Term length of US Senator	6 years	0.34	0.19
Office of William Rehnquist	Chief Justice	0.29	0.13

Table 5.1: Political knowledge question asked to 212 black respondents and 221 white respondents, and proportion answered correctly

(segregation), social awareness, and how seriously respondents take the survey. Indeed, more than three times as many white respondents answered "don't know" to the question, indicating that actual responses as to race are not likely to be conditionally random.

Fortunately, in the political knowledge survey we also observe the instrument of race of the interviewer which ensures balance of observed and *unobserved* covariates. This permits us to explicitly test the inferences about racial perceptions, in way that is robust to these problems of endogenous perceptions.

5.3 Framework of Causal Inference

To leverage randomized assignment, we use a framework of modeling compliance (Imbens and Rubin, 1997a). Compliance behavior here may be interpreted as respondent perception of race in response to assignment of interviewer race. Let $D_i(z)$ represent an indicator of perception of race as a function of assignment. $D_i(0)$ equals 1 if i would perceive the interviewer as black if i's interviewer was in fact white and, 0 if i would perceive the

		Interviewer race			Interviewer perceived		
		white	black		white	black	
	Mean	($Z_i = 0$)	($Z_i = 1$)	t-stat	($D_i = 0$)	($D_i = 1$)	t-stat
Black respondents (n=212)							
Interviewer black (Z_i)	0.68	0.00	1.00		0.37	0.85	-7.63
Perceived black (D_i)	0.64	0.29	0.81	-7.90	0.00	1.00	
Questions correct ($\sum_j^7(\frac{Y_{ij}}{7})$)	2.97	2.84	3.03	-0.87	2.75	3.10	-1.68
Male (X_{i1})	0.33	0.28	0.35	-1.00	0.33	0.32	0.08
College degree (X_{i2})	0.20	0.15	0.23	-1.47	0.16	0.23	-1.26
Some college (X_{i2})	0.37	0.41	0.35	0.80	0.43	0.34	1.37
White respondents (n=221)							
Interviewer black (Z_i)	0.56	0.00	1.00		0.23	0.79	-9.62
Perceived black (D_i)	0.59	0.29	0.83	-9.46	0.00	1.00	
Questions correct ($\sum_j^7(\frac{Y_{ij}}{7})$)	3.74	3.68	3.79	-0.49	3.81	3.69	0.51
Male (X_{i1})	0.43	0.41	0.44	-0.34	0.48	0.39	1.30
College degree (X_{i2})	0.37	0.37	0.36	0.13	0.36	0.37	-0.28
Some college (X_{i2})	0.28	0.30	0.26	0.67	0.27	0.28	-0.26

Table 5.2: Summary statistics of political knowledge survey

interviewer as white. $D_i(1)$ equals 1 if i would perceive the interviewer as black if i's interviewer was in fact black and, 0 if i would perceive the interviewer as white. Four types of compliance behavior are then defined by C_i:

$$C_i = \begin{cases} c & \text{if } D_i(z) = z \quad \text{(Compliers)} \\ n & \text{if } D_i(z) = 0 \quad \text{(Never-takers)} \\ a & \text{if } D_i(z) = 1 \quad \text{(Always-takers)} \\ d & \text{it } D_i(z) = 1 - z \quad \text{(Defiers)} \end{cases}$$

for $z = 0, 1$. Compliance behavior, i.e., the type of respondent, is not directly observable, but causal effects are only properly defined within each type (Frangakis and Rubin, 2002). This is because an "as-treated" analysis fails to account for the fact that compliers, the subpopulation of respondents who change their racial perception due to assignment of interviewer, may be substantially different from noncompliers (i.e., always-takers, never-takers, and defiers) in observed and unobserved factors. The instrument of interviewer race circumvents this selection bias between respondents, since perception for this subgroup differs only as a result of assignment, which is random.

We can now also define the potential outcomes $Y_i(z, D_i(z))$ under each assignment, suppressing j for notational simplicity. We treat the counterfactual responses as missing data. The overall intention-to-treat effect (ITT) is then defined as a weighted average of ITT effects for each subtype:

$$\begin{aligned} ITT &= \sum_{t \in \{c,n,a,d\}} ITT_t * N_t/N \\ ITT_t &= \sum_{i \in \{i:C_i=t\}} \left[Y_i(1, D_i(1)) - Y_i(0, D_i(0))\right]/N_t \end{aligned}$$

where $t \in \{c, n, a, d\}$, N_t denotes the number of respondents in each subtype, and N denotes the total number of respondents.

The identification assumptions usually imposed in IV studies to estimate ITT_c are as follows:

ASSUMPTION 3 (NO INTERFERENCE AMONG UNITS) *$Y_i(z, D_i(z)) \perp\!\!\!\perp Z_{i'}$ for all $z = 0, 1$ where $i' \neq i$.*

where $\perp\!\!\!\perp$ denotes independence. This assumption implies that the potential responses to survey questions by one respondent are not affected by the assignment of interviewer race of another respondent. Where the assignment of one might affect outcomes of other units, this assumption is violated. In political science, for example, the assignment of municipality services as an instrument for voter information may also affect voters outside of the assigned municipality, if voters move to different municipalities as a result of assignment(Lassen, 2003).

ASSUMPTION 4 (RANDOM ASSIGNMENT) $Y_i(z, D_i(z)), D_i(z) \perp\!\!\!\perp Z_i$ *for all i and* $z = 0, 1$.

Random assignment of interviewer race is arguably met by virtue of the standard random digit dialing of the survey: "[a]ssignment [of respondents] was not based on either the characteristics of the interviewer or the characteristics of the respondents" (Davis and Silver, 2003, p. 37). Accordingly, Table 5.2 provides no evidence of incomplete randomization within substrata of respondent races.

Nonetheless, it remains curious that the proportion of self-identified black interviewers differs appreciably across black and white respondents. While roughly 67% of interviewers of black respondents identified themselves as black, only 56% of interviewers of white respondents identified themselves as black (t-stat=2.5). This may be due to (a) clustering of 73 interviewers, which would bias our standard errors downwards, or (b) undocumented effects arising from the regional stratification in the survey designed to ensure roughly equal number of black and white respondents.

As a result, we relax the randomization assumption to ignorability of assignment conditional on covariates by stratifying exactly on respondent race, as in the DS specification with endogenous racial perception. To the degree that there are further unspecified errors in randomization of interviewer race with respect to unobserved covariates, causal inferences may be substantially more difficult.

ASSUMPTION 5 (MONOTONICITY) $D_i(1) \geq D_i(0)$ *for all i.*

Monotonicity rules out respondents who would perceive black interviewers to be white, and white interviewers to be black. While plausible here, where there is no evidence of hostile respondent behavior which might imply the existence of defiers, it may not be in other political science settings. Take the case of randomized direct mailing of campaign advertisement as instruments for voter information (Gerber and Green, 2000). Defiers might be less inclined to inform themselves about the campaign as a result of reading the direct mailing, leading classic IV estimates to be biased.

ASSUMPTION 6 (EXCLUSION RESTRICTIONS ON SUBTYPES)

(A) Never-takers: $P(Y_i(1,D_i(1)) = 1|X_i, C_i = n) = P(Y_i(0,D_i(0)) = 1|X_i, C_i = n)$

(B) Always-Takers: $P(Y_i(1,D_i(1)) = 1|X_i, C_i = a) = P(Y_i(0,D_i(0)) = 1|X_i, C_i = a)$

Exclusion restrictions (A) and (B) rule out any effect of race on those respondents for whom the race of the interviewer has no effect on racial perception. This formalizes the DS hypothesis that race should affect responses only via the perception of race. More generally, these restrictions in conjunction with monotonicity hold that the instrument affects outcomes only in the subpopulation of compliers who are treated as a result of assignment.

The virtue of a Bayesian framework lies in the fact that it permits (a) relaxing monotonicity to allow for defiers and (b) relaxing the exclusion restrictions to test whether race affects respondents whose racial perceptions are not altered by assignment. Why might ITT effects occur in these subpopulations who are unaffected by interviewer race in their racial perceptions? Theories emphasizing linguistic differences in interviewers might predict large ITT effects for respondents who do not perceive racial cues, while theories emphasizing social desirability with particular social groups would predict no ITT effects on these subpopulations. In addition, the assignment of a black interviewer may induce informal changes in the administration of the survey script because of *interviewer* perceptions irrespective of respondent perceptions. Whether race itself or race mediated by racial perceptions affects responses is a question of potentially wide interest, which may be explicitly tested in this framework.

Of course, ITT effects for non-compliers might alternatively indicate that the method of assessing racial perception is simply unreliable, due to self-censoring when respondents do in fact *perceive* racial cues. ITT effects for non-compliers thereby indicate that either the theory or the measurement is violated in some fashion. Both should be of interest to researchers in this field.

5.4 A Mixture Model

We follow Hirano et al. (2000) and model the outcomes as a mixture of three types. We let θ denote the vector containing all parameters. Each type of respondent has an outcome distribution that is modeled as a logistic regression:

$$P(Y_i|C_i = t, Z_i = z, X_i, \theta) = \pi_{tzi} = \frac{\exp(\alpha_{tz} + X_i\beta)}{1 + \exp(\alpha_{tz} + X_i\beta)}$$

where α_{tz} represents the intercept for each type $t \in \{c, n, a\}$ and assignment $z = 0, 1$, and β represents the vector of three slope parameters for pre-treatment covariates. We assume equal slopes across subtypes and z, which may be relaxed. The distribution of types is modeled as a multinomial logit:

$$P(C_i = t|X_i, \theta) = \Psi_{ti} = \frac{\exp(X_i \psi_t)}{\sum_{k \in \{c,n,a\}} \exp(X_i \psi_k)}$$

where for identification purposes we restrict, as usual, each element of ψ_n to 0, and ψ_c and ψ_a are vectors of four parameters corresponding to covariates and an intercept. As a result, there are 17 parameters to be estimated in $\theta = \{\alpha_{c1}, \alpha_{c0}, \alpha_{a1}, \alpha_{a0}, \alpha_{n1}, \alpha_{n0}, \beta, \psi_a, \psi_c\}$ under no exclusion restrictions. Exclusion restriction (A) implies that $\alpha_{n1} = \alpha_{n0}$, and (B) implies that $\alpha_{a1} = \alpha_{a0}$. We can now write the complete data likelihood as:

$$\begin{aligned} \mathcal{L}(\theta|Z, D, Y, X, C) &= \prod_{i \in \{i:C_i = c\}} \Psi_{ci}(\pi_{czi})^{Y_i}(1-\pi_{czi})^{1-Y_i} \times \prod_{i \in \{i:C_i = n\}} \Psi_{ni}(\pi_{nzi})^{Y_i}(1-\pi_{nzi})^{1-Y_i} \\ &\quad \times \prod_{i \in \{i:C_i = a\}} \Psi_{ai}(\pi_{azi})^{Y_i}(1-\pi_{azi})^{1-Y_i} \end{aligned}$$

We use the following conjugate prior, which adds fractional observations to the data to ensure proper form of the posterior (Clogg et al., 1991):

$$p(\theta) \propto \prod_{i=1}^{N} \prod_{t \in \{c,n,a\}} \prod_{z=0,1} \prod_{y=0,1} \left\{ \Psi_{ti}(\pi_{tzi})^{y}(1-\pi_{tzi})^{1-y} \right\}^{2.5/N}$$

Similar to Jeffrey's prior in a binomial model, this prior shrinks the likelihood towards an equal number of subtypes.

5.5 Results

We were able to substantially replicate the main findings with the original data (using the endogenous perception of race). While the original DS analysis uses OLS to model the total number of correct answers, for this analysis we use a more appropriate logit model

	White		Black	
	ITT Effect	S.D.	ITT Effect	S.D.
Voting age	0.13	0.03	0.01	0.03
Pres. term	−0.01	0.02	0.03	0.03
Maj. party	−0.02	0.03	−0.07	0.04
Veto override	0.08	0.03	0.02	0.03
No. Justices	−0.09	0.03	−0.03	0.03
Senate term	0.06	0.03	0.03	0.03
Rehnquist	−0.00	0.03	0.07	0.03

Table 5.3: Intention to treat results from 14 logistic models, with flat prior and 10,000 draws of the Metropolis sampler, using the multivariate normal as the jumping distribution, scaled by the asymptotic variance-covariance matrix with maximum likelihood estimates as starting values.

for binary responses to each of the questions. We first present results from an ITT analysis on the effect of interviewer race on the survey answers. We then present IV results of the effects of racial perception, testing sensitivity to the exclusion restrictions.

5.5.1 ITT Analysis

Table 5.3 presents posterior distributions of the the ITT effects estimated with a logistic model, with covariates and interviewer race as explanatory variables, for each of the seven political knowledge questions for black and white respondents.

The results are surprising in light of the DS theory. We detect substantial ITT effects in the direction anticipated by DS for only one of the seven questions for black respondents. In fact, interviewer race appears to have greater effects on white respondents than on black respondents. In three of the seven questions, the posterior probability of a positive effect is greater than 97%, while the effect is negative and significant for one model (veto override).

A black interviewer increases the probability of a white respondent answering the Senate term question correctly by roughly 13%. While none of these results are suggested by the DS theory, which posits that white respondents should remain unaffected by interviewer race, they are consistent with other studies finding that white respondents fare better with black experimenters (e.g., Danso and Esses, 2001).

5.5.2 IV Analysis

To turn now explicitly to the quantity of interest professed by DS, we examine whether the ITT effects are induced by racial perceptions only. Separate mixture models were estimated for each question for black and white respondents. We considered four combinations of exclusion restrictions: (a) restrictions on always and never-takers, (b) always-takers only, (c) never-takers only, and (d) none.

We estimate the parameters of the mixture model by MCMC with a Gibbs sampler iterating between (a) drawing the latent subtypes and (b) drawing parameters conditional on the subtypes. Step (a) is simplified by the monotonicity assumption, which implies that $P(C_i = a|Z_i = 0, D_i = 1) = 1$ and $P(C_i = n|Z_i = 1, D_i = 0) = 1$. Hence, we draw from the mixture of compliers and always takers when $Z_i = 1, D_i = 1$ and mixture of compliers and never takers when $Z_i = 0, D_i = 0$. For efficiency, draws in step (b) were undertaken in sequential Metropolis steps for first the multinomial parameters $\{\psi_a, \psi_c\}$, second the subtype intercepts $\{\alpha_{c1}, \alpha_{c0}, \alpha_{a1}, \alpha_{a0}, \alpha_{n1}, \alpha_{n0}\}$, and finally the covariate slopes $\{\beta\}$, using a t-distribution with five degrees of freedom scaled by tuning constants as the jumping distribution. Tuning constants were chosen to minimize autocorrelation of the posterior draws. Four overdispersed chains of 14,500 draws each were run for each model, discarding the first 2,000 burn-in draws, saving every fifth draw, and combining the four chains for a posterior sample of 10,000 draws. Convergence was monitored by examining autocorrelation of all parameters and $\hat{R}$ statistics for overdispersed chains (Gelman and Rubin, 1992). The sampling algorithm was implemented in C++, using the Scythe matrix library (Martin, Quinn and Pemstein, 2002), and convergence diagnostics were conducted in R.

Table 5.4 presents results of quantities of interest from the posterior distribution of models imposing exclusion restrictions for always and never-takers and no exclusion restrictions. For space limitations and since results are materially the same, we do not report models relaxing exclusion restrictions (A) and (B) separately. Consider first the effects estimated on white respondents, for whom DS posits race or racial perception should not affect responses. Imposing both exclusion restrictions, we estimate the perception of a black interviewer by a white respondent to cause a roughly 25% increase in the probability of answering the voting age question correctly. Yet, relaxing the exclusion restrictions demonstrates that never-takers (respondents that perceive the interviewer to be white irrespective of assignment) exhibit substantial ITT_n effects as well. In other words, race appears to affect responses not only through racial perceptions.

The top panel of Figure 5.1 plots simulation scatterplots of the joint posterior distribution of ITT_n and ITT_c from that question. Most of the posterior mass for both never-takers and compliers is concentrated above the origin, and the effect appears to be larger for never-takers, albeit with higher variance. Due to the relatively small sample size we unfortunately cannot estimate the parameters with a high degree of precision, particularly given the small proportions of noncompliers. Nevertheless, this demonstrates the flexibility of this framework: imposing exclusion restrictions that are inherent in most IV studies can lead us to draw overconfident inferences on compliers. Examining the results for the question on the number of Justices, the negative ITT effect appears to be driven by compliers. White respondents who correctly perceive a black interviewer are roughly 30% less likely to answer the question correctly.

Consider now the effects estimated on black respondents. The only substantial complier effect imposing all exclusion restrictions is in the opposite direction anticipated by DS for the majority party question. Even worse, after relaxing the exclusion restrictions, the only substantive effect of black interviewers in the posited direction is not on compliers but on always- and never-takers for the voting age question. The bottom panel of Figure 5.1 shows that while the ITT_c effect for that question is centered around the origin, over 90% of the posterior mass of ITT_n is positive.

5.6 Conclusion

We have illustrated an application of an exciting new area of applied research, spawned by reinterpretations of traditional IV estimation strategies in a potential outcomes framework. Our analysis shows that researchers can leverage randomized instruments to test hypotheses with greater scientific credibility. Our findings are largely inconsistent with

DS, whose findings are likely to be an artifact of endogenous racial perceptions. Instead, consistent with (Aronson et al., 1999) and (Danso and Esses, 2001), we show that white and black respondents alike appear to be affected by race at times, and that these effects do not appear to be mediated by measured racial perceptions alone.

	ITTc	S.D.	$P(C_i = c)$	ITTa	S.D.	$P(C_i = a)$	ITTn	S.D.	$P(C_i = n)$	ITT	S.D.
White respondents (exclusion restrictions on always and never takers)											
Voting age	0.25	0.08	0.51	0	0	0.30	0	0	0.18	0.13	0.04
Pres. Term	-0.01	0.06	0.51	0	0	0.31	0	0	0.18	-0.00	0.03
Maj. party	0.00	0.12	0.50	0	0	0.31	0	0	0.19	0.00	0.06
Veto override	0.16	0.10	0.51	0	0	0.31	0	0	0.18	0.08	0.05
No. Justices	-0.23	0.10	0.50	0	0	0.31	0	0	0.18	-0.11	0.05
Senate term	0.10	0.11	0.51	0	0	0.31	0	0	0.18	0.05	0.05
Rehnquist	-0.02	0.10	0.51	0	0	0.31	0	0	0.18	-0.01	0.05
White respondents - no exclusion restrictions											
	ITTc	S.D.	$P(C_i = c)$	ITTa	S.D.	$P(C_i = a)$	ITTn	S.D.	$P(C_i = n)$	ITT	S.D.
Voting age	0.17	0.12	0.52	-0.01	0.10	0.30	0.23	0.28	0.18	0.13	0.04
Pres. Term	-0.01	0.07	0.53	-0.06	0.10	0.30	0.11	0.17	0.17	-0.00	0.03
Maj. party	-0.02	0.20	0.51	-0.17	0.21	0.30	0.23	0.31	0.18	-0.02	0.05
Veto override	0.10	0.17	0.51	0.06	0.23	0.31	0.08	0.22	0.18	0.08	0.04
No. Justices	-0.29	0.14	0.52	0.17	0.17	0.30	0.08	0.24	0.18	-0.09	0.04
Senate term	0.09	0.18	0.51	0.02	0.25	0.31	0.05	0.24	0.18	0.06	0.04
Rehnquist	-0.09	0.14	0.53	0.24	0.19	0.30	-0.16	0.22	0.18	-0.00	0.04
Black respondents (exclusion restrictions on always and never takers)											
	ITTc	S.D.	$P(C_i = c)$	ITTa	S.D.	$P(C_i = a)$	ITTn	S.D.	$P(C_i = n)$	ITT	S.D.
Voting age	0.03	0.08	0.48	0	0	0.31	0	0	0.21	0.02	0.04
Pres. Term	0.01	0.08	0.50	0	0	0.30	0	0	0.20	0.01	0.04
Maj. party	-0.19	0.14	0.47	0	0	0.32	0	0	0.21	-0.09	0.06
Veto override	0.06	0.09	0.48	0	0	0.31	0	0	0.21	0.03	0.04
No. Justices	-0.03	0.09	0.50	0	0	0.30	0	0	0.20	-0.02	0.04
Senate term	0.08	0.09	0.48	0	0	0.31	0	0	0.21	0.04	0.04
Rehnquist	0.06	0.08	0.49	0	0	0.31	0	0	0.20	0.03	0.04
Black respondents - no exclusion restrictions											
	ITTc	S.D.	$P(C_i = c)$	ITTa	S.D.	$P(C_i = a)$	ITTn	S.D.	$P(C_i = n)$	ITT	S.D.
Voting age	-0.02	0.09	0.50	-0.05	0.12	0.30	0.27	0.19	0.20	0.03	0.04
Pres. Term	-0.05	0.09	0.50	0.13	0.15	0.30	0.13	0.18	0.20	0.04	0.04
Maj. party	-0.24	0.21	0.48	0.02	0.20	0.31	0.14	0.26	0.21	-0.08	0.05
Veto override	0.11	0.11	0.49	-0.07	0.12	0.30	-0.13	0.13	0.20	0.00	0.04
No. Justices	0.02	0.11	0.49	-0.10	0.13	0.30	-0.11	0.16	0.20	-0.04	0.05
Senate term	0.14	0.10	0.49	-0.10	0.12	0.31	-0.13	0.15	0.20	0.01	0.04
Rehnquist	0.03	0.08	0.50	0.10	0.11	0.30	-0.02	0.11	0.20	0.04	0.04

Table 5.4: Summary statistics of posterior distributions of models with exclusion restrictions for always and never-takers and no exclusion restrictions

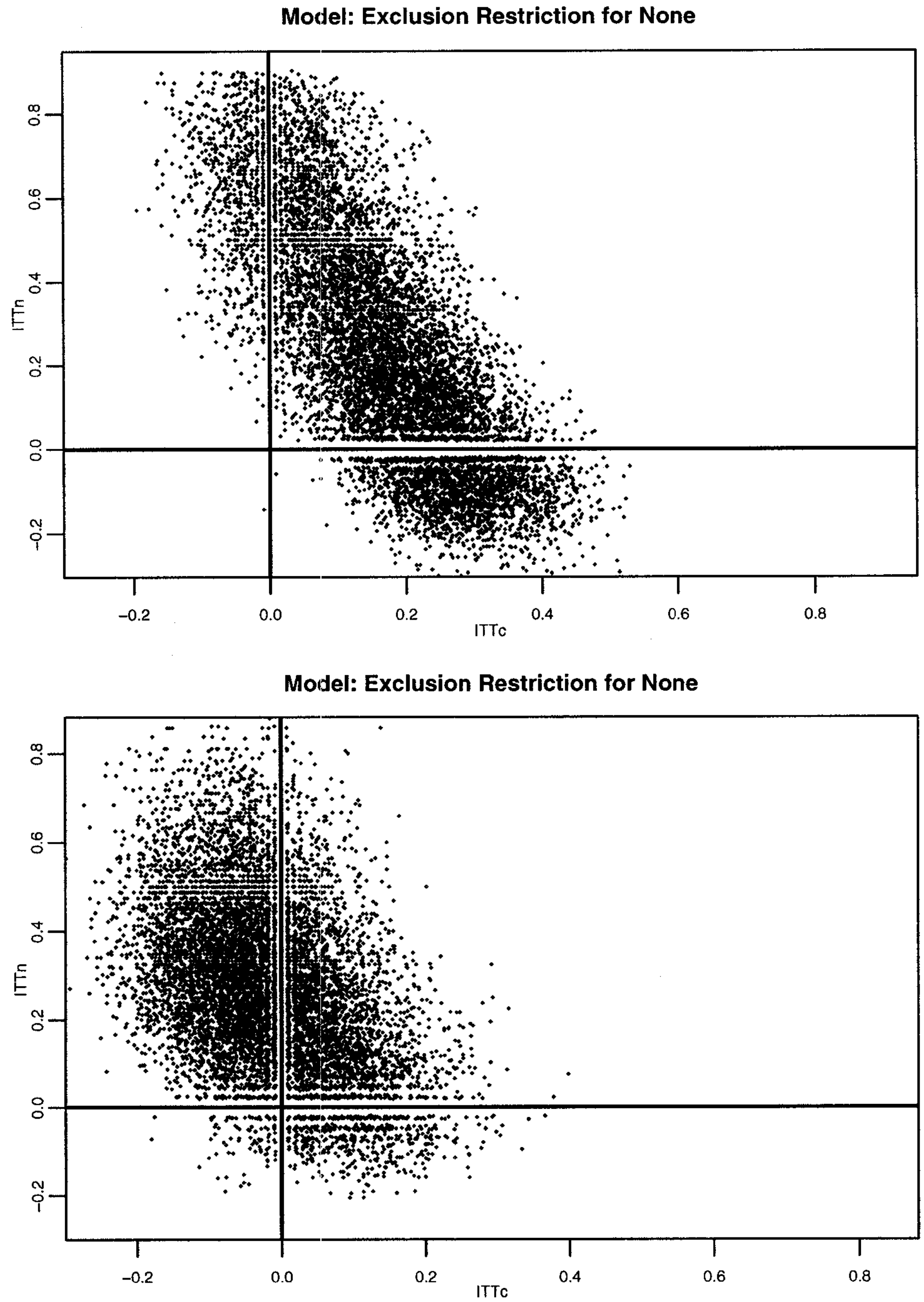


Figure 5.1: Simulation scatterplot of the joint posterior distribution of (a) ITT_n and ITT_c in the model of white respondents to the minimum voting age question with no exclusion restriction for never-takers (top panel) and (b) ITT_a and ITT_c in the model for black respondents to the minimum voting age question with no exclusion restrictions (bottom panel)

Bibliography

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80:1284–1286.

Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables (with discussion)." *Journal of the American Statistical Association* 91:444–455.

Ansolabehere, Stephen, Shigeo Hirano, Jim Snyder and Michiko Ueda. 2003. "Voting Cues and the Incumbency Advantage: Non-Partisan and Partisan Elections to the Minnesota State Legislature, 1950-1988." *Technical Report.*

Aronson, Joshua, Michael J. Lustina, Catherine Good, Kelli Keough, Claude M. Steele and Joseph Brown. 1999. "When White Men Can't Do Math: Necessary and Sufficient Factors in Stereotype Threat." *Journal of Experimental Social Psychology* 35:29–46.

Bagley, C.R. 1966. "Does Candidates' Position on the Ballot Paper Influence Voters' Choice? A Study of the 1959 and 1964 British General Elections." *Parliamentary Affairs* 19:162–174.

Bain, Henry M. and Donald S. Hecock. 1957. *Ballot Position and Voter's Choice.* Detroit: Wayne State University.

Brady, Henry E., Michael Herron, Walter R. Mebane, Jasjeet S. Sekhon, Kenneth Shotts and Jonathan Wand. 2001. "Law and Data: The Butterfly Ballot Episode." *PS: Political Science and Politics* 34:59–69.

Caltech and MIT Voting Technology Project. 2001. "Voting: What is, What Could be." *Technical Report.*

Clogg, Clifford C., Donald B. Rubin, Nathaniel Schenker, Bradley Schultz and Lynn Weidman. 1991. "Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression." *Journal of the American Statistical Association* 86:68–78.

Cochran, William G. 1977. *Sampling Techniques.* 3rd ed. New York: John Wiley & Sons.

Cox, David. R. 1958. *Planning of Experiments.* New York: John Wiley & Sons.

Cox, Gary. 1990. "Centripetal and Centrifugal Incentives in Electoral Systems." *American Journal of Political Science* 34:903–35.

Danso, Henry A. and Victoria M. Esses. 2001. "Black Experimenters and the Intellectual Test Performance of White Participants: The Tables Are Turned." *Journal of Experimental Social Psychology* 37:158–165.

Darcy, R. 1986. "Position Effects with Party Column Ballots." *Western Political Quarterly* 39:648–662.

Darcy, R. and Ian McAllister. 1990. "Ballot Position Effects." *Electoral Studies* 9:5–17.

Davis, Darren W. and Brian D. Silver. 2003. "Stereotype Threat and Race of Interviewer Effects in a Survey on Political Knowledge." *American Journal of Political Science* 47:33–45.

Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94:1053–62.

Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score Matching Methods for Non-Experimental Causal Studies." *Columbia University, Department of Economics Discussion Paper* 0102-14.

Epstein, Richard A. 1987. "Causation in Context: An Afterword." *Chicago-Kent Law Review* 63:653–80.

Fiona, McGillivray. 2002. "Government Hand-Outs, Political Institutions, and Stock Price Dispersion." Manuscript, available at `http://www.nyu.edu/gsas/dept/politics/seminars/fiona.pdf`.

Fisher, Ronald A. 1935. *The Design of Experiments.* London: Oliver and Boyd.

Forsythe, R., R. B. Myerson, T. A. Rietz and R. J. Weber. 1993. "An experiment on coordination in multi-candidate elections - the importancel of polls and election histories." *Social Choice and Welfare* 10:223–247.

Frangakis, Constantine E. and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58:21–29.

Frangakis, Constantine E., Donald B. Rubin and Xiao-Hua Zhou. 2002. "Clustered Encouragement Designs with Individual Noncompliance: Bayesian Inference with Randomization, and Application to Advance Directive Forms (with discussion)." *Biostatistics* 3:147–164.

Gelman, Andrew and Donald B. Rubin. 1992. "Inference from Iterative Simulations Using Multiple Sequences (with Discussion)." *Statistical Science* 7:457–472.

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2004. *Bayesian Data Analysis, 2nd Edition.* London: Chapman & Hall.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94:653–663.

Gold, David. 1952. "A Note on the "Rationality" of Anthropologists in Voting for Officers." *American Sociological Review* 17:99–101.

Gourevitch, Peter A. 2003. "The Politics of Corporate Governance Regulation: Political Determinants of Corporate Governance: Political Context, Corporate Impact (book review)." *Yale Law Journal* 112:1829–80.

Gourevitch, Peter, Richard Carney and Michael Hawes. 2003. "Testing Political Explanations of Corporate Governance Patterns." Manuscript, available at `http://www.stanford.edu/group/sshi/Finance2003/gourevitch_paper.pdf`.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66:1017–1098.

Heckman, James J., Hidehiko Ichimura and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 64:605–54.

Heckman, James J. and Richard Robb. 1985. *Longitudinal Analysis of Labor Market Data (eds. J. Heckman and B. Singer).* New York, NY: Cambridge Chapter Alternative Methods for Evaluating the Impact of Interventions, pp. 63–113.

Hirano, Keisuke, Guido W. Imbens, Donald B. Rubin and Xiao-Hua Zhou. 2000. "Assessing the effect of an influenza vaccine in an encouragement design." *Biostatistics* 1:69–88.

Holland, Paul W. 1986. "Statistics and Causal Inference (with Discussion)." *Journal of the American Statistical Association* 81:945–960.

Holland, Paul W. 2003. "Causation and Race." *Educational Testing Service: Research Report* 03-03.

Howell, William and Paul E. Peterson. 2002. *The Education Gap: Vouchers and Urban Schools.* Brookings.

Imai, Kosuke. 2004. "Do Get-Out-The-Vote Calls Reduce Turnout?: The Importance of Statistical Methods for Field Experiments." *American Political Science Review, to appear.*

Imai, Kosuke and David A. van Dyk. 2004. "A Bayesian Analysis of the Multinomial Probit Model Using Marginal Data Augmentation." *Journal of Econometrics, to appear.*

Imbens, Guido W. 2002. "Sensivity to Exogeneity Assumptions in Program Evaluation." Manuscript, available at `http://emlab.berkeley.edu/users/imbens/imbens.pdf`.

Imbens, Guido W. and Donald B. Rubin. 1997a. "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." *Annals of Statistics* 25:305–327.

Imbens, Guido W. and Donald B. Rubin. 1997b. "Estimating Outcome Distributions for Compliers in Instrumental Variables Models." *Review of Economic Studies* 64:555–574.

Iversen, Torben and David. Soskice. 2002. "Insurance and Representation: Why do Some Democracies Redistribute More than Others?" Manuscript, presented at 2002 Annual Meeting of the American Political Science Association, Boston, MA.

Kimball, David and Martha Kropf. 2003. "Ballot Design and Unrecorded Votes in the 2002 Midterm Election." *Technical Report, `http://www.umsl.edu/~kimballd/dk_vote.htm`.*

Kinder, Donald R. and Thomas R. Palfrey, eds. 1993. *Experimental Foundations of Political Science.* University of Michigan Press.

King, Gary. 1989. "Representation through Legislative Redistricting: A Stochastic Model." *American Journal of Political Science* 33:787–824.

King, Gary and Langche Zeng. 2002. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." *Manuscript, available at `http://Gking.harvard.edu`.*

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton, NJ: Princeton University Press.

Koppell, Jonathan G.S. and Jennifer A. Steen. 2004. "The Effects of Ballot Position on Election Outcomes." *Journal of Politics* 66:267–281.

Krosnick, Jon A., Joanne M. Miller and Michael P. Tichy. 2003. *Rethinking the Vote (eds. Ann W. Crigler, Marion R. Just, and Edward J. McCaffery).* New York: Oxford University Press Chapter An Unrecognized Need for Ballot Reform, pp. 51–74.

Lassen, David Dreyer. 2003. "The Effect of Information on Voter Turnout: Evidence from a Natural Experiment." Manuscript, available at `http://www.econ.ku.dk/ddl/files/votingnov03.pdf`.

Lee, Jong Wha. June 1993. "International Trade, Distortions and Long-run Economic Growth." *IMF Staff Papers.*

Lijphart, Arend and Rafael Lopez Pintor. 1988. "Alphabetic Bias in Partisan Elections: Patterns of Voting for the Spanish Senate, 1982 and 1986." *Electoral Studies* 7:225–231.

MacKerras, Malcolm H. 1970. "Preference Voting and the 'Donkey Vote'." *Politics: The Journal of the Australasian Political Studies Association* 5:69–76.

Madow, William G. and Lillian H. Madow. 1944. "On the Theory of Systematic Sampling, I." *Annals of Mathematical Statistics* 15:1–24.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Martin, Andrew D., Kevin M. Quinn and Daniel Pemstein. 2002. Scythe Statistical Library, release 0.3. Technical Report.

Masterman, C. J. 1964. "The Effect of the Donkey Vote on the House of Representatives." *Australian Journal of Politics and History* 10:221–225.

McCullagh, Peter and John A. Nelder. 1989. *Generalized Linear Models (Second edition)*. London: Chapman & Hall.

McFadden, Daniel. 1973. *Frontiers of Econometrics (eds. P. Zarembka)*. New York: Academic Press Chapter Conditional Logit Analysis of Qualitative Choice Behavior, pp. 105–142.

Mebane, Jr., Walter R. and Jasjeet Sekhon. 2004. "Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data." *American Journal of Political Science*.

Mebane, Walter R. Jr. 2000. "Coordination, Moderation and Institutional Balancing in American Presidential and House Elections." *American Political Science Review* 94:480–512.

Miller, Joanne M. and Jon A. Krosnick. 1998. "The Impact of Candidate Name Order on Election Outcomes." *Public Opinion Quarterly* 62:291–330.

Milner, Helen V. and Benjamin Judkins. 2002. "Partisanship and Trade Policy: Is there a Left-Right Divide on Trade Policy?" Manuscript, presented at 2002 Annual Meeting of the American Political Science Association, Boston, MA.

Niemi, Richard G. and Paul S. Herrnson. 2003. "Beyond the Butterfly: The Complexity of U.S. Ballots." *Perspectives on Politics* 1:317–326.

Pablo T. Spiller, Ernesto Stein and Mariano Tommasi. 2003. "Political Institutions, Policymaking Processes, and Policy Outcomes: An Intertemporal Transactions Framework." Manuscript, available at `http://www.wws.princeton.edu/~rppe/apr2003conf/Tommasi.doc`.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

Persson, Torsten and Guido Tabellini. 2003. *The Economic Effects of Constitutions.* Cambridge, MA: MIT Press, Forthcoming.

Robson, Christopher and Brendan Walsh. 1973. "The Importance of Positional Voting Bias in the Irish General Election of 1973." *Political Studies* 22:191–203.

Rogowski, Ronald, Eric C. C. Chang and Mark Andreas Kayser. 2002. "Electoral Systems and Real Prices: Panel Evidence for the OECD Countries, 1970-2000." Manuscript presented at 2002 Annual Meeting of the American Political Science Association, Boston, MA.

Rogowski, Ronald and Mark Andreas Kayser. 2002. "Majoritarian Electoral Systems and Consumer Power: Price Level Evidence from the OECD Countries." *American Journal of Political Science* 46:526–39.

Rosenbaum, Paul R. 2nd edition, 2002. *Observational Studies.* New York, NY: Springer Verlag.

Rosenbaum, Paul R. and Donald B. Rubin. 1983a. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society, Series B* 45:212–18.

Rosenbaum, Paul R. and Donald B. Rubin. 1983b. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

Rosenbluth, Frances and Ross Schaap. 2002. "The Domestic Politics of Banking Regulation." Manuscript, presented at 2002 Annual Meeting of the American Political Science Association, Boston, MA.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and non-randomized studies." *Journal of Educational Psychology* 66:688–701.

Rubin, Donald B. 1980. "Comments on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu." *Journal of the American Statistical Association* 75:591–593.

Scartascini, Carlos G. 2002. "Political Institutions and Regulation: An Inquiry on the Impact of Electoral Systems on the Regulation of Entry." Manuscript, available at `http://www.gmu.edu/departments/economics/bcaplan/scar.pdf`.

Scartascini, Carlos G. and W. Mark Crain. 2002. "The Size and Composition of Government Spending in Multi-Party Systems." Manuscript.

Schaffner, Brian F. and Matthew J. Streb. 2002. "The Partisan Heuristic in Low-Information Elections." *Public Opinion Quarterly* 66:559–581.

Scott, James W. 1972. "California Ballot Position Statutes: An Unconstitutional Advantage fo Incumbents." *Southern California Law Review* 45:365–395.

Snyder, James M. and Michael M. Ting. 2002. "An Informational Rationale for Political Parties." *American Journal of Political Science* 46:364–378.

Tomz, Michael and Robert P. Van Houweling. 2003. "How Does Voting Equipment Affect the Racial Gap in Voided Ballots?" *American Journal of Political Science* 47:46–60.

US General Accounting Office. 2001. "Statistical Analysis of Factors That Affected Uncounted Votes in the 2000 Presidential Election." *Report to the Ranking Minority Member, Committee on Government Reform, House of Representatives.*

US Office of Management and Budget. 2003. "Informing Regulartory Decisions: 2003 Report to Congress on the Costs and Benefits of Federal Regulations and Unfunded Mandates on State, Local, and Tribal Entities." *Office of Information and Regulatory Affairs Report.*

Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane Jr., Michael C. Herron and Henry Brady. 2001. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* 95:793–810.

Wilcox, Rand R. 1997. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press.

Wolter, Kirk M. 1984. "An Investigation of Some Estimators of Variance for Systematic Sampling." *Journal of the American Statistical Association* 79:781–790.

Yee, T. W. and C. J. Wild. 1996. "Vector Generalized Additive Models." *Journal of the Royal Statistical Society, Series B, Methodological* 58:481–493.