Thesis submitted in partial fulfilment of the requirements for the degree of DPhil in Politics in the Department of Politics and International Relations at the University of Oxford

Three Essays on Measuring Political Behaviour

Thomas Samuel Robinson

Trinity Term 2020

Word Count: 75,725

Wadham College, University of Oxford

Abstract

Fundamental issues of political representation persist in modern democratic states. Policies favoured by large majorities of the population are not acted upon by elected representatives. Direct forms of democratic policymaking are increasingly used to settle policy disputes. And the amount of money in politics continues to grow. At the same time, new sources of data allow us to test the political behaviour of democratic actors in novel ways. Across three papers, this thesis explores how representation and policymaking operate in modern democratic systems and how researchers should measure political behaviour in these contexts.

In Chapter 2, I ask why popular legislation is passed by direct democratic means instead of by the legislature? I use a combination of district-level voting data, ideological estimates of financial donors, and a survey of state legislators to test competing theories of ballot initiative success in the United States. I argue that direct democratic policymaking is necessary when issues lie outside mainstream policy networks. Legislators appear averse to policies that are novel, untested, or go against the preferences of their financial supporters. In Chapter 3, I show how experimentalists can make valid statistical inferences when their data contains group-level structures. I illustrate this guidance using both Monte Carlo simulations and a review of experimental publications over the past three years. In Chapter 4, I examine how campaign finance disclosure affects vote choice in both representative and direct democratic contexts. Applying the lessons of the previous chapter, I use conjoint experiments to show that while voters can use disclosure information to inform vote choice, these effects are crowded out once partisan, ideological and issue cues are made explicit. Across all three chapters, I develop inferential techniques designed to account for limitations of observed data and, in particular, incomplete data.

Contents

1 Introduction **1**

1.1 Representation, responsiveness, and measurement 1

1.2 Three essays on measuring political behaviour 5

1.3 Key concepts 7

1.4 Contributions 9

References 10

2 Direct democracy in representative systems: Understanding breakdowns in responsiveness through ballot initiative success **13**

2.1 Selection problems and direct democratic policymaking 19

2.1.1 What are the relevant comparison cases? 19

2.1.2 Valid inferences with systematically missing data 21

2.2 Theories of electorally popular but legislatively unpopular policy 25

2.3 Empirical evidence: district level voting behaviour 36

2.3.1 Explanation 1: Disproportionality of preferences across legislative districts 36

2.3.2 Explanation 2: Orthogonal political conflict 39

2.3.3 Explanation 3: Conservative Bias 41

2.4 Empirical evidence: campaign donors' ideology 42

2.4.1 Data and case selection 43

2.4.2 Explanation 5: Active legislative aversion 45

2.4.3 Explanations 3 and 4: conservative bias and partisan legislatures . . . 53
2.5 Empirical evidence: Surveying state legislators directly . . . 56
2.6 Discussion . . . 59
References . . . 61

3 When should we cluster experimental standard errors? 65
3.1 What is uncertainty in an experimental context? . . . 69
3.2 Group-level structure in experimental data . . . 74
3.3 Simulation evidence in favour of the design-based approach . . . 84
3.4 Review of Experimental Papers . . . 93
3.5 Practical guidance for experimentalists . . . 102
3.6 Discussion . . . 105
References . . . 105

4 Knowledge or ignorance? Assessing the causal effects of campaign finance disclosure on vote choice 111
4.1 The informational benefit of disclosure . . . 115
4.1.1 Effects of disclosure with known information . . . 119
4.1.2 Effects of disclosure across types of electoral campaign . . . 121
4.1.3 Relevant facets of disclosure . . . 122
4.2 Conjoint experiment to assess impact of disclosure on vote choice . . . 127
4.2.1 Randomising the *number* of attributes . . . 128
4.2.2 Design . . . 130
4.3 Results . . . 135
4.3.1 The effect of aggregate disclosure . . . 135
4.3.2 Priming subjects' perceptions of candidates' ideology and valence . . . 138
4.3.3 Disclosure's effect on initiative outcomes . . . 141
4.4 Discussion . . . 145

References 147

5 Conclusions **151**

Appendix A Supporting information for Chapter 2 **153**

A.1 Estimating legislator's ideal responses: robustness tests 153

A.2 Ideological distribution of support: further analyses 156

A.2.1 Proportion of extreme initiatives by year 156

A.2.2 Proportion of extreme initiatives by issue area 156

A.3 Robustness tests: stringency of exclusion criteria 162

A.4 Robustness tests: Bootstrapped difference-in-means estimates 166

A.5 Kernel density plots 169

A.6 Factors deterring legislative action, by issue area 182

References 182

Appendix B Supporting information for Chapter 3 **183**

B.1 Uncertainty as a function of randomisation 183

B.2 Derivation of variance 187

B.3 *JEPS* review articles 189

References 193

Appendix C Supporting information for Chapter 4 **197**

C.1 Subject description 197

C.2 Citizens United and its implications for disclosure 200

C.3 Summary of state campaign finance laws 201

C.3.1 Example of initiative conjoint round 203

C.4 Causal assumptions 204

C.5 Additional models 219

References 223

Acknowledgements

The research that ultimately led to this thesis started in 2015, in a meeting with my then undergraduate supervisor, Ursula Hackett. Little did I know that this conversation would lead to five years of research into ballot initiative policymaking, campaign finance regulations, and experimental methods. Along the way, I have been helped by so many kind and inspirational academics. I am indebted to Ursula, Nigel Bowles, Nancy Bermeo, Raymond Duch, Nelson Ruiz, Anna Petherick, Ranjit Lall, and Noah Bacine, all of whom I have learned invaluable lessons from. To others who have provided me with helpful feedback at various stages of this research, including Rachel Bernhard, Karine Van Der Straeten, Lucas Leemann, Arndt Leininger, and Jan Stuckatz, I am exceptionally grateful.

Of course, this thesis would not have been possible without the encouragement, criticism, and patience of my supervisor Andy Eggers. All the errors contained within are mine, but all the good ideas have his fingerprints on them.

Along the way, I have been supported by a whole array of wonderful friends. To Abdul, Musashi, Leonie, Raluca, Florian, Klaudia, Raquel, Benoit, Ilona, Laure, Marco, and Julia – thank you for all your support and friendship, making my time in the Department that much better. To Ali, who has suffered living with me for almost our entire student lives, I will never forget the kindness, distractions, and friendship you have shown me.

To the wonderful Lucy, thank you for listening to me talk about ballot initiatives (more than you thought humanly possible). You have always helped me see the wood from the trees. I love you.

Finally, to my family, who have unceasingly made the sacrifices necessary to enable me to reach this point. I owe you more than I can ever repay.

1 Introduction

1.1 Representation, responsiveness, and measurement

At the heart of the study of political economy is an attempt to understand what mediates the relationship between what citizens want and what political elites are willing to provide. This area of study raises several questions about representation and responsiveness. How do institutions, actors and contextual dynamics enhance or mitigate the political voice of actors within democratic regimes? What affects the decisionmaking of both elites and voters? How do we quantify the extent to which actual political outcomes correspond to our normative expectations about democratic policymaking?

Not only are these questions colossal in scale, but researchers must seek answers to them from finite and often incomplete data. Voter and elite motivations are typically hidden from view and must be inferred through indirect observations. Some political outcomes are systematically unobserved, limiting our ability to observe variance in outcomes. The challenge when studying latent concepts like representation is how to measure the phenomena and its mechanisms from typically incomplete, truncated, or indirect data. In response, new theories (Broockman, 2016), new measurements (Hertel-Fernandez, Mildenberger and Stokes, 2019; Simonovits, Guess and Nagler, 2019), and new sources of data (Bonica, 2014) have been developed in recent years that enable scholars to examine this crucial area of study in more detail.

As our study of the interactions between elites and voters has developed, so too have the

systems we seek to understand. Advanced democratic systems are now huge structures comprised of hundreds of formal and informal institutions, voting populations that number in the tens or hundreds of millions, and permeated by technologies and resources that did not exist even twenty years ago. Consequently, scholars of democratic performance and policymaking have had to keep pace with rapid changes to how democratic politics is conducted in the modern world. The ease with which information can be transmitted digitally, and the scale at which this can be conducted, fundamentally alter how citizens and political elites interact in contemporary systems. Concurrently, the amount of money in politics has grown substantially (Briffault, 2011; Rhodes et al., 2019), and inequalities in political representation are more pronounced (Schlozman, Brady and Verba, 2018).

High financial costs of running for office mean that candidates are heavily reliant on donors to sustain their campaigns (Stratmann, 2005). This is particularly true in the United States where loosening campaign finance laws enable well-funded interest groups and corporations to spend increasingly large sums of money in support of their favoured candidates (Briffault, 2011). The removal of limits on spending have coincided with widening economic inequality across advanced democratic states (Bartels, 2016). Taken together, these changes may ultimately affect the political equality of individuals within democratic states, even as formal political rights remain unchanged (Bonica et al., 2013; Przeworski, 2015; Page and Gilens, 2017; Schlozman, Brady and Verba, 2018).

Despite the increased scale of modern democracy, some elements of modern democratic policymaking are distinctly ancient. In particular, direct democratic modes of policymaking, based on the Ancient Athenian model of governance, are an increasingly common way of altering the policy status quo (Qvortrup, 2014). Finland, for instance, adopted a citizens initiative provision in 2012, using it to pass same-sex marriage legalisation into law. In 2016 the Brexit referendum (initiated by political elites) resulted in fundamental and far-reaching changes to the United Kingdom's constitutional order. In the United States, citizens and interest groups pass policy at each electoral cycle directly on the ballot, bypassing

representative policymaking institutions. Citizens themselves have written and passed public smoking bans, limits on the government revenue generation, wildlife and animal protections, and the legalisation of formally prohibited drugs like marijuana. In each case, policy change was achieved through direct, rather than representative, means.

The rising use of direct democratic policymaking is not surprising. Citizen satisfaction with democracy in general has been declining over the past two decades (Foa et al., 2020). At the same time, large majorities of Americans demand quite considerable policy change on issues that are not being met by conventional, legislative policymaking processes (Gilens, 2015; Page and Gilens, 2017). Direct, initiated means of policymaking, whereby citizens write and pass their own laws, allow voters to bypass intransigent legislators (Magleby, 1994). Ballot initiative provisions and referendums provide concurrent policymaking venues, leading to interactive effects between policymakers within and outside of the legislature (Gerber, 1996).

In general, these contemporary dynamics have forced scholars to reconsider questions about democratic representation and responsiveness in modern democratic systems. Recent work has sought to quantify how well elected officials respond to public opinion (Caughey, Warshaw and Xu, 2017; Caughey and Warshaw, 2018), and highlight instances where legislators do not always respond to voters' demands (Lax and Phillips, 2012; Hertel-Fernandez, Mildenberger and Stokes, 2019). New work on policy congruence has sought to quantify the the (financial) extent of incongruence between policy status quo and majority preferences (Simonovits, Guess and Nagler, 2019). At the voter-level, too, changes to campaign finance regulation and the amount of money in politics have spurred new research not only into how spending affects the outcomes of political races (Stratmann, 2005, 2006, 2010; Figueiredo et al., 2011; Brown, 2013), but also the decisionmaking of voters themselves (Dowling and Wichowsky, 2014, 2015; Wood and Spencer, 2016; Wood, 2017).

Renewed interest in these questions has also highlighted important gaps in our knowledge. We still know very little about the interaction between direct and representative forms of

policymaking, despite the prominence of referendum and ballot initiatives in the world today. Why do citizens have to resort to legislating themselves when they elect democratic representatives to act on their behalf? While the growth of super-PACs, unlimited donations, and "dark money" entities in the United States have fuelled new research on campaign finance transparency, the extent of disclosure's effects on voters behaviour remains contested, the interaction with partisanship unclear, and the differential effects across types of election understudied.

Concurrently, new sources of data and statistical methods have become increasingly accessible to scholars of political representation. Campaign finance records at the individual-level have been digitised across states in the US. Sub-constituency level voter returns are more common, allowing researchers to analyse actual voter behaviour in greater detail than before. New quantitative methods that can handle larger quantities of data are now commonplace in our analyses (Grimmer, 2015). More concretely, the abundance of political data means we can now estimate the ideology not just of political elites but of individual citizens too (Bonica, 2014). Advances in experimental designs like conjoint analyses allow researchers to recover multiple causal estimates of how phenomena affect vote choice efficiently (Hainmueller, Hopkins and Yamamoto, 2014).

Limitations of finite data, systematically unobserved components, and alternative causal pathways continue to exist, even as the depth of data available to researchers has increased. With the advent of new data sources and methods come concerns about their inferential validity. In particular, insufficient attention has been paid to substantive questions where missing or unobserved data are a consequence of the phenomena under study. In other words, how do we make valid inferences when some portion of the observed data is systematically missing? Separately, how should we treat and therefore estimate the uncertainty of experimental estimates? How do we isolate alternative causal pathways or mitigating factors that could undermine observed causal effects? Answering these questions are crucial to improving the validity of our claims not only within the study of representation and political economy,

but across the social sciences more broadly too.

1.2 Three essays on measuring political behaviour

In this thesis, I focus on two foundational issues within the study of political representation. First, what explains breakdowns in the responsiveness of legislators to the preferences of their constituents? Second, how do voters respond to information about the funding profiles of political campaigns? Across both questions, I seek to characterise how different forms of democracy interact within the same system, and how actors' behaviours differ between them. In doing so, I advance our understanding of the mechanisms of representation and responsiveness, drawing on new insights from across the political economy literature.

These two questions also prompt a series of inferential and methodological challenges, that I seek to resolve in this thesis. Across the three individual papers, I develop a series of inferential tools that extend our ability to make claims about representation, policymaking and electoral dynamics. More specifically, I focus on how scholars can and should handle counterfactual cases within empirical research. Counterfactual reasoning rests on inferences about cases we cannot observe. It is inherent to any statistical procedure that seeks to estimate the uncertainty around a point estimate (as discussed in Chapter 3). Counterfactual reasoning also extends to cases where portions of the data are systematically unobserved – for instance, in the study of rare events where we cannot isolate similar cases from other observable data (King and Zeng, 2001). Counterfactual reasoning occurs in many experimental works, where efficient estimation strategies aim to recover multiple causal effects from limited numbers of observations (Hainmueller, Hopkins and Yamamoto, 2014).

Breakdowns in policy responsiveness. In Chapter 2, I examine why some popular policies pass by direct democracy rather than through the legislature. In this paper, I develop and test a series of competing theories, using novel empirical data, to advance our understanding of policy responsiveness. I use district-level voting data, nearly 3 million estimates of cam-

paign donors' ideological positions, and a small survey of state legislators to test why state policy is incongruent with citizens' preferences, and thus why direct democracy succeeds.

I argue that initiatives are used to pass popular policy when the issue has not taken root in the mainstream policy community. Moderate correlations between partisan and initiative voting suggest that initiative issues are not fully captured by the partisan dimension of conflict. Many initiatives, but not all, moderate away from the position of the majority party and towards the minority party. Evidence from campaign finance data suggests successful initiative campaigns attract more ideologically extreme donors than do successful legislative candidates in the same cycle. Legislators themselves report being aware of policy incongruence and point to the political sensitivity of issues and countervailing interests of donors.

Clustering experimental standard errors. In Chapter 3, I provide a simplified explanation of when experimentalists should cluster standard errors. This research incorporates recent insights from the econometrics literature revising how researchers should interpret uncertainty in randomised settings (Abadie et al., 2014, 2017, 2020). In contrast to the technical guidance in these papers, I addresses the paucity of applied guidance in the literature on this crucial feature of estimation. I show how randomisation affects the data generating process in the presence of group-level features using Monte Carlo simulations. I present a comprehensive review of all experiments published between 2017 and 2019 in the *Journal of Experimental Political Science* to illustrate this issue in practice, replicating three studies to show that incorrectly estimating errors can lead to variance estimates that can be substantially biased. While this research is methodologically focussed, its conclusions have important implications for how we treat uncertainty in applied research. As experimental research methods become more and more common in political economy (and the social sciences more broadly), researchers should take seriously the problem of variance estimation.

The effects of campaign finance disclosure. Finally, Chapter 4 explores how voters use campaign-finance information to inform their voting decisions. I focus in particular on the United States, where limits on donations and spending have been progressively weakened over recent decades (Jiang, 2018). I address several limitations of existing studies on the effects of disclosure. First, few studies have tested the impact of disclosure on voters' electoral decisionmaking. Instead most papers have sought to test whether disclosure affects voters' perceptions of candidates, irrespective of whether that affects their vote choice. Second, little work has sought to assess directly how partisanship and ideology interact with disclosure mechanisms. Finally, to the best of my knowledge, no studies have compared whether the effects of disclosure differ between representative and direct democratic electoral campaigns.

To address these gaps in the literature, I conduct a series of conjoint experiments, applying the estimation guidance in Chapter 3. I compare how the same disclosure mechanisms affect vote choice across both gubernatorial and ballot initiative races. For the gubernatorial experiment, I also randomise whether subjects are exposed to additional political cues. This strategy allows me to measure directly the influence of partisan signals on the causal efficacy of disclosure. I find that the effects of disclosure are limited once partisan and ideological signals are made explicit. This null effect is important precisely because I separately show that subjects can and do use various features of a candidate's financial profile to influence their vote choice when these signals are unavailable. Results from the ballot initiative conjoint experiment suggest a similar result: disclosure is ineffective in the presence of explicit policy choices that prime voters' decisionmaking.

1.3 Key concepts

Several key terms recur throughout this thesis. I briefly outline these definitions here for the sake of concision and consistency. These concepts – representation, congruence, and responsiveness – form the conceptual foundation upon which the substantive focus of this

thesis rest.

Representation can refer either to descriptive or substantive ways in which an actor, institution, or process acts on behalf of someone else (Carnes and Lupu, 2016). An actor or body is descriptively representative if they share key physical, socioeconomic, or cultural attributes. Substantive representation on the other hand is the extent to which an actor's policy preferences are represented by another actor. If a constituency demands some policy position, an elected official is substantively representative to the extent they also advocate for that position. An elected official therefore can be substantively representative, by advocating for policy on behalf of their constituents, without necessarily being descriptively representative of that constituency (and vice versa).

Congruence is a measure of the distance between policy positions or preferences. Policy congruence is typically treated as the difference between the position of enacted legislation and the ideal point the electorate (Simonovits, Guess and Nagler, 2019). Consequently, the less congruent policy is with voters' preferences, the less substantively representative the legislature is to the electorate as a whole. Legislators may nevertheless be descriptively representative of their community, or a particular voter, and substantively representative of certain interests. Crucially, how these terms coincide will depend on how substantive preferences are represented. I discuss this further in Chapter 2.

Responsiveness is an active concept. It is the extent to which an actor adjusts their own behaviour on the basis of another. A legislator, for instance, is responsive if they adjust their own legislative preferences on the basis of the preferences of their constituents. Politicians can be responsive without necessarily fully converging on the ideal point of the statewide majority (Caughey, Warshaw and Xu, 2017). For instance, voters may demand universal healthcare. Legislators are conceivably responsive if they shift policy from a private system to a hybrid public-option system, even if they resist calls to nationalise the healthcare system. Whereas congruence measures the distance between actors or preferences in policy space, responsiveness refers to movement towards or away from congruent positions.

1.4 Contributions

Chapter 5 discusses the specific conclusions of each paper and their wider implications for the study of political economy. Here I briefly note two broader contributions this thesis makes to the study of political behaviour and representation.

First, this thesis provides new ways to measure and account for political behaviour across observational and experimental contexts, providing general guidance on how to implement these frameworks in other research contexts. Not every question we ask in the study of political economy, or political science more generally, can or should be answered with a difference in means or linear regression model. This thesis makes clear that even when the predominant type of empirical evidence is quantitative and large-N, researchers have to be sensitive to underlying limitations of the data-generating process. I demonstrate several novel approaches that can be employed to make valid inferences in the face of these limitations.

Second, the substantive focus of Chapters 2 and 4 pose normative questions about the representative role of elected legislators in democratic systems. Given multiple competing principals – the parties that select candidates, donors who fund campaigns, and voters who ultimately place them in office – what should politicians do? In this thesis I provide evidence that suggests legislators face trade-offs between personal convictions about policy, the demands of constituents, and financial pressures.

Underlying this research, therefore, is a fundamental concern with political inequality in democratic systems (Bartels, 2016; Schlozman, Brady and Verba, 2018). The results in this thesis are consistent with the notion that unequal distributions of resources, and the policy networks that exclude certain actors, lead to biased representational outcomes. This thesis demonstrates how electoral dynamics like partisanship can ameliorate the informative benefit of financial disclosure that, theoretically, would allow voters to correct for the undue influence of narrow but influential groups.

These issues are pressing. Loosening campaign finance regulations and the presence of big donors in policy networks threatens to undermine the political equality that forms the

keystone of modern democratic governance. The remainder of this thesis provides novel evidence of these dynamics, and more broadly helps demonstrate how political scientists can adapt their inference to keep pace with the way in which democratic systems are developing.

References

Abadie, Alberto, Susan Athey, Guido Imbens and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" *arXiv:1710.02926 [econ, math, stat]* .

Abadie, Alberto, Susan Athey, Guido W Imbens and Jeffrey M Wooldridge. 2014. Finite Population Causal Standard Errors. Working Paper 20325 National Bureau of Economic Research.

Abadie, Alberto, Susan Athey, Guido W. Imbens and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88(1):265–296.

Bartels, Larry M. 2016. *Unequal Democracy : The Political Economy of the New Gilded Age*. Second edition. ed. New York: Russell Sage Foundation.

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58(2):367–386.

Bonica, Adam, Nolan McCarty, Keith T Poole and Howard Rosenthal. 2013. "Why Hasn't Democracy Slowed Rising Inequality?" *Journal of Economic Perspectives* 27(3):103–124.

Briffault, Richard. 2011. "Nonprofits and Disclosure in the Wake of Citizens United." *Election Law Journal: Rules, Politics, and Policy* 10(4):337–361.

Broockman, David E. 2016. "Approaches to Studying Policy Representation: Studying Policy Representation." *Legislative Studies Quarterly* 41(1):181–215.

Brown, Adam R. 2013. "Does Money Buy Votes? The Case of Self-Financed Gubernatorial Candidates, 1998–2008." *Political Behavior* 35(1):21–41.

Carnes, Nicholas and Noam Lupu. 2016. "Do Voters Dislike Working-Class Candidates? Voter Biases and the Descriptive Underrepresentation of the Working Class." *American Political Science Review* 110(4):832–844.

Caughey, Devin and Christopher Warshaw. 2018. "Policy Preferences and Policy Change: Dynamic Responsiveness in the American States, 1936–2014." *American Political Science Review* 112(2):249–266.

Caughey, Devin, Christopher Warshaw and Yiqing Xu. 2017. "Incremental Democracy: The Policy Effects of Partisan Control of State Government." *The Journal of Politics* 79(4):1342–1358.

Dowling, Conor and Amber Wichowsky. 2014. The Effects of Increased Campaign Finance Disclosure: Evaluating Reform Proposals. SSRN Scholarly Paper ID 2483194 Social Science Research Network Rochester, NY: .

Dowling, Conor M. and Amber Wichowsky. 2015. "Attacks without Consequence? Candidates, Parties, Groups, and the Changing Face of Negative Advertising: The Changing Face of Negative Advertising." *American Journal of Political Science* 59(1):19–36.

Figueiredo, De, John M, Chang Ho Ji and Thad Kousser. 2011. "Financing Direct Democracy: Revisiting the Research on Campaign Spending and Citizen Initiatives." *The Journal of Law, Economics, and Organization* 27(3):485–514.

Foa, R. S., A. Klassen, M. Slade, A. Rand and R. Williams. 2020. The Global Satisfaction with Democracy Report 2020. Technical report Centre for the Future of Democracy Cambridge, United Kingdom: .

Gerber, Elisabeth R. 1996. "Legislative Response to the Threat of Popular Initiatives." *American Journal of Political Science* 40(1):99–128.

Gilens, Martin. 2015. "Descriptive Representation, Money, and Political Inequality in the United States." *Swiss Political Science Review* 21(2):222–228.

Grimmer, Justin. 2015. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science & Politics* 48(01):80–83.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(01):1–30.

Hertel-Fernandez, Alexander, Matto Mildenberger and Leah C. Stokes. 2019. "Legislative Staff and Representation in Congress." *American Political Science Review* 113(1):1–18.

Jiang, Lear. 2018. Disclosure's Last Stand? The Need to Clarify the 'Informational Interest' Advanced by Campaign Finance Disclosure. SSRN Scholarly Paper ID 3262349.

King, Gary and Langche Zeng. 2001. "Explaining Rare Events in International Relations." *International Organization* 55(3):693–715.

Lax, Jeffrey R. and Justin H. Phillips. 2012. "The Democratic Deficit in the States." *American Journal of Political Science* 56(1):148–166.

Magleby, David B. 1994. "Let the Voters Decide - An Assessment of the Initiative and Referendum Process." *University of Colorado Law Review* 66:13.

Page, Benjamin I. and Martin Gilens. 2017. *Democracy in America?: What Has Gone Wrong and Waht We Can Do about It*. The University of Chicago Press.

Przeworski, Adam. 2015. Economic Inequality, Political Inequality, and Redistribution. Working Paper Department of Politics, New York University.

Qvortrup, Matt. 2014. *Referendums around the World: The Continued Growth of Direct Democracy*. Basingstoke: Palgrave Macmillan.

Rhodes, Samuel C., Michael M. Franz, Erika Franklin Fowler and Travis N. Ridout. 2019. "The Role of Dark Money Disclosure on Candidate Evaluations and Viability." *Election Law Journal: Rules, Politics, and Policy* 18(2):175–190.

Schlozman, Kay Lehman, Henry E. Brady and Sidney Verba. 2018. *Unequal and Unrepresented : Political Inequality and the People's Voice in the New Gilded Age*. Princeton: Princeton University Press.

Simonovits, Gabor, Andrew M. Guess and Jonathan Nagler. 2019. "Responsiveness without Representation: Evidence from Minimum Wage Laws in U.S. States." *American Journal of Political Science* 63(2):401–410.

Stratmann, Thomas. 2005. "Some Talk: Money in Politics. A (Partial) Review of the Literature." *Public Choice* 124(1-2):135–156.

Stratmann, Thomas. 2006. "Is Spending More Potent For or Against a Proposition? Evidence from Ballot Measures." *American Journal of Political Science* 50(3):788–801.

Stratmann, Thomas. 2010. Campaign Spending and Ballot Measures. In *Financing Referendum Campaigns*, ed. Karin Gilland Lutz and Simon Hug. London: Palgrave Macmillan UK.

Wood, Abby K. 2017. "Show Me the Money: 'Dark Money' and the Informational Benefit of Campaign Finance Disclosure." *SSRN Electronic Journal* .

Wood, Abby K. and Douglas M. Spencer. 2016. "In the Shadows of Sunlight: The Effects of Transparency on State Political Campaigns." *Election Law Journal: Rules, Politics, and Policy* 15(4):302–329.

2 Direct democracy in representative systems: Understanding breakdowns in responsiveness through ballot initiative success

Policy referendums around the world succeed regularly and on important policy areas. But why do these policies pass by direct democracy and not through the legislature? In this paper, I address this puzzle by focussing on ballot initiative success in the United States. Using a combination of district level voting data, ideological estimates of campaign donors, and a small survey of state legislators, I test why state policy is incongruent with citizens' preferences and thus why direct democracy succeeds. Taken together, my results suggest that initiatives are used to pass popular policy when the issue has not taken root in the mainstream policy community. Moderate correlations between partisan and initiative voting suggest that initiative issues are not fully captured by the partisan dimension of conflict. Many initiatives, but not all, moderate away from the position of the majority party and towards the minority party. And evidence from campaign finance data suggests successful initiative campaigns attract more ideologically extreme donors than do successful legislative candidates in the same cycle. Finally, I present qualitative survey evidence to show that legislators are aware of the responsiveness gap and their own aversion to passing certain popular policies.

Policy is approved regularly by voters at the ballot box across the world, and over a broad range of issue areas. From the "Brexit" referendum in the United Kingdom, to same-sex marriage legalisation in Ireland, and an array of economic and social policies passed in the United States, referendums are a powerful and frequent mode of policy change across advanced democratic nations (Qvortrup, 2014). In the United States, in particular, the initiative process provides a separate but concurrent policymaking venue (Jourdain, Hug and Varone, 2017) with limited restrictions on the types of policy citizens can propose (Smith and Tolbert, 2007). As Figure 2.1 demonstrates, citizen-initiated means of legislating in the US are a prevalent form of policymaking.[1]

Initiative success, however, is theoretically puzzling. Under the ideal conditions noted by Downs (1957) and Schumpeter (1976), elites and parties engage in a competitive struggle for the people's vote. In competitive democracies like the United States, when a policy is supported by a majority of voters we should expect a party to advocate and claim that policy position (Manza and Cook, 2002; Gerber and Lewis, 2004; Leeman and Wasserfallen, 2016; Page and Gilens, 2017). If they do, direct democratic policymaking should not succeed. Even if politicians are initially hesitant to pass the majority's preferred position, citizens can threaten to override the status quo by proposing their own legislation. This threat should force legislators to "crowd out" the initiative policy and converge towards the median voter (Gerber, 1996).

Despite these expectations, initiatives continue to pass every electoral cycle across the United States. Of course, there are numerous reasons why the Downsian logic will fail in practice, but work to date has yet to explain why in a systematic fashion. Previous explanations have focussed on the uncertainty of voters' positions (Matsusaka and McCarty, 2001; Hug, 2004). If parties are uncertain over support for a policy, it is possible that the legislature will fail to converge on the median voter allowing initiative proponents to set a

[1]State-level initiative processes exist in 24 states, and over half of all American cities have some form of initiative provision. In total, around 70 percent of the US population has access to one or more direct democratic procedures (Matsusaka, 2005).

position preferred by the electorate. But this theory does not comport with the typically high levels of support successful initiatives garner. The average vote share for successful initiatives was nearly 64 percent in 2018, and 27 of these 28 measures were supported by over 55 percent of voters. The high proportion of votes successful initiatives receive suggest parties should know that these policies are popular, and therefore whether to support their passage through the legislature.

Arguments from uncertainty also fail to capture other important aspects of the electoral context that impact parties' willingness to pass popular policies. Theories of initiative policy-making should take seriously institutional features that may preclude legislative convergence (Boehmke, Osborn and Schilling, 2015), the role interest groups play in influencing legislative outcomes (Hertel-Fernandez, Mildenberger and Stokes, 2019), the systematic biases in legislators' perceptions (Broockman and Skovron, 2018), and normative policy judgements made by legislators themselves. As initiatives continue to pass by wide margins in the US, and new sources of data become available to researchers, we should re-examine the success of initiatives taking into account these important factors.

In this paper, I explore five different explanations for why parties fail to adopt popular policies, and thus why initiatives succeed. First, does the concentration of interests within a small number of legislative districts impact the ability of legislators to converge on the statewide median voter? Second, are legislators ignoring issues where the issue dimension is orthogonal to the partisan dimension, making it hard to form coalitions capable of passing these popular policies? Third, do strong majority party agenda-setting powers preclude legislative action even when a majority of legislators would endorse policy change? Fourth, are legislators systematically biased in their perceptions of where the median voter is? Finally, might legislators be unwilling to pass certain policies even if they know they are electorally popular due to sectional interests or normative policy concerns?

Each of these explanations have distinct implications about the level and distribution of support for initiatives. To test these theories, therefore, I present empirical evidence using a

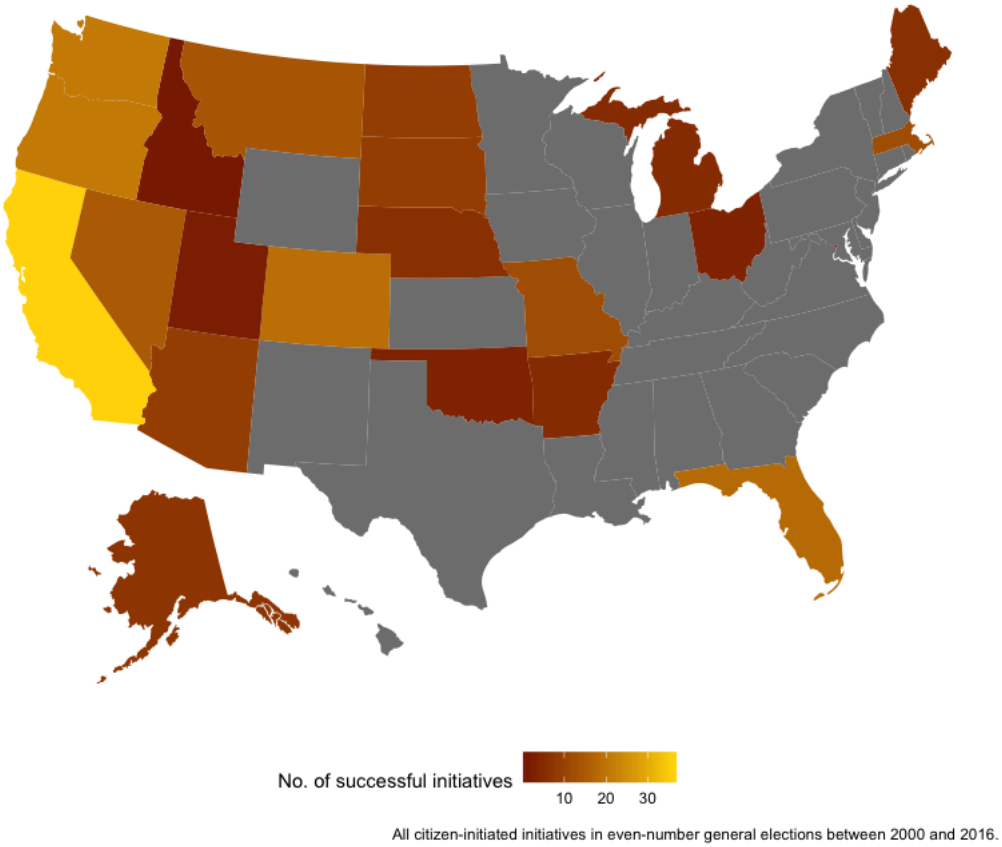


Figure 2.1: Successful initiative usage across the United States (2000-2016).

The colour gradient indicates the number of initiatives passed between 2000 and 2016. States shaded grey have no initiative provisions or did not pass an initiative during this time period. Data source: Initiative & Referendum Institute, University of Southern California (2018) and National Conference of State Legislature (NCSL; 2018).

diverse range of different data that allows me to assess the electoral, financial, and legislative behaviour of both voters and actors within state policymaking processes. In particular, this paper presents the results from three different data generating processes.

First, I compile and match precinct and district-level voting data for 194 initiatives across six states, manually scraped from the respective states' official filings. I merged these vote returns with districting information to calculate the actual district-level vote-share for initiative and presidential elections. This data enables me to estimate the "ideal" behaviour of legislators based on their own districts' majority preference. I use these calculations to estimate the proportion of legislative support each initiative *should* have received in the leg-

islature had the same initiative bill come to the floor. Finally, I compare the initiative voting patterns with corresponding precinct and district-level vote shares for presidential elections to assess the correlations between issue voting and overtly partisan races.

Second, I isolate the political action committees (PACs) that supported successful ballot initiatives across all initiative states, and merge this information with financial contribution data from the Database on Ideology, Money in Politics, and Elections (DIME; Bonica, 2016). I focus my analysis on the financial support coalitions of 77 successful initiatives to successful candidates between 2000 and 2012 (13,500 legislators across all years and states). I then compare the financial supporters of these initiatives to the financial supporters of successful candidates to estimate the relative position of support on an initiative-by-initiative basis. In total, these estimations are calculated on the basis of nearly three million unique donors.

Finally, I present small-n evidence from a survey of legislators in five high-usage initiative states. Fielded in 2019, this survey complements the quantitative analysis in the rest of the paper by directly probing the motivations and concerns of legislators. The data provides direct evidence of legislators attitudes towards not only the initiative process, but perceptions of their role as representatives. Legislators were asked, across a battery of questions, to consider why popular policies may nevertheless be acted on by the legislature.

I find that while initiative voting and partisan voting are correlated, the extent of this correlation is reasonably limited. This evidence is consistent with the theory that policy passed by initiative tends not to be fully captured by dominant partisan dimension of political conflict. Moreover, most (but not all) successful initiatives tend to moderate away from the majority party's position. The majority party seems unwilling to adjust its position on certain issues, even when a majority prefer such a shift. At the same time, supporters of successful initiative campaigns are more "extreme" than supporters of successful legislative candidates. Taken together, this evidence suggests that successful initiatives are popular policies that have not taken root in the mainstream policy community.

There are various reasons why policy might lie outside the mainstream. It could be because

they are untested, possibly unwise (at least to conventional policymakers), or because they are opposed by powerful financial interests. Using the responses to the legislative survey, I show that state legislators consistently point to both the role of financial interests and political sensitivity of the issues as reasons that deter legislative action.

Conversely, the empirical evidence suggests that how preferences are distributed across legislative districts, is not a convincing explanation for why initiatives succeed. Were legislators to vote according to the expressed preference of their constituents, almost all successful initiative policies should have been passed in the legislature. Moreover, policies do not appear to be predominantly liberal-leaning, suggesting that conservative bias plays a limited role in determining whether initiative policies are passed at the ballot.

Testing these theories is empirically challenging precisely because it involves selecting on the dependent variable. While we observe cases where legislators fail to act on majority public opinion (successful initiatives), we are unable to directly observe the relevant counterfactual – those initiatives that were not proposed because the legislature had passed the popular policy. Selection problems of this sort often pose serious inferential issues for research designs (Geddes, 1990; Mintz, 2009). In this paper, however, I show how researchers can be sensitive to selection issues when the relevant counterfactual is systematically unobserved. Using insights from Bayesian theory, I present a simple formalisation of the selection problem that better enables researchers to make inferentially valid claims in the face of incomplete data. This framework for handling difficult empirical contexts has broader utility beyond the focus of this paper.

Substantively this paper presents a new lens through which to understand policy incongruence. The broader literature on responsiveness dynamics makes clear that legislators often do not fully converge on the demands of the electorate. While state policy is responsive – more liberal states have more liberal policy and *vice versa* (Erikson, Wright and McIver, 1993; Caughey and Warshaw, 2018) – it nevertheless remains out of kilter with majority preferences even as it moves in the right direction (Lax and Phillips, 2012; Caughey, Warshaw

and Xu, 2017; Broockman and Skovron, 2018; Simonovits, Guess and Nagler, 2019). Ballot initiative success is symptomatic of (previous) policy incongruence. These cases represent instances where legislators seemingly could have passed policy with greater congruence to the state's median voter but did not. This paper shifts our attention towards cases where we observe policy incongruence being resolved: cases where ballot initiatives succeed.

In the next section, I develop a simple framework to understand policy incongruence with respect to direct democracy. I directly address the empirical challenge of constructing valid tests in the face of systematically unobserved data, and how this conditions the analysis in the remainder of this paper. In Section 3.2 I then use this framework to set out five separate explanations for why popular legislation is passed by initiative rather than by the legislature. In Sections 2.3-2.5, I test these explanations using the voting, ideological and qualitative evidence respectively.

2.1 Selection problems and direct democratic policymaking

Initiatives succeed when legislators fail to pass policy that a majority of voters demand. Policy entrepreneurs place these items of policy directly on the ballot and citizens vote to enact them. The fundamental puzzle is therefore why legislators act on some popular policies, but not on others? Put another way, why do we observe policies that are electorally popular but which are not supported by the legislature?[2]

2.1.1 What are the relevant comparison cases?

Consider the graph depicted in Figure 2.2, where policy support is disaggregated into electoral and legislative support respectively. The four quadrants indicate ideal types of policy

[2]Implicitly, this framing assumes that policies proposed directly at the ballot are done so because they were, or would be, unsuccessful within the legislature. Again, this feature is largely unobservable. Intuitively, however, the costs associated with private citizens or interested groups pursuing direct legislation are higher than the costs of legislatures doing the same (Gerber, 1996). Therefore, that policy entrepreneurs pursue this costly action suggests that, at least from their perspective, the legislature will not consider the policy in question despite its popularity.

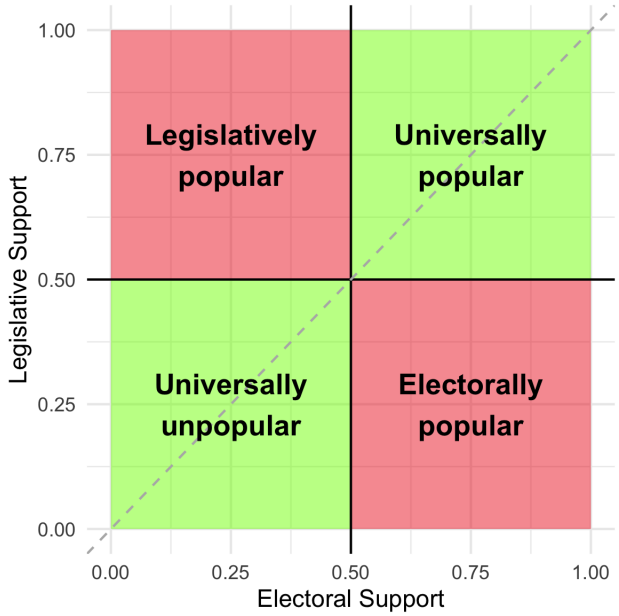


Figure 2.2: Hypothetical forms of policy proposals disaggregating legislative and popular support.

support: universally popular policies favoured by both legislators and voters, those that are electorally popular but legislatively unpopular, universally unpopular policies, and legislatively popular but electorally unpopular policies. In a majoritarian system, all policy proposals should fall along the positive diagonal - either universally unpopular or universally popular. If a majority of the public favours a proposal, a majority of legislators should also favour it (and vice versa). As a result, all policies with greater than 50 percent support in the electorate (and no others) should be passed into law by the legislature.

Breakdowns in policy congruence occur when policy proposals end up in either the top-left or bottom-right quadrant of Figure 2.2. "Legislatively popular" policies are those where legislators pass policies despite opposition from a majority of the public. I do not consider these cases in this paper. Instead I focus on those cases in the bottom-right quadrant – policies with a majority of popular support, but where legislators are resistant to change. These policies typify successful ballot initiatives. Since legislators do not support the proposals, policy entrepreneurs instead submit them to the ballot where citizens override the legislature's intransigence. But why, if a majority of the electorate support a proposal, are

legislators intransigent in the first place?

The ideal comparison to answer this question is between successful initiatives posed on the ballot (where the legislature failed to act), and those popular initiative policies *not posed* because legislators acted on the majority's preference. This is because the puzzle is *conditional on* majority electoral support for the policy. Crucially we are not trying to explain why some initiatives succeed when others fail.[3] Comparing between successful and unsuccessful initiatives is not informative because initiatives that fail are by definition unpopular. In which case, it is not puzzling that legislators do not act on these proposals. Instead, the issue is why some policies are located in the bottom-right instead of the top-right quadrant.

2.1.2 Valid inferences with systematically missing data

The ideal comparison, however, poses an interesting empirical problem for researchers. Since universally popular policies are enacted by legislators, the electorate never votes on them. Therefore, by dint of legislative support, we never observe these policies on the ballot. Consequently, we only ever observe the "negative" cases in which legislators do not support the proposal.

Given the systematic absence of data, we should rightly be concerned about inferences made on the basis of observed cases alone, or "selecting on the dependent variable" (Geddes, 1990; Mintz, 2009). Selection problems, however, do not have to be fatal to empirical research designs. In the remainder of this section, I formalise the selection problem to motivate precisely why it is problematic in comparative designs. I then show how researchers can mitigate some of these problems. In short, by accumulating evidence that refines our priors over counterfactual cases, researchers can make inferentially valid claims even when some portion of the data is systematically missing.

[3]There is an entire literature on understanding the causal drivers of initiative success (Gerber, 1999; Smith and Tolbert, 2007; Stratmann, 2010; Figueiredo et al., 2011; Adams, 2012; Richards, Allender and Fang, 2013). This question would be answered by comparing initiative policies located in the bottom-left and bottom-right quadrants.

Suppose ten children all score highly in a test, and a researcher notes that each of these children reported watching cartoons the night before. They conclude from this data that watching cartoons is good for test performance. Most of us would balk at this claim, and rightly so. How do we know that children who scored poorly did not also watch the same cartoons? In which case, cartoon-watching could not explain the difference in test scores. In other words, it is very easy to construct alternate hypotheses commensurate with the observed data but where the suggested mechanism is inert.

Naively interpreting commonalities between successful initiatives can be similarly misleading. If it is plausible that some same set of features common to successful initiatives would also be true of the counterfactual unproposed initiatives, then we should not conclude that those features cause initiative success. Since we cannot observe the counterfactual cases, we seemingly have little way of knowing whether observed commonalities across successful initiatives are also common to unproposed initiatives too. By making inferences about why successful initiatives occur only on the basis of observed cases, we risk making inferential errors associated with selecting on the dependent variable.

To generalise this problem, assume there is some comparative hypothesis H that states ϕ is responsible for a given outcome. For simplicity's sake, let us assume the outcome can either be positive $(+)$ or negative $(-)$. H is more consistent with some observed data than others. That is, if H is true, we expect some joint distribution over the outcome $D_H = \{d_{\phi,+}, d_{\neg\phi,-}\}$, those positive (negative) cases that are (not) coincident with ϕ. In Bayesian terms, this is the likelihood $P(d_{\phi,+}, d_{\neg\phi,-}|H)$. Some alternative hypothesis, H_a, is associated with a different joint distribution of the data such as $D_{H_a} = \{d_{\phi,+}, d_{\phi,-}\}$ – a hypothesis, for instance, under which the observed feature is present among both positive and negative outcomes. To assess the probability of H being true we sample observations from the real world and assess how similar the draw is to the joint distributions expected under H compared to the alternative H_a.

In ideal comparative designs, we are able to sample observations that encompass the full

joint distribution of positive and negative outcomes. We take some draw of data $D' \sim \{d_+, d_-\}$, comprised of both positive and negative outcomes, and assess the similarity of the drawn data with the expectations under H compared to H_a. For example, we may compare whether children watched cartoons across those who did well and those who did poorly in the test. Or we may compare electoral rules between systems with many and few parties. In the initiatives puzzle, however, $D' \sim \{d_-\}$; we only observe features of campaigns that actually took place, namely where legislative support was insufficient for the policy to pass in the legislature. We cannot observe those initiatives left dormant because legislators acted in accordance with the majority preference. In which case, the full joint distribution of data is unobserved and it is not possible to compare across outcomes. The fundamental problem, therefore, is making valid inferential claims when some portion of the joint distribution of cases is systematically missing.

But suppose the ten students in the previous example had all been found to have taken a rare drug known to boost attention. Again the researcher claims the drug improves test scores on the basis of these ten high-performing students alone. While the validity of a claim about the drug's effects is not as strong as if the researcher had conducted a randomised drug-trial, this claims seems more plausible than the cartoon-watching theory. Why? Because despite not sampling from the full joint distribution of data, we have some prior belief that this drug's use is highly uncommon among the population. In other words, we believe it is highly unlikely that we would observe the drug's usage in the unobserved cases.

More generally, selecting on the dependent variable is not problematic simply because the researcher made an inference on the basis of incomplete data. Rather, selection issues are problematic when we think there are many alternate and *plausible* hypotheses commensurate with the observed data in which the proposed mechanism has no effect. A researcher's claim is invalid when, with the subset of data available to them, they could not discriminate between these plausible hypotheses. Consequently, selection problems are not entirely intractable. In some instances, researchers can explicitly assess the strength of their hypotheses given the

observed data *and* a range of plausible distributions for the unobserved data. If the theory is resilient to these likely alternatives, there is a useful, substantive claim to be made using the incomplete data.

The magnitude of the selection problem, or inversely the strength of our convictions in the face of incomplete data, depends on how many plausible alternatives we can rule out. Suppose we have the hypothesis that A is smaller than B, where $A, B \in [0, 1]$. We observe that $A = 1$, but B remains unobserved. Note that given A is the maximum value, there is no plausible joint distribution over A and B in which B is larger than A. Therefore, despite not observing B, we can logically refute the hypothesis. In this instance, there is no selection problem despite systematically missing data.

If $A < 1$, there are possible values of B greater than A and so we cannot deductively falsify the hypothesis. We can assess the probability of $B > A$, however, given some prior over B (even though we do not observe B in the data). Most simply, we might assume an uninformative uniform prior over the possible values of B and let $A = 0.95$. The probability that the hypothesis holds is $1 - A = 0.05$. It is reasonably unlikely, in other words, that B is larger than A. The uniform prior, however, may not be a good assumption about the distribution of B. If, in fact, B is more likely to be close to 1 then the uniform prior is not a useful way of assessing the posterior probability of the hypothesis. The supplementary objective for researchers working with incomplete data, therefore, is to produce informative priors over the missing component.

Put simply, when a hypothesis cannot be deductively falsified, the validity of that claim will rest on the validity of our priors. To overcome the selection problem, researchers need to assemble evidence to inform these prior beliefs, which in turn improve the validity of our inferences. As I demonstrate at various points in this paper, this additional evidence can come from alternative sources of data that indirectly strengthen our prior beliefs. The goal is not to quantify the unobserved values themselves, but to refine our beliefs over their possible values. To the extent this is possible, researchers can make informative claims from

incomplete data that would typically be dismissed due to selection issues.

In practical terms, this inference strategy requires a two-pronged approach. First, and as with typical analyses, empirical tests should be designed to discriminate between theories given the observable data. Second, when making inferences about a given theory researchers should incorporate assessments of how likely the theory is to hold across a range of *plausible* values for the unobserved portion of the data. To do so, they can assemble evidence that helps generate informative priors over the systematically unobserved component of the model.

This strategy provides researchers with a principled way to approach inference when some portion of the data is systematically unobserved. The claims made using this approach are invariably weaker than when researchers can sample from the full joint distribution of data. They are nonetheless a useful way to make sense of important political phenomena while taking selection problems seriously. Systematically missing data occur in substantively important areas of study, like rare event studies (King and Zeng, 2001), and selection issues often arise in small-n analyses more generally. This approach is also useful for larger quantitative studies, as I illustrate in the remainder of this paper. Careful analyses can help researchers learn important things about the world even in the presence of incomplete data.

2.2 Theories of electorally popular but legislatively unpopular policy

This section outlines five different explanations for why some popular policies have to be passed by ballot initiative. These theories capture separate intuitions about policy responsiveness, building on recent theoretical advances in the literature. In each case, I outline the empirical expectations of the theory, which I subsequently test using voting, donations, and survey data in Sections 2.3-2.5.

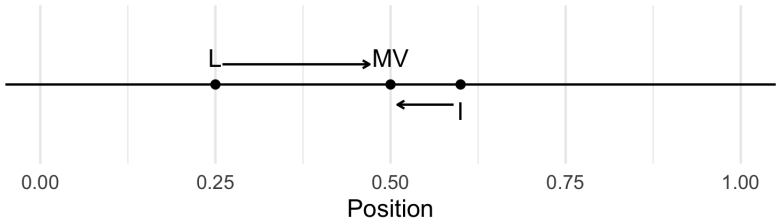


Figure 2.3: Uni-dimensional model of initiative-legislature interplay, following the reduced logic of Gerber (1996). Individual points indicate the ideal policy position of the legislature (L), median voter (MV), and initiative proponent (I) respectively. Arrows indicate the hypothetical movement of policy positions by L and I as they converge on MV.

The foundational model. Formalisations of initiative-legislative interactions typically model policy contestation in a one-dimensional policy space, along which the legislature, initiative proponent and median voter are arrayed (Gerber, 1996). The objective for both the legislature and initiative proponent in these models is to minimize the distance between the final policy position and their ideal point. All players have perfect information of each others' positions.

For instance, Figure 2.3 depicts a hypothetical set of ideal points. Assuming actors are rational, L will shift their position to converge on MV (since I strictly prefers MV to all positions between L and MV) – at which point, there is no pareto optimal policy that can trump L at the ballot and I withdraws their threat of initiative.[4] In other preference orderings, L will adjust their position just until the initiative proponent is no longer the most proximate point to the median voter. Of course, the legislature may well find itself having to offset the positions of various potential proponent groups. In which case, they may still end up converging on MV. In all cases, as Gerber (1996) demonstrates, the legislature will revise the status quo to crowd out the initiative proponent. As a result, initiatives should never succeed.

Few expect this rational actor model to comport with reality and, indeed, ballot initiatives succeed every electoral cycle in the United States. Revisions to the foundational model have

[4]The additional assumption here is that I and L incur some cost c to legislating, but $c_I > c_L$.

predominantly focussed on limiting the extent to which the legislature can converge on the true position of the median voter. In the case of imperfect information, the legislature may simply misjudge how far it must shift its policy proposal (Matsusaka and McCarty, 2001; Hug, 2004). In pivotal actor models, supermajority requirements limit the ability of the legislature to pass new policy (Boehmke, Osborn and Schilling, 2015). These modifications suggest that convergence dynamics are disrupted by institutional or behavioural features of legislative politics that complicate how and whether legislators respond to majority demands in the electorate.

These broader policymaking dynamics deserve further attention. New theoretical developments in the literature on policy responsiveness can be applied to refine our understanding of initiative success, including: the role of parties and partisanship (Cox and McCubbins, 2005; Shor, Berry and McCarty, 2010); systematic biases in legislators' perceptions (Broockman and Skovron, 2018); the influence of interest groups (Hertel-Fernandez, Mildenberger and Stokes, 2019); and normative policy judgements by legislators themselves. The five explanations I discuss below incorporate these concerns, challenging or extending the foundational model, to help us better understand why initiatives succeed. Collectively, I focus on explanations that encompass instances where legislators are responsive (but the electoral system is not), where they are insufficiently responsive, or where they are deliberately unresponsive to the majority's preferences.

Electoral Disproportionality. While the entire state electorate is eligible to vote on ballot initiatives, individual legislators are selected by non-overlapping subsets of the same electorate. The legislature may not act if a majority of these districts do not favour a policy. This feature is problematic if a policy nevertheless is favoured (relative to the status quo) by a majority of the state electorate as a whole.

Assume legislators are only concerned about their constituents' preferences, and are perfectly responsive to them. A legislator will vote for any proposal their constituency median

voter prefers to the status quo, and reject any policy further away. Individual legislators do not take into consideration the statewide median voter when deciding whether to support a proposal. Sufficient imbalances in the density of a proposal's support across districts will lead to instances where each legislator is fully responsive to their district's median, but the legislature as a whole is not responsive to the demands of the statewide median.

To illustrate this implication, suppose we have a state with ten electoral districts, each comprising 100 voters.[5] Three districts have large majorities in favour of the proposed policy - say, 80 percent. In the other seven districts, only 40 percent favour the initiative policy. If all voters turn out, 52 percent of the state electorate would be in favour of the initiative policy. Seven out of the ten legislators, however, would vote against that policy given their districts' median preferences. Most importantly, these seven state legislators are still perfectly responsive with respect to their own constituents.

At the state level, the breakdown in responsiveness is a function of how preferences are aggregated at the state level and the imbalance of support across legislative districts. Because support for a proposal is concentrated in relatively few districts, a majority of legislators oppose policy change that is nevertheless supported by a majority of voters in the state.

The observable expectation for this theory is simple: successful initiatives should have majority support in less than half of all legislative districts. More broadly, the theory entails that for those counterfactual (and unobserved) initiative policies that never get proposed, majority support should more often be distributed across a majority of legislative districts. In other words, legislators act in the interest of their constituents, and policy gets passed by the legislature when it is supported by a majority of district median voters, irrespective of the preference of the statewide median voter.

Issue orthogonality. Alternatively, what if the particular issue does not relate to the partisan dimension along which other major issues are decided? If an issue is orthogonal to

[5]Note this argument does not rely on imbalances in population across districts. Of course, if voters are unevenly distributed across legislative districts, the potential for incongruent state level outcomes increases.

the partisan dimension, and legislators are unwilling to engage in bipartisan action, popular proposals may be ignored in the legislature.

Consider a specific issue dimension, D_I, and the liberal-conservative partisan dimension, D_P. The correlation between voting along these two dimensions, ρ, ranges between 0 to 1 (in absolute terms).[6] Let us also assume that there are two parties in the legislature – Party A and Party B – and along D_P the rightmost legislator in A is to the left of the leftmost legislator in B. The two parties in other words are polarised along the partisan dimension - mirroring contemporary legislative dynamics (Barber and McCarty, 2015; Thomsen, 2014; Andris et al., 2015; McCarty, Poole and Rosenthal, 2016).

If $\rho = 1, D_I = D_P$ and legislators in Party A and Party B are perfectly divided on the issue. As ρ approaches 0, however, the extent to which D_P determines D_I diminishes, and as a consequence we would expect the composition of support coalitions across this issue to be comprised more equally of members of Party A *and* B. Suppose some policy on D_I, p_I, commands a majority of support in the legislature and the electorate (it is universally popular). If ρ is sufficiently low, then the legislative coalition behind this support is likely to be bipartisan.

Formally, nothing above precludes the legislature passing p_I. However, party polarisation has led to 'concrete behavioural effects' within American legislatures (Miller and Conover, 2015, p.225), increasing hostility towards the opposition and reducing cooperation across party lines (Mann and Ornstein, 2012; Gluck, O'Connell and Po, 2015; Aldrich and Battista, 2002; Shor, Berry and McCarty, 2010). These dynamics, therefore, combined with the assumptions made above may preclude bipartisan action irrespective of the size of the supportive legislative coalition. As a result, the legislature may be unable to shift its position towards p_I. The resultant incongruence between legislative and electoral preferences means initiative proponents can pass the policy by ballot initiative.[7]

[6] A correlation of $-1 < \rho < 0$ indicates that the issue dimension is inverted with respect to the direction in which the partisan dimension is ordered.

[7] A weaker version of this theory would simply state that orthogonal issues are less valuable to parties, because the partisan signal they send to voters is weaker. Party labels in low-information contexts like state

If this theory holds, we should expect successful initiatives to be uncorrelated with the partisan dimension of conflict. Moreover, this theory implies that as the relationship between partisan and issue dimensions increases, the ability of a party to "own" and pass a policy issue without having to create bipartisan linkages also increases. Therefore counterfactual initiative policies that we do observe are ones where the corresponding issue dimension is more partisan.[8] Moreover, the extent of the relationship between these two dimensions helps us to gauge the extent to which any given issue dimension has been captured by conventional partisan forces.

Partisan legislatures. Other features of partisan politics may prevent legislative convergence towards the median voter. One potential explanation is that the majority party refuses to engage in legislation where the median voter lies away from its own ideal position. Given its agenda-setting powers, it can prevent the legislature from shifting the policy – even if legislators' positions on the whole would support a policy shift towards the state median.

Suppose the majority party has near total control of the legislative agenda (cf. Cox and McCubbins, 2005), and that the majority party acts under a Hastert Rule (HR) – only legislating on positions favoured by a majority of the majority party. There is no guarantee that this position, as determined by HR, includes the electoral median voter. For a given proposal, there may be sufficient support from a small number of members of the majority party, as well as the minority party, to command an overall legislative majority, despite a majority of the majority party favouring the status quo.

Suppose Party A has 60 seats and Party B has 40 seats. Some policy p is favoured by the electorate over the status quo. All members of Party B are in favour of p, and so too are 20 members of Party A. In total, a clear majority in the legislature would like to pass p (thus improving congruence with the state electorate). But this policy won't pass under

legislative races provide key information shortcuts to voters (Schaffner and Streb, 2002; Coan et al., 2008). In order to focus on the salient partisan priorities, parties may want to preserve the apparent distinction between their brands by ignoring orthogonal issues.

[8]Note, this theory is not a comparative claim against unsuccessful initiatives. Unsuccessful initiatives may simply be orthogonal cases that also happen to be electorally unpopular.

the conditions outlined above. Since 40 members of Party A, and thus a majority of the majority party, favour the status quo, the majority party under HR will not move p to the floor. And since the minority party has little control over the agenda, they cannot force the issue. p is not passed, leaving open the option for an initiative proponent to submit the legislation directly to the ballot.

Under the foundational model, the majority party should anticipate this chain of events and shift its position away from the party's median position. But if individual legislators are resistant to shifting their position (because they are unwilling to abandon their party's programme), then these convergence mechanisms may fail, at least in the short run. Ultimately, this theory rests on the assumption that legislators have party concerns in addition to the proximity of legislation to their individual ideal points. Consequently, at time, they are stubborn.

The key implication of this explanation is that successful initiatives should be those policies where support for the policy lies away from the majority party. In Democrat-controlled states, initiatives should be supported by more centrist and Republican-leaning voters, and vice versa in Republican-controlled states. Those unproposed popular initiatives are, conversely, ones where a majority of the majority party support the policy preferred by the state median voter.

Conservative bias. Alternatively, legislators may attempt to converge on the median voter but be systematically mistaken about where the median voter lies. Legislators (of both parties) are known to overestimate the extent of voters' conservatism (Broockman and Skovron, 2018; Hertel-Fernandez, Mildenberger and Stokes, 2019). If legislators are wrong about the position of the median voter, then the legislature will not adjust its position fully. This lack of full convergence leaves a policy winset open to initiative proponents to pass majority-preferred legislation.

Imagine a single legislator who attempts to maximise policy congruence by passing all

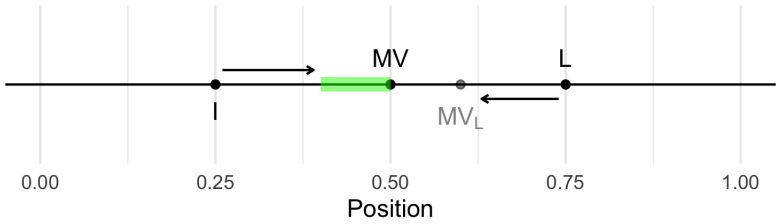


Figure 2.4: Uni-dimensional model of initiative-legislature interplay when the legislature faces a conservative bias constraint. L misperceives the median voter's position as MV_L. The green shaded area indicates the possible policy space in which an initiative proponent can pose a policy favourable to the median voter, which the legislature will not crowd out.

policies they believe to be favoured by the majority. An unbiased legislator (with perfect information) is able to decide accurately which policies to adopt and which to reject. In this world, there would be perfect policy congruence. Any policy favoured by a majority in the electorate would be favoured by the legislator (as represented by the diagonal line in Figure 2.2.)

But if the legislator's responsiveness function is biased to the right, then there is a "blindspot" where the legislator will not respond sufficiently. Consider Figure 2.4, where the preferences are ordered $I < MV < L$. Given the conservative bias, L fails to converge fully on the actual position of MV. The shaded green area represents policy positions initiative proponents can propose at the ballot but which the legislator will not converge on. The ordering of preferences does matter here. When $L < I < MV$, the legislature will still converge on I even though it perceives MV to be further to the right than it actually is. In general, if legislators exhibit conservative bias and have an ideal position to the right of the median voter, there will exist some winset of policy positions that command majority support in the electorate but which the legislature will not converge.[9]

By implication, successful initiative policies should be left-leaning. In contrast to the partisan legislatures explanation, this expectation has a singular direction (towards the liberal

[9]Legislators in this model are not wilfully unresponsive. Each legislator is assumed to respond as best as they can, resulting in a convergence on MV_L. In some sense, therefore, this explanation reintroduces uncertainty into the model. Unlike generic plane uncertainty, however, the form considered here is bounded with a definite direction.

end of the policy spectrum) irrespective of which party holds a majority within the legislature. And in terms of the relevant counterfactual cases, these should be instances where legislators possess less conservative bias and therefore the proportion of left-leaning cases should be essentially zero. In other words, different issues or policies are affected by conservative perceptual biases to different extents, and legislators are more able to adequately converge on certain issues than others.[10]

Active legislative aversion to policies outside the political mainstream. What if legislators simply refuse to act on certain issues? Legislators may be averse to certain policy areas or policies, despite their *known* electoral popularity. In particular, legislators may resist passing new legislation when the policy lies outside the mainstream policy networks – those actors, parties, and institutions that regulate conventional legislative processes.

There are numerous reasons why legislators may be averse to particular policies that are nevertheless popular. Policies that are relatively novel, and thus untested, may have uncertain social or economic implications, and not be incorporated into party's agendas. In which case, risk-averse legislators (and parties) may decide not to pass new legislation despite broad electoral support. Legislators make a trade off between immediate electoral reward (the conventional payoff associated with policy convergence) and the potential long term impact on their electoral support from enacting risky policy.

Alternatively, it may be that legislators exercise their own policy expertise (acting as "Burkean" trustees) to arrive at different policy preferences to their constituents. In this sense, legislators may consider certain policies unwise even if they are electorally popular. Again, these policies are likely to be those that lie outside the policy mainstream (and are thus not captured by either party). In which case, legislators face an intrinsic trade-off between setting policy positions they deem to be normatively acceptable and what a majority want at any given time.

[10]This adaptation to the conservative bias theory can be extended and formalised in greater detail. Given that the full implications of this theory for ballot initiatives success are unobserved, I leave this avenue open for future research.

Or, the status quo may be preferred by interested groups who donate to political parties. These unelected actors in policy networks influence conventional policymaking through their control of political resources. Politicians, dependent on donations to finance campaigns, may discount the preferences of the majority in favour of their donors (Barber, 2016*b*). If candidates' responsiveness prioritises the interests of donors, then legislators may oppose certain policy positions even if they are electorally popular (Francia et al., 2005; Przeworski, 2015; Hill and Huber, 2017; Page and Gilens, 2017; Klüver and Pickup, 2019). Recent experimental survey evidence suggestS that legislators' knowledge of constituency policy preferences are, in fact, biased by interest group contact at both the national and state level (Hertel-Fernandez, Mildenberger and Stokes, 2019; Stokes, Mildenberger and Hertel-Fernandez, N.d.).

In sum, the policy responsiveness of legislators may be mediated by other factors – such as the issue's novelty, the economic costs of the policy, unknown policy implications, or resistance by influential donors. These considerations lessen the importance of the median voter's demands on legislators and parties. And common across these factors, the policy defies aspects of the mainstream policy community – it is not supported by the major parties or donors. Successful initiatives, therefore, should be supported by actors who lie outside this mainstream network.

On the other hand, unproposed popular initiatives are those policies that were already incorporated into mainstream policy networks. This intuitively makes sense – legislators are happy to act on policies that are commensurate with, and accepted by, those actors that operate within regular policy networks. The corresponding empirical expectation, that we cannot observe, unproposed initiatives lie closer to the mainstream policy community than do unproposed popular initiatives.

Summary. Table 2.1 summarises the expectations of each of the five theories, and briefly describes the empirical tests I use to test the expectations of each. Given the range of different

explanations, I use three different sources of data – actual voter returns at the district or, in some cases, precinct level; donations data and corresponding ideological estimates of these donors; and direct survey evidence from state legislators themselves.

Using different sources of data also enable me to test the same explanations at different levels. For example, under the conservative bias theory – supportive evidence can be observed in terms of the correlation of support between partisan and initiative voting, and separately in the ideological position of financial donors. Collectively, the evidence from each test helps inform our broader understanding about the conditions under which popular policies pass by initiative.

Explanations	**Tests**			
	Aggregation	*Partisan Correlations*	*Donor Support*	*Legislator Survey*
1: Electoral disproportionality	Proportion of legislative districts in favour of an initiative is less than 0.5	–	–	–
2: Orthogonal issue dimensions	–	Low correlations between partisan and initiative voting behaviour	Uniform support of donors across ideological dimension	–
3: Conservative bias	–	Democrat-leaning support for initiatives	Liberal-leaning support coalitions for successful initiatives	–
4: Partisan legislatures	–	–	Ideological support closer to minority party position	–
5: Active aversion to policies outside the political mainstream	–	Correlations between partisan and initiative voting that are less than perfect (i.e. $\rho < 1$)	Support coalitions lie away from partisan positions	Awareness of incongruence, role of interest groups and political sensitivity of issues

Table 2.1: Summary of plausible explanations and corresponding expectations.

2.3 Empirical evidence: district level voting behaviour

In this section, I use district level voting data to assess whether our expectations hold for the first three explanations. Data were collected from the Secretary of State websites for each of six states – Alaska, California, Colorado, Massachusetts, Ohio, and Washington – for initiatives held between 2000 and 2016. These states vary in terms of legislative control (Republican, Democratic and split control) and are geographically diverse. Where district level data were not directly available (Alaska, Colorado, Ohio and Massachusetts), I computed district-level aggregations from the precinct-level returns.[11]

2.3.1 Explanation 1: Disproportionality of preferences across legislative districts

The first theory argues legislators fail to pass the policy because a majority of legislative districts have median voters who do not favour the policy in question. To test this explanation, I calculate each district's level of support for an initiative. If a district's level of support is greater than 50 percent, the ideal response of that legislator is to vote in favour of the initiative (coded as "1", otherwise "0"). The mean ideal response is the proportion of legislators who should have voted in favour of the initiative had the initiative been posed as a bill within the legislature. Proportions less than 0.5 are evidence in favour of the disproportionality explanation. I calculate the ideal response of individual state legislators for every ballot initiative proposal where adequate data was available (194 initiatives in total).[12]

Figure 2.5 plots the results graphically against the statewide vote share each initiative

[11] Acquiring precinct-level data for a large number of states was very difficult. Electoral returns for statewide races like initiatives are typically reported by state counties, not all of which report precinct or state district level results. Moreover, one cannot simply infer district vote breakdowns from county data alone since a single legislative district may overlap portions of two (or more) counties.

[12] I present results for the ideal behaviour of the lower chamber alone. For the states in question, with the exception of Massachusetts, the upper chamber districts are made up of two or more lower chamber legislative districts. The difference between the chambers is typically, but not always, the length of legislative term. Given the strength of findings for the lower chamber, and for the sake of brevity, I do not address bicameral issues in this paper.

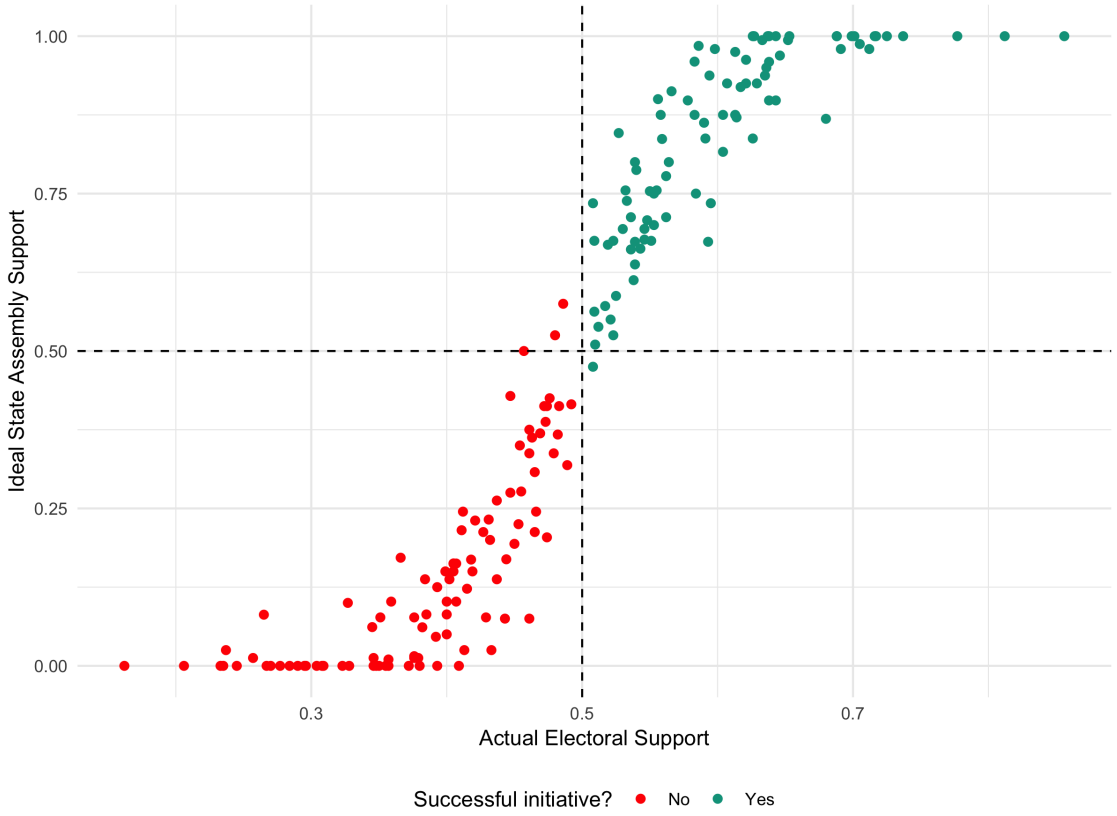


Figure 2.5: Comparison of the ideal support within the legislature to actual support in the electorate for initiatives. Ideal support is calculated as the proportion of legislative districts that voted in favour of the initiative proposal. Green (red) points indicate initiative proposals that received more (less) than 50 percent of the statewide vote.

received. If the theory holds, for the observed data, we should observe successful initiatives occupying the lower-right quadrant of the voting space. In other words, while a majority of voters within the state favoured the initiative over the status quo, a majority of district median voters did not. In which case, a majority of legislators (acting on behalf of their constituencies) should have voted against the hypothetical bill (had it been posed on the floor).

The results across all six states overwhelmingly, however, reject the electoral disproportionality argument. Every *successful* initiative except one (N = 90) across the six states should have had more than 50 percent support in the legislature, had the bills been posed in the legislature. In fact, given the steepness of the trend line it is clear that the aggregation of

electoral preferences works near perfectly. As electoral popularity increases, ideal legislative support rises at a 'faster' rate. By about 65 percent electoral popularity, we should expect the legislature to be near unanimous in its approval of the policy proposal.

Consistent with the pattern for successful initiatives, those issues that lack majority support in the electorate also lack a majority of ideal support in the legislature, for all but two initiatives (both in California) while one further Californian initiative would command exactly half of the legislature's support. Despite the apparent majority of ideal support in the legislature, representatives evidently did not pass these bills, since each ended up on the state ballot.

These findings are robust to more risk-averse assumptions about legislative behaviour. Suppose that legislators abstain if their district's support for an issue is marginal (operationalised as a vote share less than 52.5 percent). Excluding these legislators, the results are unchanged and we observe no cases where an initiative is electorally popular but legislatively unpopular[13] If legislators vote against rather than abstaining on any marginally popular policies, six successful initiatives are pushed into bottom-right quadrant that reflects electorally popular but legislatively unpopular policy. Hence, there are a small handful of cases where the electoral disproportionality may impact vote choice, but this only occurs when we impose conservative criteria about legislative behaviour. The vast majority of initiatives, nevertheless, remain in the universally popular quadrant of support.

The correlation between ideal legislative support and popular vote shares is so strong that despite the absence of counterfactual cases the likelihood of the full hypothesis is very low. In fact, since all bar one successful initiative had a majority of support from legislative districts, the full hypothesis would only hold if every unproposed initiative had both majority statewide and district support. Only in that case would it be true that successful initiatives had more concentrated support. Even then, however, the difference would be negligible and insubstantial.

[13] All robustness tests are reported in the Appendix.

These findings therefore suggest that popular policies do not end up on the ballot because preferences for change are unequally distributed across legislative districts. The almost perfect clustering of successful initiative cases within the upper-right quadrant of Figure 2.5, in fact, only reifies the puzzle at hand. Why are legislators not acting on clear majority preferences within the state electorate when their local constituents desire change?

2.3.2 Explanation 2: Orthogonal political conflict

To assess the orthogonality of initiative policy I compare voters' behaviour on an overtly partisan electoral decision to their behaviour on initiatives. Specifically, I regress precinct/district level vote shares for initiatives on presidential vote shares in the same election. Presidential elections are highly salient and partisan contests that activate first-dimension, liberal-conservative political values. If policies that end up on the ballot are those pertaining to orthogonal issue areas, we would not expect to observe significant or substantive correlations between presidential vote choice and ballot initiative support (or opposition).

As above, I look at all initiative elections held in the same six states, except I only take the subset that occurred in presidential election years (N=110). For each initiative, I regress the precinct/district presidential vote share (the smallest unit available) on the initiative vote share:

$$V_{ij} = \alpha + \beta \times \text{Dem}_j + \epsilon,$$

where V_{iJ} is the vote share in favour of initiative i in precinct/district j, and Dem_j is the two-way Democratic vote share (versus Republicans) in the presidential election in the same area.[14] I include unsuccessful initiatives for the sake of comparison.

Overall, 101 of the 110 initiatives have a statistically significant relationship between Democratic presidential vote share and the initiative vote share. Moreover, looking only at

[14]I note that there is some potential for an ecological inference fallacy here. There is no guarantee that as the absolute number of voters correlate, that it is the same voters across the two contests. The evidence presented here is purely descriptive, and inferences from these relationships should be treated with caution.

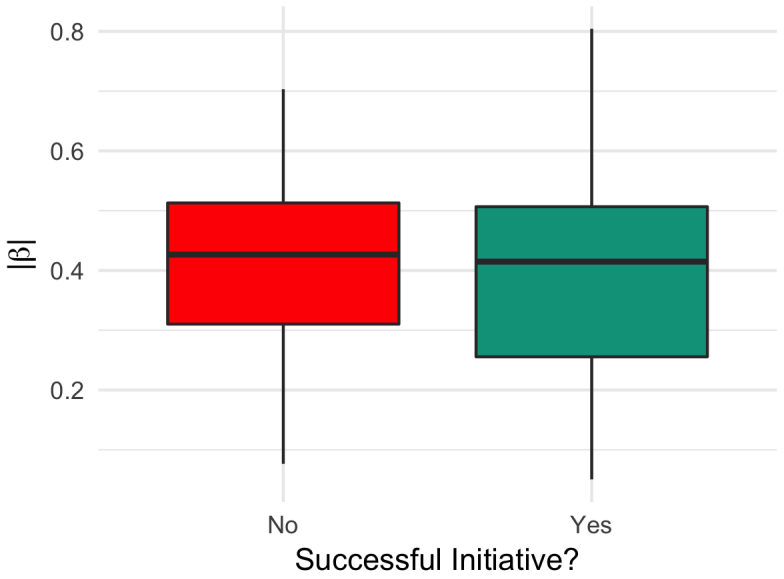


Figure 2.6: Boxplot of initiative-specific bivariate regression on Democratic presidential vote share.

those successful initiatives, 57 out of 59 return a significant coefficient. Figure 2.6 shows the interquartile ranges and median value for both successful and unsuccessful initiatives' absolute beta coefficients (excluding non-significant results).

The median absolute value of the beta coefficient across these successful initiatives is 0.41, which is sizeable though clearly not perfectly correlated. Voter's decisions on initiatives appear to partially covary with Democratic presidential candidate vote shares. Most successful ballot initiatives have a correlation between 0.25 and 0.6, which is suggestive that, while voting across initiative and presidential elections are not like-for-like, at least a substantial part of voters' behaviour corresponds to the partisan dimension of conflict. The strict claim that successful initiative policies are entirely orthogonal from the partisan dimension of conflict does not seem to hold. The median absolute size of the beta coefficient for unsuccessful initiatives is very similar at 0.43, with a narrower interquartile range.

The *observed* data is not consistent with the theoretical expectations of the orthogonality explanation. Therefore, it is less essential to examine the potential partisan correlations for unproposed initiatives. We may still, however, want to gauge how significant the correlations

are compared to other types of political races. It may be that no race correlates very highly and therefore the observed coefficients are near the top of the plausible range. In other words, do other overtly partisan races exhibit much higher correlations with presidential voting? Finding a consistent comparison race across states and electoral cycles is beyond the scope of this paper. There is, however, some suggestive evidence from individual races.[15] In Washington state, for instance, voting for the Lieutenant Governor had a regression coefficient of 0.89 – higher than any initiative correlation and almost double the average. While a single datapoint is not generalisable, it is at least further suggestive evidence that initiatives are not decided on fully partisan grounds alone – partisan correlations nearing 1 are feasible and observed.

2.3.3 Explanation 3: Conservative Bias

Next I test whether the evidence supports the conservative bias explanation: that legislators overestimate voters' conservatism on the liberal-conservative dimension (Broockman and Skovron, 2018). If this explanation holds, the legislature should set policy positions too far to the right. Initiative proponents can therefore propose policy more *liberal* than this legislative point, and garner majority electoral support.

To do so I analyse the direction of the coefficients generated in the previous section. Where initiative vote shares at the district level are positively (negatively) correlated with democratic vote shares, I take these to be indicative of liberal-leaning (conservative-leaning) measures.

In total, 40 of those 57 significant correlations exhibit positive correlations with Democratic presidential vote share. That is, around 70 percent of initiatives are liberal-leaning. There is at least *prima facie* evidence to suggest the conservative bias explanation has some leverage.

As a robustness check, I disaggregate this distribution by state legislative control: if the conservative bias holds, we would expect there to be a strong liberal-lean across all three

[15]Not all state legislative seats are contested, senate and gubernatorial races are staggered across cycles in different states, limiting the ability to generate comparable data.

types of control (Democratic, Republican, and split control).

Table 2.2 displays these results. Clearly, there is an imbalance in the sample towards initiatives passed in Democratic-controlled states. This feature is unsurprising; California has a disproportionate number of successful initiatives and it also has persistent Democratic control of the legislature over the time period in question. Under split control, the vast majority of these (76 percent) have a positive correlation with Democratic vote share. The number of initiatives passing under Republican-controlled states in the sample is too small to make valid inferences from.

Legislative Control	$\beta > 0$	$\beta < 0$
Democrat	23	12
Republican	4	1
Split	13	4

Table 2.2: Distribution of correlation coefficients by legislative control.

In summary, therefore, the evidence is mixed. Initiatives do tend to correlate with Democrat rather than Republican presidential voting, even in states not controlled by the Democratic Party. However, a considerable number of conservative-correlated initiatives still pass in Democratically-controlled states, which is inconsistent with the conservative bias explanation of initiative success. This explanation requires more evidence, therefore, and in the next section I use individual-level donor estimates to further examine the relative support of initiatives vis-a-vis parties.

2.4 Empirical evidence: campaign donors' ideology

In this section, I compare the ideology of donors to successful initiative campaigns to that of the financial backers of seated legislators. This data helps us assess the active aversion theory. And when merged with contextual data on legislative control, I can assess whether this divergence is correlated with majority status (partisan legislatures explanation), or whether support leans towards the liberal pole (conservative bias explanation).

I first show how support bases for successful initiatives differ from those of successful candidates. I demonstrate that the large majority of successful initiatives have mean ideological support more "extreme" than that of either legislative party. Moreover, I show that there is some evidence for the partisan legislatures explanation, but that this evidence does not support the conservative bias theory.

2.4.1 Data and case selection

Between 2000 and 2012 (for which the relevant data is available), 148 *successful* initiatives were passed across all states that have the initiative provision.[16] For each campaign I used the respective Secretary of State website and PAC listings in NIMSP to generate a list of supportive political action committees for each successful ballot initiative. I merge this data with ideal point estimates of campaign finance donors (Bonica, 2014) to assess the relative distribution of ideological support for successful initiatives and legislative candidates within the same electoral cycle. For each supportive PAC, I generate a distribution of donors' ideological positions. In total, this data captures 94,289 initiative donors who gave a cumulative 152,241 times.

I compare these initiatives to the relevant state-cycle subsets of the 13,506 successful legislative candidates who ran for state legislative office in the same time period and states – including both lower and upper chambers. For each legislator I include all donations made by individuals to their respective candidate PAC. In total, this legislative set of ideology estimates captures 2.9 million contributors who made 3.8 million donations collectively to successful legislative candidates.

All ideological estimates – common-space campaign finance scores (CFscores) – are taken from the Database on Ideology, Money in Politics, and Elections (DIME), which contains contribution-level data for state and federal level campaigns between 1979 and 2014.[17] The

[16]This section focusses solely on successful political campaigns - that is, candidates who are elected to state legislative office and initiatives that pass the ballot. Failed initiatives represent cases where legislators were right not to act since initiative proponents misjudged the majority preference of the electorate.

[17]In particular, I leverage the state-level subset of this database which draws data from the National

CFscore assesses ideological similarity based solely on donation activity. Donors who donate to the same political entities (candidates or PACs) are assumed to be ideologically similar. The CFscore is highly correlated with existing ideal point estimation techniques for legislators yet the universe of estimated positions is much larger (Bonica, 2014).[18]

Out of the 148 initiatives that passed in this time-period, I was able to isolate the key support groups for 124 campaigns. The missing 24 cases lacked donor or PAC information on both the respective state repository and NIMSP. This missing data can largely be attributed to initiatives held earlier in the time period, where records have not been digitised. The DIME database also lacks donor data for some initiatives held earlier in the time period: for instance the database does not include any records for Washington ballot initiative PACs in the 2000 and 2002 general elections, for instance.[19] Finally, the number of unique contributors to each initiative campaign varies substantially.[20] To ensure my results are meaningful, I exclude those initiatives with less than 40 unique contributors, leaving 77 initiative campaigns in the dataset.

Institute for Money in Politics (NIMSP) (National Institute on Money in State Politics, 2017).

[18]The CFscore has come under criticism for deriving candidate point estimates from donors who are a potentially biased subset of the voting population (Hill and Huber, 2017). This criticism focusses on the subject of the measurement itself, rather than the methodology of inferring ideological positions from the actions of candidate's supporters. If donors are atypical in terms of their political support, then measuring ideology based on donation data may produce biased estimations of ideology. Essentially, under this view, the CFscore mistakes donors for random draws from the larger distribution of voters' ideological positions.

There are, however, *prima facie* reasons to think that ideal point estimates *should* focus on donors (Barber, 2016*a*,*b*). Donors to political campaigns are predominantly individuals and not corporations (Barber, Canes-Wrone and Thrower, 2017). This subset of the electorate are active participants in political contests (Ansolabehere et al., 2003), who use financial contributions to support candidates and campaigns within electoral races. Crucially for the purposes of this paper, moreover, the measure is consistent across electoral races. Even if the point estimates are not accurate, that should not impact the relative position of CFScores vis-a-vis initiative and candidate campaigns.

[19]Ballot initiative PACs often do not need to be registered specifically to a single campaign, since initiatives are submitted by a chief petitioner. Thus, multiple committees may register in favour of (or in opposition to) an initiative proposal. I include all committees registered in support of an initiative. Moreover, committees may register in support and opposition to multiple campaigns. In these cases, donors to these PACs are not explicitly supporting a single issue. I include these PACs since I assume that donors would not donate if they were opposed to the passage of a specific initiative, and that said PACs act based on the interests of their donors. Similarly, this analysis does not take into account the amount of money donated (for which multiple-issue PACs would be more problematic), but with the ideological support base of donors.

[20]Those with very low numbers of contributors represent cases either where financial disclosure was poor, where campaigns were funded solely by single individuals/groups, or where campaigns were financed other than through PACs (such as independent expenditures). These "reporting leakages" are a potential limitation of this analysis, but are broadly unavoidable given American campaign finance regulation.

2.4.2 Explanation 5: Active legislative aversion

Distributions of support. As a first check of the distribution of support for initiatives, I generate non-cumulative kernel density plots for each initiative and the two corresponding state legislative parties. For instance, Figure 2.7 demonstrates the relative densities of support along the ideological axis for initiatives held in Florida. In both Figure 2.7, and across the entire set of initiatives considered in this paper, the vast majority of initiatives exhibit support clustered around a distinct region of the ideological axis, with limited diffusion of ideological positions.[21]

I classify the position of initiatives dependent on relative position of the initiative's mean ideological support *vis-a-vis* the mean support for both parties. "Left" ("right") positions indicate those initiatives whose mean ideological support is more liberal (conservative) than both party means. Since both positions lie outside the interval between the two parties, "left" and "right" initiatives are also "extreme". Initiatives with mean ideological support between the parties are labelled "centrist". Table 2.3 reports the calculated mean ideological position of support for each ballot initiative, as well as a brief description of the bill's content.[22] The position column refers to whether the mean ideological support of the initiative is to the 'Left' of the Democratic Party, 'Right' of the Republican Party, or in the 'Centre' between the two party distributions.

[21] A full set of kernel density plots for each state are located in the Appendix.

[22] Descriptions adapted from information provided by the National Conference of State Legislatures' ballot measures database (National Conference of State Legislatures, 2018).

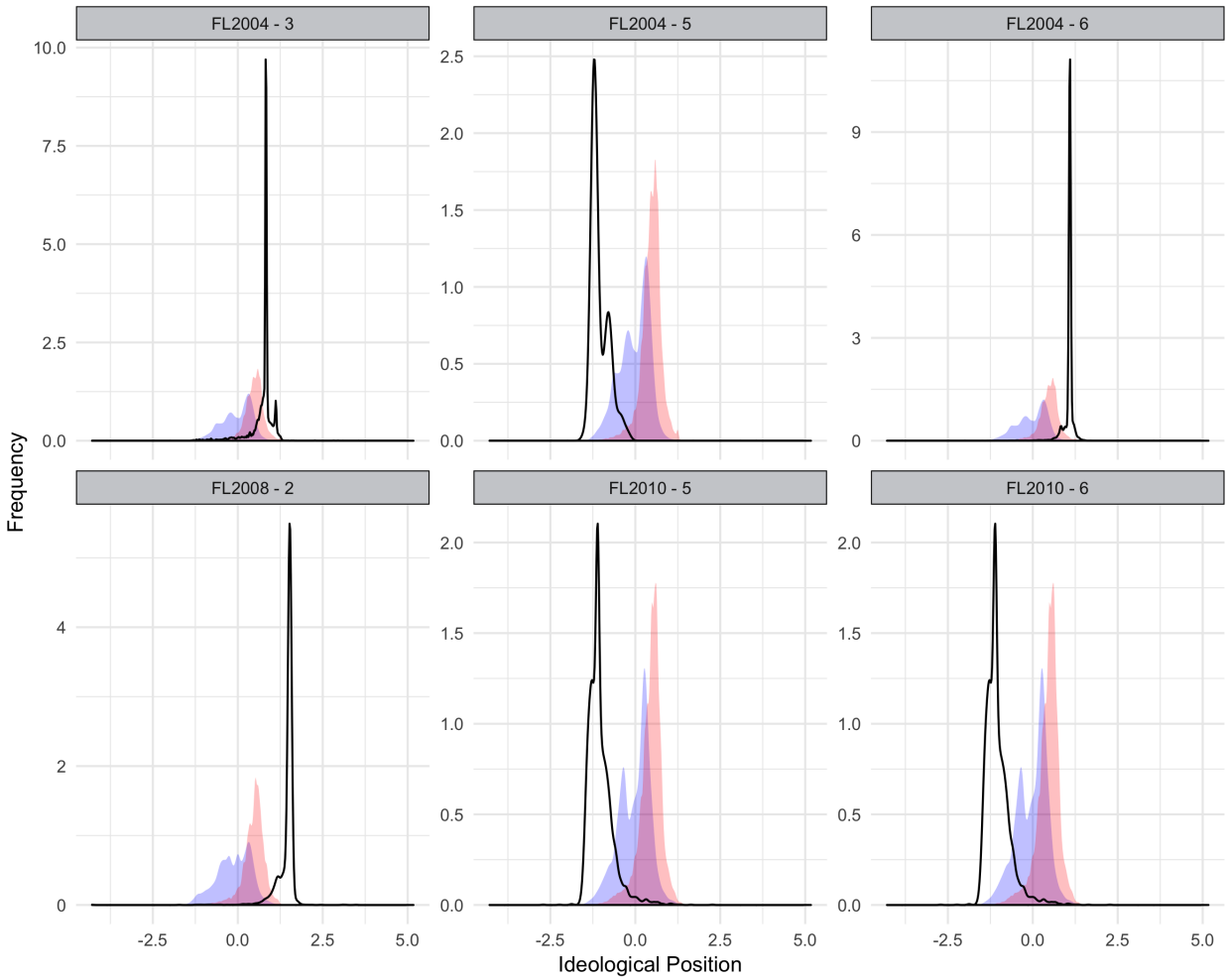


Figure 2.7: Kernel density plots of the ideological distribution of financial donors to initiatives held in Florida, compared against the respective support for both parties for the corresponding electoral cycle.

Table 2.3: Descriptions of successful initiatives and the calculated ideological positions

State	Year	Ballot No.	Description	Mean Ideology	Position
Arizona	2004	200	Implement voter ID law	1.82	Right
Arizona	2006	201	Public smoking ban	-0.66	Left
Arizona	2006	203	Funding for early childhood development and health programs	0.24	Centre
Arizona	2006	204	Prevent cruel and inhumane confinement of animals	-1.05	Left
Arizona	2006	207	Reimburse lost property value when imposing land usage restrictions	1.27	Right
Arizona	2010	203	Legalise medical marijuana	-0.44	Centre
Arkansas	2004	3	Same-sex marriage ban	1.22	Right
Arkansas	2008	1	Prevent fostering or adoption by non-married cohabiters	1.28	Right
California	2000	35	Use of private contractors in public works projects	0.22	Right
California	2000	39	Bond for school facilities	-0.41	Centre
California	2002	49	State grant for after school activities	0.37	Right
California	2002	50	Bonds to improve water quality	-0.43	Centre
California	2004	64	Limit enforcement of business competition laws	0.21	Centre
California	2004	63	Mental health services expansion and funding via millionaire's tax	-1.16	Left
California	2004	71	Stem cell research funding	-0.57	Centre
California	2006	83	Punishment, residence restrictions and monitoring of sexual offenders	0.65	Right
California	2006	84	Bond to improve water quality	-0.50	Centre
California	2008	8	Same-sex marriage ban	1.87	Right
California	2008	11	Redistricting to be decided by citizens commission	0.09	Centre
California	2008	2	Confined farm animal standards	-1.39	Left
California	2010	25	Change legislative budget passage vote to 50% threshold	-0.86	Left
California	2010	22	Prevent appropriation of funds from transportation or local government projects	-0.33	Centre
California	2010	20	Redistricting of congressional districts	0.30	Right
California	2010	26	Require two-thirds approval to increase state and local fees	0.23	Centre
California	2012	30	Temporary tax to fund education spending	-1.18	Left
California	2012	35	Penalties for human trafficking and offender registration	-1.16	Left
California	2012	36	Three strikes law	-1.13	Left
California	2012	39	Clean energy funding	-1.86	Left
Colorado	2002	27	Implement campaign finance limits	-1.36	Left
Colorado	2004	37	Renewable energy requirement	-1.32	Left
Colorado	2004	35	Increase tobacco tax	-0.88	Centre
Colorado	2006	43	Same-sex marriage ban	1.50	Right
Colorado	2008	54	Campaign finance regulations for government contractors	1.21	Right
Colorado	2012	64	Legalis use and regulation of marijuana	-0.91	Centre

Table 2.3: Descriptions of successful initiatives and the calculated ideological positions

State	Year	Ballot No.	Description	Mean Ideology	Position
Colorado	2012	65	Charge congressional delegation with supporting campaign finance limits	-1.40	Left
Florida	2004	3	Limit fees paid to attorneys in medical malpractice suits	0.74	Right
Florida	2004	6	High speed rail amendment repeal	1.06	Right
Florida	2004	5	Increase minimum wage	-1.05	Left
Florida	2008	2	Same-sex marriage ban	1.44	Right
Florida	2010	5	Legislative redistricting to be 'fair' and use geographical boundaries when possible	-1.03	Left
Florida	2010	6	Congressional redistricting as in Measure 5	-1.03	Left
Maine	2012	1	Same-sex marriage legalisation	-1.44	Left
Massachusetts	2008	3	Prohibit dog racing	-1.17	Left
Massachusetts	2008	2	Decriminalize marijuana	-0.92	Left
Massachusetts	2010	1	Salex tax on alcoholic beverages	-0.17	Centre
Massachusetts	2012	3	Legalise medical use of marijuana	-0.92	Left
Michigan	2004	2	Same-sex marriage ban	1.45	Right
Michigan	2006	2	Ban affirmative action	1.17	Right
Michigan	2008	2	Allow human embryo and human embryonic stem cell research	-0.65	Centre
Michigan	2008	1	Permite the use and cultivation of marijuana for medical conditions	-0.88	Left
Missouri	2004	3	Transportation funds	-0.07	Centre
Missouri	2006	2	Stem cell research	-0.28	Centre
Missouri	2006	B	Increase minimum wage	-0.79	Left
Missouri	2008	C	Create a renewable energy standard	-1.01	Left
Missouri	2010	B	Improve dog breeding welfare standards	-0.80	Left
Missouri	2010	3	Prevent new real estate taxes	1.69	Right
Montana	2004	149	Increase tobacco tax	-2.78	Left
Montana	2006	151	Increase minimum wage	-1.10	Left
Montana	2012	166	Prevent corporations from being considered 'human beings'	-1.36	Left
Nevada	2004	1	Fund education before any other budget item at state-level	0.30	Right
Nevada	2004	3	Limit fees paid to attorneys in medical malpractice suits	0.47	Right
Ohio	2004	1	Same-sex marriage ban	1.44	Right
Ohio	2006	2	Increase minimum wage	-0.89	Left
Ohio	2006	5	Ban smoking in public places	-0.35	Centre
Oregon	2002	25	Increase minimum wage	-1.02	Left
Oregon	2004	36	Same-sex marriage ban	1.24	Right
Oregon	2004	37	Reimburse lost property value when imposing land usage restrictions	0.78	Right
Oregon	2006	44	Expand state drug prescription program	-3.21	Left

Table 2.3: Descriptions of successful initiatives and the calculated ideological positions

State	Year	Ballot No.	Description	Mean Ideology	Position
Oregon	2012	79	Prevent new real estate taxes	0.88	Right
Washington	2004	297	Added regulations about (mixed-) radioactive waste	-1.41	Left
Washington	2004	872	Reform electoral system to use "top-two" primary	0.40	Right
Washington	2006	937	Energy resource use by utilities companies	-1.33	Left
Washington	2008	1000	Establish right to die	-1.13	Left
Washington	2010	1053	Require two-third vote to increase state taxes and fees	1.10	Right
Washington	2012	502	Decriminalize marijuana	-1.25	Left
Washington	2012	1240	Create a public charter school system	0.11	Centre
Washington	2012	1185	Two-thirds vote for tax increases	1.34	Right

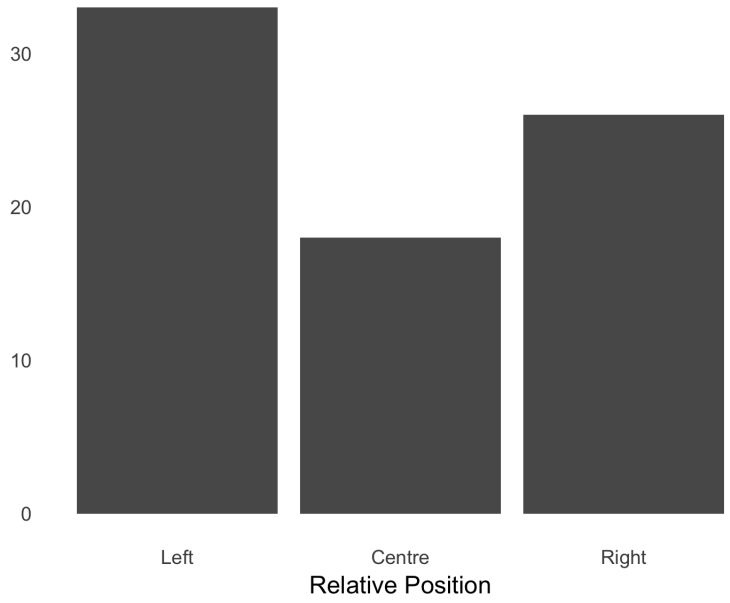


Figure 2.8: Histogram of the relative position of initiatives compared to the legislative parties.

Figure 2.8 summarises the distribution of the relative positions of initiatives. Initiatives succeed at all three relative positions, but a clear majority of initiatives are "extreme". Collectively, 77 percent of initiatives have a mean ideological score that is greater (in absolute terms) than the closest party mean. These "extreme" initiatives (in relative terms) are like those displayed in Figure 2.7 where the ideological support for each initiatives lies predominantly outside the respective distributions for both major parties. Indeed, this imbalance of extreme initiatives is found within each state, and for all but the 2000 electoral cycle where data is more sparse (see Table A.2 in the Appendix.) In Appendix Table A.3 I also show that this pattern occurs within social, economic, and governmental issue areas. For social and economic issues, most initiatives are further left than the Democratic Party but governmental issues (policies that determine how politics is conducted) tend to be right-leaning.

Initiatives labelled "extreme" are substantively further to the left or right than either party. On average, extreme initiatives are 1.2 standard deviations (in terms of the party distribution) away from the legislative support of the closest party. These measures are therefore much closer to the extremes of the ideological dimension than are the two respective

parties. This is a non-trivial finding: successful initiatives that command majority support in the electorate often have a support base that is substantively more extreme along the ideological dimension than that for either party.

To compare the distributions of donors' ideology between initiative and successful party candidates further, I run Mann-Whitney two-sample U-tests to determine whether these ideological values are drawn from the same population distribution (Lewis-Beck, Bryman and Futing Liao, 2004). This nonparametric test assesses the likelihood that a randomly-selected CFscore from one distribution is higher than a randomly selected CFscore from the other. The resultant p-value corresponds to the null hypothesis that the two distributions of ideological support are identical. Full results can be found in Appendix Table A.1.[23]

Overwhelmingly, these tests show that initiative and legislative party support are drawn from separate distributions. We can reject the null hypothesis that the ideological support for initiatives is distinct from both parties in 71 out of 77 cases. Overall, not only are the vast majority of initiatives statistically distinguishable from support for the two parties respectively, but the statistical confidence in these results is high too. Even after Bonferroni-adjustments, $p < 0.001$ for 144/154 tests. In general, therefore, support for initiatives are drawn from distinct bases of ideological support relative to those of the two legislative parties.

Robustness tests

Projected donors and CFscore confidence. The CFscore metric relies on the shared donation history between donors. Individual point estimates in turn rely on the number of donations that each separate donor makes across the pooled universe. As a consequence, the greater the number of distinct donations to different campaigns made by an individual, the greater our confidence in the ideological point estimate of that actor. Individuals who make a single distinct donation – that is, to one candidate only across the entire donor universe –

[23]Since the number of pairings between the two groups is large, the reported p-values are calculated using a normal approximation. Moreover, running 154 separate Mann-Whitney U-tests (77 initiatives compared against the two parties separately) increases the likelihood of Type-I errors through multiple comparisons. I therefore report corrected p-values using the Bonferroni adjustment (Sedgwick, 2014).

are excluded from the scaling, and only reintroduced through projection. These donors are assigned the ideological score of the candidate or group to which they donate.

This projection is potentially problematic. While a donation to a liberal Democrat might indicate that that donor is also a left liberal, projection assumes their ideologies are exactly the same when in reality there might be substantial divergence. Projected donors pose a robustness problem if they substantively impact the categorisation and distance measurement of support for legislative and initiative campaigns alike. Conventional confidence intervals of the individual ideological positions are not possible since the scaling algorithm used to determine the score relies on the *non-independence* of observations. To test the robustness of my findings, therefore, I rerun my analysis imposing increasingly stringent restrictions on the number of distinct donations required to be included within the analysis - a minimum of two, four, and eight distinct donations respectively (similar to tests in (Hill and Huber, 2017)).

Table 2.4 summarises the results of these robustness checks. As in the main analysis, I do not consider any initiative where the total number of donors is below 40. Therefore, there are fewer initiatives considered as the stringency increases. The average extremeness of measures is stable as the stringency of the exclusion criteria is increased - increasingly slightly from 77 to 82 percent over the four specifications. There is a slight decrease in the average distance (in standard deviations) as the stringency of inclusion increases, but the average extremity remains around 1 standard deviation away from the closest party. Overall, this test suggests that projected donors are not unevenly distributed between centrist and extreme positions, which would result in larger shifts in the number of extreme initiatives relative to centrist ones. If anything, including projected donors is a more conservative estimate of this ratio. Full initiative-level results are available in Table A.4 in the Appendix.

Bootstrapped means. The distribution-test results are also robust to a direct comparison of means, as opposed to Mann-Whitney U-tests. In the Appendix I compute a difference-

Minimum donations	Distance (σ) Min.	Max.	Mean	No. of extreme initiatives	Extreme %
$n \geq 1$	0.019	5.355	1.269	59	0.766
$n \geq 2$	0.006	4.525	1.058	59	0.787
$n \geq 4$	0.075	2.32	0.975	54	0.783
$n \geq 8$	0	2.524	0.96	53	0.815

Table 2.4: Robustness test: comparison of SD distances for extreme measures.

in-means estimate between the initiative and legislative positions using bootstrapping. The results are substantively unchanged. Full details of these calculations and discussion of the results are provided in the Appendix alongside the full table of results (Table A.5).

2.4.3 Explanations 3 and 4: conservative bias and partisan legislatures

The donations data can also be used to inspect the conservative bias and partisan legislatures explanations further. If there is conservative bias to legislators' beliefs, we would expect to observe initiative support clustered to the left of the majority party's position. If the partisan legislatures explanation holds, then support for successful initiatives should cluster in the opposite ideological direction to the party that controls the legislative agenda. To assess these hypotheses, I merge the initiative-level data with information on state-level party control taken from Caughey, Warshaw and Xu (2016).

Table 2.5 reports the relative position of successful initiatives by party control of the state legislature in the same year (the legislative session prior to the November election). A chi-squared test of the two categorical variables fails to reject the null ($\chi^2 = 1.91$, df $= 4$, p-value $= 0.75$).[24] This association is similarly weak if you aggregate initiative positions into a binary measure of extreme or centrist ($\chi^2 = 1.45$, df $= 2$, p-value $= 0.48$). Null results are also found for gubernatorial control and state divided government.

Initiative proponents appear able to build majority coalitions at the relative fringes of

[24] As a directional test of association, I also run a Goodman-Kruskal tau test to check for the presence of an asymmetric relationship between the two variables (Pearson, 2016). In both directions, there is no substantive correlation between state legislative control and the relative position of initiatives ($\tau_{xy} = 0.012, \tau_{yx} = 0.011$).

Initiative	**Legislative Control**		
Position	Democrat	Republican	Split
Left	13	14	6
Centre	10	6	2
Right	11	9	6

Table 2.5: Legislative control and initiative position.

the ideological spectrum regardless of party control within the state legislature. But since both the conservative bias and partisan legislatures arguments centre around ideological divergence relative to the majority party's position, I assess these mechanisms further by comparing the relative position of support for initiatives to the *majority* party's position.

With respect to conservative bias, successful initiatives do not garner more liberal support than the majority party. Ignoring cases of split legislative control, 33 of 63 cases exhibit more liberal-leaning supporters ($p > 0.8$). Contrary to expectations, the likelihood of being liberal leaning is essentially indistinguishable from chance.

On the other hand, 41 cases exhibit moderation towards the minority party (as expected by the partisan legislatures explanation). That is, in 65 percent of cases ($p < 0.05$), successful initiatives' mean ideological support is to the right of the Democratic party in Democrat-controlled legislatures, and *vice versa*. This finding is supportive of the partisan legislatures explanation. The majority party's control of agenda-setting procedures precludes convergence on median electoral preferences that are moderated towards the other side of the ideological spectrum. This moderation is not necessarily small; moderated positions include those Right extreme measures (initiatives with ideological support more conservative than the Republican party) in Democratic-controlled legislatures.

Still 35 percent of initiatives do not exhibit support away from the majority party and towards the minority party – and this is clear from Table 2.5, where there are a considerable number of Right extreme initiatives in Republican legislatures and Left extreme measure in Democratically-controlled legislatures. This explanation therefore finds some supportive evidence, but it is far from conclusive.

Summary. The financial support bases for successful initiatives and candidates are, in almost all cases, ideologically distinct. Moreover, robust to specification changes, a large majority of successful initiatives have, on average, more ideologically extreme donors than do the respective state legislators of each party. The observed differences in mean ideological position are substantial and suggestive of differences in the underlying support coalitions for issues that have to be passed by initiative. This feature could be interpreted in various ways. Broadly, it suggests that these popular policies have not taken root in the mainstream policy community. This could be because they are untested, possibly unwise (at least to conventional policymakers), or because they are opposed by powerful financial interests. Moreover, there is some suggestive evidence that successful initiatives have support coalitions that mediate towards the direction of the minority.

How do we make sense of these findings in terms of the broader theory that includes those unproposed initiatives? In one sense, comparisons against party distributions of support are an inherent check of whether successful initiatives lay outside the political mainstream (although the party distributions act to define this extremity too). The outstanding concern is whether other issues that legislators are willing to pass could also be as outlying – thus the extremity itself would not be an intrinsic reason for legislators' aversion. This unobserved portion of the joint distribution is critical for valid inference. The sheer magnitude of the differences in support suggests the unproposed initiatives would also have to be very outlying for the hypothesis to be falsified. And if that were the case, who are these party donors who differ so much from the supporters for *all* popular policies? These questions cannot be fully resolved here, and merit much more research. In the final empirical section, however, I present the results from a bespoke small-n survey to assemble further data in support of this active aversion theory.

2.5 Empirical evidence: Surveying state legislators directly

We cannot observe the motivations behind legislators' apparent intransigence directly using either the voting data or contributions information. We can, however, solicit legislators' attitudes towards the policies and processes that structure and define both conventional and ballot initiative policymaking. To this end, I conducted a small survey on state legislators to test their attitudes on the initiative process and its role vis-a-vis the legislature. All state legislators in Arizona, California, Massachusetts, Oregon, and Washington were contacted at their official legislative email address and asked to complete a short structured survey. Given the likely policy knowledge of staffers close to state legislators (Hertel-Fernandez, Mildenberger and Stokes, 2019), I indicated that senior legislative staffers could respond on behalf of their representative. In total, 27 legislators responded to these invitations.[25] Limits to the generalisability of this evidence go without saying. I use these structured responses as a qualitative complement to the analysis in the rest of this paper.

Legislators were asked to indicate their level of agreement with respect to five statements about the initiative process and when it is best used. Over 60 percent of respondents agreed that 'there are times when legislators have to be more cautious than voters would be', and a majority indicated that the initiative process itself 'takes the pressure off legislators to act on certain issues.' High levels of support for both statements indicate that legislators acknowledge both policy incongruence, and that the initiative process can act as a pressure valve in these instances. No respondent, however, agreed that the initiative process was best used when the legislative session was busy.

Over half of all respondents also indicated that the initiative process was *not* a good way to make 'important political decisions'. And only a minority of respondents (38.5 percent) agreed that the initiative process was best used for 'politically sensitive' issues. The surveyed legislators do not simply defer to the initiative process because they perceive normative

[25]The response rate was 5.4 percent, which is in line with similar studies at the national and state level (Hertel-Fernandez, Mildenberger and Stokes, 2019; Stokes, Mildenberger and Hertel-Fernandez, N.d.), but the size of the population smaller is in this case.

benefits to direct democratic action. Instead, legislators appear to prefer, hypothetically, to handle all policy issues themselves. Their ability to do so, however, is mediated by other factors.[26]

To understand these factors better, I asked legislators to rate the importance of seven different factors that "deter the legislature from passing legislation" (on a scale from 0 = "not at all important", to 7 = "very important"). This question was asked abstractly without reference to any particular policy or legislation, and to avoid possible social desirability bias the question was framed in terms of the legislature as a whole.

Factor	Mean Importance	σ	N
Political sensitivity of the policy	5.72	1.17	25
Pressure from interest groups	5.48	1.61	25
Policy stalemate within the legislature	5.17	1.75	23
The relative importance of the policy	5.00	2.20	25
Economic costs of the policy	4.96	1.81	25
Unstable public opinion on the policy	4.71	2.12	24
How well the policy is written	4.29	1.88	24

Table 2.6: Factors contributing to legislative inaction.

Table 2.6 ranks these factors by their average importance to legislators. Each factor was ranked above the scale's midpoint, suggestive of the complexity behind why legislatures act on some issues but not others. That said, political sensitivity was ranked the most important reason for inaction (with the smallest variance in ranking across respondents). Legislators appear risk-averse to some issue areas, dampening legislative enthusiasm towards pass new policy even if its popular. While this factor was consistently rated as important for deterring legislative action objectively, legislators did not believe that the initiative process was better able to handle these types of sensitive issues.

[26]When asked if there were any particular issue areas that are better resolved by ballot initiative, half of all respondents answered in the affirmative. These respondents then had the opportunity to select from a list of 26 areas derived from the National Conference of State Legislature's policy categorisations. 15 of the areas were chosen at least once suggesting inaction is not clustered in a small number of policy areas. The most frequently chosen topics were "lobbying and campaign finance" and "ethics" (six times); "drug policy" and "alcohol and tobacco policy" (five times); "criminal justice" and "tax and revenue" policy (four times). As a note of caution, there may be some confirmation bias to this question given the recent initiative history of the states in question.

Interest group pressure was the second most important reason for inaction, on average. Given the sensitivity of this subject, it is notable that legislators were willing to rate the effect of interest group influence so highly.[27] These responses accord with the findings of the previous section: disjunctures in financial support bases appear to influence the ability of legislators to act on certain, popular policy proposals. Legislators feel the pressure of interested groups who may not necessarily be aligned with popular opinion, providing further evidence that sectional interests may diminish policy responsiveness.

Less influential (in relative terms) were the quality of the written policy and unstable public opinion on a given issue. At least relative to the other factors, legislators did not see inactivity as a result of uncertainty about the position of the median voter. Given the ideal legislative responses computed earlier in the paper, it is not surprising that legislators do not rank this factor as highly as the others.[28]

These findings complement the analysis of the previous section. Issues are less likely to be acted on if they alienate key interest groups or are politically sensitive, features we would expect of issues not incorporated into mainstream political networks. Legislators, while sceptical of the initiative process, recognise moreover that the initiative process alleviates the pressure to act on issues. Again, the evidence suggests that the initiative process allows legislators to defer judgement on policy that may be novel, untested or have negative policy consequences despite its electoral popularity – shifting responsibility for its passage directly

[27] As Stokes, Mildenberger and Hertel-Fernandez (N.d.) suggest, the importance of this factor may be more important than this direct question suggests. They compare the response rate of overt questions about interest group influence to inferred ratings via a list experiment. The list experiment, in which respondents do not need to actively indicate the importance of interest groups, leads to higher levels of perceived importance of these actors' role. In the survey conducted for this paper, while it is clear legislators are not entirely averse to admitting the role of interest groups, they may nevertheless downplay their importance though this is of course speculative.

[28] Respondents were also asked about specific initiative policies passed in the last two electoral cycles (2016 and 2018) in their state. Respondents were asked to rank their relative importance of the same factors as in Table 2.6. In the Appendix, I group these issues into six policy areas - drug, election, environmental, gun, tax, and wage policies – and compare rankings across issue areas. Notably when respondents were asked about specific issues, interest group pressure is the most consistently important factor. Respondents rate political sensitivity as the most important factor for drugs and gun policy, but this factor is less important for elections and wage policy respectively. Legislative time and uncertain public opinion are consistently ranked the least important across issue areas. Table A.6 in the Appendix gives a more detailed breakdown.

to voters.

2.6 Discussion

In this paper, I explore five different reasons why initiatives might be expected to succeed, given our intuitions about legislative convergence in representative democratic systems. These explanations cover a range of different explanatory factors that would distort or limit the responsiveness of legislators. I leverage new data sources to challenge existing explanations for initiative success at both the individual and institutional level.

In doing so, I develop and incorporate hypothesis testing strategies that are sensitive to the systematic incompleteness of data, and illustrate its implications for valid inference beyond the scope of this paper. Researchers should not shy away from substantively important topics where selection issues are prevalent, but instead construct empirical tests that assess the likelihood of hypotheses taking into account plausible distributions for the missing cases. This paper demonstrates how such strategies can be applied in a context where one portion of the joint distribution of data is systematically unobserved.

In substantive terms, this paper presents evidence to suggest that the initiative process is used predominantly when issues lie outside the political mainstream. While there are substantial correlations between initiative and presidential voting, it does not appear that initiative issue dimensions are fully captured by this partisan dimension of conflict. Moreover, supporters of initiatives are, typically, much more extreme than supporters of the two major parties. This evidence suggests that legislators deliberately avoid passing popular legislation on some policy areas, particularly when that policy lies outside mainstream policy networks. Qualitative survey evidence suggests that both the financial interests and political sensitivity of the issues condition legislators' hesitance to act.

This paper also demonstrates important null results. Popular policy does not appear to end up on the ballot due to electoral disproportionality. In a counterfactual sense, legislators

should have supported legislation that went on to be passed at the ballot. The significance of this result is noteworthy. Not only do legislators appear to, at times, ignore majority demands of the statewide electorate, but more particularly representatives appear to ignore the majority demands of the very voters upon whom their legislative careers depend.

Future work should seek to verify further the underlying mechanisms that drive legislative aversion. What types of popular issues, in particular, are legislators keen to avoid? And can selection problems be overcome by comparing successful legislative policymaking efforts to initiatives?

More generally, this paper helps to develop our understanding of representation in modern democratic systems. It is clear from the evidence presented in this paper that legislators do not always act on popular policies. Our focus on political representation to date has focussed on identifying when, and for how long, incongruences exist between voters and their elected representatives (Caughey, Warshaw and Xu, 2017; Simonovits, Guess and Nagler, 2019). New theories of representation have highlighted impediments, both strategic and psychological, that limit this convergence (Broockman and Skovron, 2018; Hertel-Fernandez, Mildenberger and Stokes, 2019). The findings of this paper suggest we should also look at normative components of representation too – what legislators think they *should* do, and how the composition of the wider policy community conditions their willingness to act on majority policy demands.

Ballot initiatives serve to resolve incongruence, and this paper provides novel evidence about the conditions under which this occurs. Initiatives succeed when legislators are averse to owning certain issues, and this appears to be driven both by party dynamics and the location of policy support vis-a-vis the political mainstream. The implications of this research are important, particularly when disillusionment with politics today is often attributed to failures in democratic representation and responsiveness. At a time when politics has become distinctly 'populist', understanding the relationships between these fundamental types of democratic policymaking is crucial to the public's trust in democracy.

References

Adams, Brian E. 2012. "Citizens, Interest Groups, and Local Ballot Initiatives." *Politics & Policy* 40(1):43–68.

Aldrich, John H. and James S. Coleman Battista. 2002. "Conditional Party Government in the States." *American Journal of Political Science* 46(1):164–172.

Andris, Clio, David Lee, Marcus J. Hamilton, Mauro Martino, Christian E. Gunning and John Armistead Selden. 2015. "The Rise of Partisanship and Super-Cooperators in the U.S. House of Representatives." *PLOS One* 10(4):e0123507.

Ansolabehere, Stephen, De Figueiredo, John M, Snyder Jr and James M. 2003. "Why Is There so Little Money in U.S. Politics?" *Journal of Economic Perspectives* 17(1):105–130.

Barber, Michael J. 2016*a*. "Ideological Donors, Contribution Limits, and the Polarization of American Legislatures." *The Journal of Politics* 78(1):296–310.

Barber, Michael J. 2016*b*. "Representing the Preferences of Donors, Partisans, and Voters in the US Senate." *Public Opinion Quarterly* 80(S1):225–249.

Barber, Michael J., Brandice Canes-Wrone and Sharece Thrower. 2017. "Ideologically Sophisticated Donors: Which Candidates Do Individual Contributors Finance?" *American Journal of Political Science* 61(2):271–288.

Barber, Michael J. and Nolan McCarty. 2015. Causes and Consequences of Polarization. In *Political Negotiation: A Handbook*, ed. Jane Mansbridge and Cathie Jo Martin. Brookings Institution Press pp. 37–90.

Boehmke, Frederick J., Tracy L. Osborn and Emily U. Schilling. 2015. "Pivotal Politics and Initiative Use in the American States." *Political Research Quarterly* 68(4):665–677.

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58(2):367–386.

Bonica, Adam. 2016. "Database on Ideology, Money in Politics, and Elections: Public Version 2.0." Stanford University Libraries.

Broockman, David E. and Christopher Skovron. 2018. "Bias in Perceptions of Public Opinion among Political Elites." *American Political Science Review* 112(3):542–563.

Caughey, Devin and Christopher Warshaw. 2018. "Policy Preferences and Policy Change: Dynamic Responsiveness in the American States, 1936–2014." *American Political Science Review* 112(2):249–266.

Caughey, Devin, Christopher Warshaw and Yiqing Xu. 2016. "Replication Data for: Incremental Democracy: The Policy Effects of Partisan Control of State Government.". **URL:** *https://doi.org/10.7910/DVN/14RSGX*

Caughey, Devin, Christopher Warshaw and Yiqing Xu. 2017. "Incremental Democracy: The Policy Effects of Partisan Control of State Government." *The Journal of Politics* 79(4):1342–1358.

Coan, Travis G., Jennifer L. Merolla, Laura B. Stephenson and Elizabeth J. Zechmeister. 2008. "It's Not Easy Being Green: Minor Party Labels as Heuristic Aids." *Political Psychology* 29(3):389–405.

Cox, Gary W. and Mathew D. McCubbins. 2005. *Setting the Agenda: Responsible Party Government in the U.S. House of Representatives*. Cambridge: Cambridge University Press.

Downs, Anthony. 1957. *An Economic Theory of Democracy*. New York: Harper & Row.

Erikson, Robert S., Gerald C. Wright and John P. McIver. 1993. *Statehouse Democracy: Public Opinion and Policy in the American States*. Cambridge University Press.

Figueiredo, De, John M, Chang Ho Ji and Thad Kousser. 2011. "Financing Direct Democracy: Revisiting the Research on Campaign Spending and Citizen Initiatives." *The Journal of Law, Economics, and Organization* 27(3):485–514.

Francia, Peter L., John C. Green, Paul S. Herrnson, Lynda W. Powell and Clyde Wilcox. 2005. "Limousine Liberals and Corporate Conservatives: The Financial Constituencies of the Democratic and Republican Parties*." *Social Science Quarterly* 86(4):761–778.

Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics." *Political Analysis* 2:131–150.

Gerber, Elisabeth R. 1996. "Legislative Response to the Threat of Popular Initiatives." *American Journal of Political Science* 40(1):99–128.

Gerber, Elisabeth R. 1999. *The Populist Paradox: Interest Group Influence and the Promise of Direct Legislation*. Princeton University Press.

Gerber, Elisabeth R. and Jeffrey B. Lewis. 2004. "Beyond the Median: Voter Preferences, District Heterogeneity, and Political Representation." *Journal of Political Economy* 112(6):1364–1383.

Gluck, Abbe R., Anne Joseph O'Connell and Rosa Po. 2015. "Unorthodox Lawmaking, Unorthodox Rulemaking." *Columbia Law Review* 115(7):1789–1865.

Hertel-Fernandez, Alexander, Matto Mildenberger and Leah C. Stokes. 2019. "Legislative Staff and Representation in Congress." *American Political Science Review* 113(1):1–18.

Hill, Seth J. and Gregory A. Huber. 2017. "Representativeness and Motivations of the Contemporary Donorate: Results from Merged Survey and Administrative Records." *Political Behavior* 39(1):3–29.

Hug, Simon. 2004. "Occurrence and Policy Consequences of Referendums: A Theoretical Model and Empirical Evidence." *Journal of Theoretical Politics* 16(3):321–356.

Initiative & Referendum Institute, University of Southern California. 2018. "State I&R." http://www.iandrinstitute.org/states.cfm.

Jourdain, Charlotte, Simon Hug and Frédéric Varone. 2017. "Lobbying Across Venues: An Issue-Tracing Approach." *State Politics & Policy Quarterly* 17(2):127–153.

King, Gary and Langche Zeng. 2001. "Explaining Rare Events in International Relations." *International Organization* 55(3):693–715.

Klüver, Heike and Mark Pickup. 2019. "Are They Listening? Public Opinion, Interest Groups and Government Responsiveness." *West European Politics* 42(1):91–112.

Lax, Jeffrey R. and Justin H. Phillips. 2012. "The Democratic Deficit in the States." *American Journal of Political Science* 56(1):148–166.

Leeman, Lucas and Fabio Wasserfallen. 2016. "The Democratic Effect of Direct Democracy." *American Political Science Review* 110(4):750–762.

Lewis-Beck, Michael, Alan Bryman and Tim Futing Liao, eds. 2004. *The SAGE Encyclopedia of Social Science Research Methods*. SAGE Publications, Inc.

Mann, Thomas E. and Norman J. Ornstein. 2012. *It's Even Worse than It Looks: How the American Constitutional System Collided with the New Politics of Extremism*. New York: Basic Books.

Manza, Jeff and Fay Lomax Cook. 2002. "A Democratic Polity?: Three Views of Policy Responsiveness to Public Opinion in the United States." *American Politics Research* 30(6):630–667.

Matsusaka, John G. 2005. "Direct Democracy Works." *The Journal of Economic Perspectives* 19(2):185–206.

Matsusaka, John G. and Nolan M. McCarty. 2001. "Political Resource Allocation: Benefits and Costs of Voter Initiatives." *Journal of Law, Economics, & Organization* 17(2):413–448.

McCarty, Nolan, Keith T. Poole and Howard Rosenthal. 2016. *Polarized America: The Dance of Ideology and Unequal Riches*. Walras-Pareto Lectures second edition. ed. Cambridge, Massachusetts: The MIT Press.

Miller, Patrick R. and Pamela Johnston Conover. 2015. "Red and Blue States of Mind: Partisan Hostility and Voting in the United States." *Political Research Quarterly* 68(2):225–239.

Mintz, Alex. 2009. "Methodological Issues in Studying Suicide Terrorism." *Political Psychology* 30(3):365–371.

National Conference of State Legislatures. 2018. "Ballot Measure Database." http://www.ncsl.org/Default.aspx?TabId=16580&language=en-US.

National Institute on Money in State Politics. 2017. "Ballot Measures." https://www.followthemoney.org/tools/ballot-measures/.

Page, Benjamin I. and Martin Gilens. 2017. *Democracy in America?: What Has Gone Wrong and Waht We Can Do about It*. The University of Chicago Press.

Pearson, Ron. 2016. "Association Analysis for Categorical Variables." The Comprehensive R Archive Network.

Przeworski, Adam. 2015. Economic Inequality, Political Inequality, and Redistribution. Working Paper Department of Politics, New York University.

Qvortrup, Matt. 2014. *Referendums around the World: The Continued Growth of Direct Democracy*. Basingstoke: Palgrave Macmillan.

Richards, Timothy, William J. Allender and Di Fang. 2013. "Media Advertising and Ballot Initiatives: The Case of Animal Welfare Regulation." *Contemporary Economic Policy* 31(1):145–162.

Schaffner, Brian F. and Matthew J. Streb. 2002. "The Partisan Heuristic in Low-Information Elections." *Public Opinion Quarterly* 66(4):559–581.

Schumpeter, Joseph A. 1976. *Capitalism, Socialism, and Democracy*. 5th ed. ed. London: George Allen & Unwin.

Sedgwick, Philip. 2014. "Multiple Hypothesis Testing and Bonferroni's Correction." *BMJ* 349:g6284.

Shor, Boris, Christopher Berry and Nolan McCarty. 2010. "A Bridge to Somewhere: Mapping State and Congressional Ideology on a Cross-Institutional Common Space." *Legislative Studies Quarterly* 35(3):417–448.

Simonovits, Gabor, Andrew M. Guess and Jonathan Nagler. 2019. "Responsiveness without Representation: Evidence from Minimum Wage Laws in U.S. States." *American Journal of Political Science* 63(2):401–410.

Smith, Daniel A. and Caroline Tolbert. 2007. "The Instrumental and Educative Effects of Ballot Measures: Research on Direct Democracy in the American States." *State Politics & Policy Quarterly* 7(4):416–445.

Stokes, Leah, Matto Mildenberger and Alexander Hertel-Fernandez. N.d. Interest Groups and Representation among US State Legislatures.

Stratmann, Thomas. 2010. Campaign Spending and Ballot Measures. In *Financing Referendum Campaigns*, ed. Karin Gilland Lutz and Simon Hug. London: Palgrave Macmillan UK.

Thomsen, Danielle M. 2014. "Ideological Moderates Won't Run: How Party Fit Matters for Partisan Polarization in Congress." *The Journal of Politics* 76(3):786–797.

3 When should we cluster experimental standard errors?

When researchers suspect that error terms are correlated by group in observational research the standard correction is to cluster the standard errors. But what about in experimental contexts where treatment is randomised? Despite their ubiquity in analyses with group-constant variables, the rationale for using clustered standard errors in experimental contexts remains underdeveloped. In this paper I present an intuitive and applied explanation of when clustering is appropriate, building on recent contributions in the statistics and econometrics literatures. I demonstrate why randomisation does not lead to identical variance estimates across estimation strategies, and conduct a review of experimental studies published between 2017 and 2019 to show that these differences can be considerable. Finally, I provide practical guidance for when and why to cluster standard errors for common experimental designs.

Standard errors help indicate the probable bounds of a point-estimate under the assumption of repeat sampling. We use these bounds to limit the likelihood of committing a Type I error (wrongly rejecting a true null hypothesis), and thus to assess our confidence over a given statistical finding. Standard errors are integral to our practice as experimentalists. Without them point-estimates of the average treatment effect (ATE) from finite samples are meaningless. We would not know the precision of the ATE and, crucially, whether it is distinguishable from a null effect.

Despite their foundational importance to statistical analysis, guidance for applied researchers on *which* estimator of uncertainty to use is still developing. Conventional estimates of variance are known to have serious limitations – for instance when observations

in a dataset are clustered (Moulton, 1986; Liang and Zeger, 1986) or when residuals are heteroskedastic (Huber, 1967; Eicker, 1967; White, 1980). But we know less about when to justify the inclusion of modified variance estimators in our analysis, particularly outside of observational research contexts (Blattman, 2015; Abadie et al., 2020). Too often researchers seem to default to whichever variance estimator provides the most conservative (widest) uncertainty bound. The problem with this approach is that researchers often neglect that variance is a quantity to be estimated. Choosing the wrong estimator leads to incorrect inferences.

Neglecting the estimation of uncertainty is particularly acute in experimental research. Advances in the design of both in-the-field and online experimental apparatus have enabled scholars to estimate causal point-estimates across a wide variety of topics. The experimental method allows researchers to assume away the problem of selection bias when treatment assignment is randomised. But what about bias in the standard error? Does randomisation allow us to assume away potential biases of the corresponding variance estimates too?

There are numerous contexts in which it appears that experimental researchers should use some non-standard variance estimation strategy. In particular, observations within experimental research often involve multiple observations within groups. When individuals take part in repeat-observation tasks, randomised control trials (RCT) measure multiple observations within a "treated" geographical area, or researchers leverage some natural randomisation, there is likely to be a group-structure to statistical models' error terms. Randomisation of treatment in these protocols ensures regression coefficients are unbiased. But despite the temptation to think that this also solves the problem of clustered data, randomisation of treatment does not guarantee that errors terms will be uncorrelated within groups (Abadie et al., 2017).

Developed in the 1980s (Moulton, 1986; Liang and Zeger, 1986), cluster-robust estimates of variance continue to be refined and compared in the statistical and applied literatures (Esarey and Menger, 2019; Jackson, 2019; D'Adamo, 2019). It is now commonly accepted

that failing to account for clustering can lead to downwardly biased standard errors increasing the likelihood of a Type I error (Wooldridge, 2003).

The incorrect application of clustering estimators, however, can itself result in biased standard errors. While the conventional expectation is that cluster-robust estimation inflates variance compared to conventional estimation (Wooldridge, 2003; Cameron and Miller, 2015), studies show that clustering can downwardly bias variance estimates too (Esarey and Menger, 2019; Jackson, 2019). Even when clustered standard errors are larger than their conventional counterparts, they may nevertheless substantially underestimate the true variance (Imbens and Kolesar, N.d.).

Fewer works, however, have examined variance estimation issues from the perspective of randomised experiments specifically (Green and Vavreck, 2008; Schochet, 2013; Cameron and Miller, 2015; Abadie et al., 2017). In particular, there is little specific guidance for experimental researchers on whether to use modified variance estimators when the data generating process includes some group-level, clustered component. Those that have considered this issue have offered disparate and often conflicting advice. While Cameron and Miller (2015) argue that clustering is inconsequential in experimental contexts due to randomisation, Abadie et al. (2017) argue that the decision to cluster is a "design-based" decision dependent on both how observations are sampled and how treatment is assigned. When should experimental researchers, therefore, cluster standard errors?

In this paper, I distil the latest research from the statistics and econometrics literatures on variance estimation (Cameron and Miller, 2015; Abadie et al., 2014, 2017, 2020) to provide an intuitive and practical guide to when and why experimental research designs should use clustered standard errors. In particular I make three contributions.

First, I provide a simplified argument about when to use cluster-robust variance estimation in experimental contexts. Building on the design-based approach to clustering developed by Abadie et al. (2017), I show how both sampling frames and treatment assignment matter for correct variance estimation. In short, researchers should cluster experimental errors when

either sampling is done at the group-level (and one wants to infer to groups beyond the data) or treatment randomisation is intentionally correlated within groups.

Second, I demonstrate the implications of this argument using Monte Carlo simulations and a comprehensive literature review of experimental studies published in the *Journal of Experimental Political Science*. I show that even in completely randomised experimental designs there are likely to be considerable differences in the magnitude of conventional and cluster-robust standard errors. Correcting variance estimations in three published articles, moreover, I demonstrate that in practice these errors can be more than twice their original reported size.

Third, I emphasise how making valid statistical inferences in experimental research has to focus on both $\hat{\beta}$ and $var(\hat{\beta})$. One tempting approach to variance estimation is simply to report whatever variance estimator produces the largest standard error. We do so to hedge our bets against committing a Type I error. But this "largest error" approach makes the implicit judgement that a Type II error (failing to reject a false null hypothesis) is the lesser statistical evil compared to its Type I counterpart. This paper resists this claim on two fronts. From a statistical perspective, without considering the underlying data generating process, cluster-robust estimates may not be a conservative estimate of the true variance.[1] From a hypothesis testing perspective, correctly estimating variance (irrespective of its comparative size compared to other estimators) is an intrinsic part of making unbiased, scientific claims. A Type II error should concern researchers just as much as a Type I error. Over-estimating variance can lead to "file-drawer" problems – where true null results go unreported, biasing our research findings – if incorrect variance estimation leads to overly cautious estimates of standard errors.

The remainder of this paper proceeds as follows. Section 3.1 provides an account of why

[1]Note for instance that Cameron and Miller (2015) argue that in contexts where there is potential clustering one should compare conventional and cluster-robust variance estimators and opt for the cluster-robust estimate if there is an 'appreciable difference' (p.17). But this is only the correct strategy if we know that errors are in fact clustered at some group level and, moreover, that the separate assumptions for valid clustered variance estimation hold for the given data.

uncertainty matters, and the different forms it can take, in experimental research. Section 3.2 motivates the design-based logic of Abadie et al. (2017) to provide a simple, principled and easily-implementable variance estimation strategy for experimentalists. Sections 3.3 and 3.4 then demonstrate the importance of correctly estimating variance using Monte Carlo simulations and replication data respectively. Finally, Section 3.5 provides concrete guidance for two common types of experiment – randomised controlled-trials and conjoint experiments.

3.1 What is uncertainty in an experimental context?

Suppose an experimentalist wants to estimate a simple parameter like the average treatment effect (ATE; τ_{ATE}). To do so, they draw a series of observations – from a convenience sample like a student subject pool, or from a representative sample of a national population – and randomly assign individual units to treatment and control. Taking the mean difference in outcome between these two groups of observations, the researcher recovers an *estimate* of the ATE ($\hat{\tau}_{\text{ATE}}$).

Estimation implies uncertainty. It is an approximation of the true value. What the "true value" corresponds to, and therefore the extent of uncertainty over an estimate, will depend on how the researcher wishes to generalise their findings. From a very narrow perspective, the researcher may claim their is no uncertainty over their estimate. Given the specific set units in the sample (S), their assignment to treatment and control (Z), and the experimental context (E), we can calculate the average effect on the outcome exactly. In other words, conditional on the sample, randomisation, and experimental conditions, the recovered ATE estimate is the true value conditional on those same features:

$$E[Y(1) - Y(0)|S, Z, E] = \hat{\tau}_{\text{ATE}} = \tau_{\text{ATE}|S,Z,E} \tag{3.1}$$

Researchers, however, typically do not want to simply estimate this conditional effect and instead wish to generalise their result. In other words, researchers want to drop at least

one of S, Z, and E from their claims; doing so is equivalent to generalising over a wider population, different treatment assignment schedules, and or contexts.[2]

$$E[Y(1) - Y(0)|S, Z, E] = \hat{\tau}_{\text{ATE}} \approx \tau_{\text{ATE}} \tag{3.2}$$

Generalisation involves uncertainty because we try to infer the parameter value for data which the researcher does not observe. Note in Equation 3.2 that the same estimate of the ATE is only an approximation of the unconditional ATE. Typically we are concerned with uncertainty that arises due to finite sampling issues. Out of the universe of cases (the population), researchers are only about to collect data on a small proportion of these (the sample). The uncertainty of a parameter estimate corresponds to the fact that we wish to make a generalisable claim about the population from the data in our sample. In other words, we want to make a claim that includes units $s' \notin S$.[3] Since in practice the researcher typically only samples from the population once, uncertainty over the estimate quantifies the probable bounds of the parameter given those units not in the sample.

This "sampling uncertainty" is the basis of variance and standard error calculation in most statistical research. However, it is just one particular form of uncertainty as a consequence of finite data. Various other factors may not be fully observed as a result of the one-shot nature of data collection, including how the estimate of the average treatment effect might differ if the treatment assignment schedule (Z) were different.

From a vast number of equally-probable assignment schedules we only observe one particular draw (Abadie et al., 2014). The notion of "causal" uncertainty, where there are many potential assignments, is the foundational basis of Neyman's randomisation inference (Ney-

[2] E captures features of the experimental environment – the mode or lab conditions, general experimental protocols, and other contextual factors – that condition how the experiment was conducted. Generalising over E captures external validity concerns – how well does the parameter estimate approximate the true effect in other experimental or non-experimental contexts. For instance, would a causal effect of an information prime delivered on a screen in a lab hold in a "real-world" context like a get-out-the-vote field experiment? For a more in-depth discussion of replicating effects across experimental contexts see, for example, Duch et al. (2020).

[3] For instance, consider the set of numbers $3, 2, 4, 9, 1$. The true mean of the set is 3.8. If we only ever sample two values our estimate will range from 1.5 ($\frac{1+2}{2}$) to 6.5 ($\frac{4+9}{2}$).

i	$Y_i(1)$	$Y_i(0)$
1	10	5
2	10	5
3	**10**	**10**
4	10	5
5	10	5
6	10	5

Table 3.1: Potential outcomes for six individuals under treatment and control conditions

man, 1990) yet receives less attention in current experimental research. Given a distribution of potential treatment assignments and resultant outcomes, how outlying and therefore significant is the *observed* difference between treatment and control? The substantive concern is, holding constant the subjects within our sample but re-randomising them into new treatment and control groups, how much would we expect the observed effect to differ? This question is inherently counterfactual. Unlike in the case of sampling uncertainty, where it is plausible to draw new observations from the population, we cannot draw new, unconfounded assignment schedules for the same individuals because they would already have been exposed to one treatment arm or another.

The counterfactual nature of this uncertainty does not negate the importance of it to researchers interested in estimating the true causal effect of a stimulus. To see why, consider the potential outcomes in Table 3.1. Note that all individuals perform uniformly except individual ($i = 3$) who always reports the same outcome regardless of the assigned treatment condition.

The true average treatment effect is $\frac{1}{M}\sum_{i=1}^{M} Y_i(1) - Y_i(0) = 4.2$. Suppose that the randomly drawn assignment vector $Z = \{1, 1, 1, 0, 0, 0\}$. Our best, unbiased estimate of the ATE is the difference in means between those actually treated and those actually in control, $\overline{Y_{i1}}(1) - \overline{Y_{i0}}(0) = 10 - 5 = 5$. But now suppose we draw a different assignment vector $Z^{'} = \{0, 1, 0, 1, 0, 1\}$. In which case, the ATE estimate becomes $10 - 6.7 = 3.3$. Z and $Z^{'}$ are equally likely assignment vectors given equiprobable treatment assignment. Neither

perfectly estimates the true causal effect for this population of six individuals. Different potential assignments therefore result in different ATE estimates even though the sample (and potential outcomes) are constant. Hence, for any single random draw of assignments there will be uncertainty over the estimate as a result of the randomisation procedure.

Variance in the estimated ATE for different assignment schedules is explained by treatment effect heterogeneity. In our simplified case, one of the six individuals is impervious to the treatment. The effect of treatment on this individual is zero, while for the other five individuals the effect is a uniform increase of 5. Causal uncertainty in the *average* treatment effect is a function of treatment effect heterogeneity – a common, and increasingly important, concern in experimental research (Grimmer, Messing and Westwood, 2017). When a given treatment has heterogeneous effects on a given sample, then the specific random assignment of treatment will result in an unbiased albeit uncertain estimate of the true ATE.

In general, when we extrapolate from a finite set of observations to a broader phenomenon, there will always be some form of uncertainty. This can be as a result of a finite number of units sampled from a wider population, or it can be as a result of inherently stochastic procedures like randomisation. Our statistical estimation in both cases are a best effort to infer, from incomplete data, a generalisable claim. Causal uncertainty is frequently ignored by experimentalists, despite its natural connection to randomisation procedures and our growing focus on heterogeneity in treatment effects. In fact, claims about the nature of the uncertainty estimated by researchers are rarely made explicit in research. Despite this, causal uncertainty is important to experimentalists – the controlled randomisation of a prime, vignette, or other condition means there are unrealised potential assignments over the estimate as a function of how treatment was randomised – irrespective of the number of observations relative to the population.

Moreover, understanding what type of uncertainty one is estimating has a substantive impact on the values in question. As Abadie et al. (2014) show, conventional estimates of uncertainty (based around the assumption of sampling uncertainty) can be overly conser-

vative estimates of the causal uncertainty – namely, the uncertainty due to other potential random treatment assignments. And as other statistical features are added to the data, like clustering or multiple observations per individual, researchers want to avoid erroneous claims about this uncertainty due to assumptions in the variance estimator about either the assignment or sampling procedures.

Sometimes experimentalists may simply want to ensure that their intervention worked – that is, for those sampled, to be confident that the observed change in outcome is a result of the treatment rather than a fluke of randomisation. In which case, experimenters are not worried about sampling uncertainty at all. The settings may be so far removed from "natural" populations that the question is not whether others' would react in the same way, but whether the actual instruments are effective at all.[4]

On the other hand, experimenters often want to estimate how likely the observed effect holds across a wider population. This is most often the case in randomised survey experiments where participants are sampled from much larger populations like general electorates. In which case, the more frequent concern about sampling uncertainty is important and should be embedded in the variance estimation strategy. Crucially, this concern is separate from the "causal" uncertainty in the parameter estimate.

In summary, experimentalists should choose their variance estimator dependent on the type(s) of uncertainty they want to account for. Two separate types of uncertainty – causal and sampling uncertainty – are particularly relevant to experimentalists. They also help structure how we appreciate extensions to variance estimation in the presence of grouped, or clustered data. The remainder of this paper focusses on this specific issue. How should experimentalists select their variance estimator, and justify its use, based on the experimental design? In the next section, I present principled guidance on variance estimation in

[4]Note also that sampling uncertainty is not the same as external validity – it is not a claim about how likely any observed effect is "in the wild". In the case where a researcher is concerned about whether the effect would hold given a different sample, the validity of that claim is still only with respect to the identical experimental conditions as it was originally conducted – visiting a lab or online portal, reading the same vignettes, exposure to the same rules and so forth.

experimental contexts where there are group-level features to the data.

3.2 Group-level structure in experimental data

Motivating examples

Study 1. Consider an observational study that assesses how a specific teaching resource correlates with student learning outcomes. Researchers collect student-level educational attainment measures across multiple classrooms, and record whether each classroom is using the resource or not. Note in this observational context access to the teaching resource is not controlled by the researcher. To estimate a correlation with students' performance, the researchers regress individual students' educational attainment on this binary indicator:

$$Y_i = \beta \times R_{C_i} + \epsilon_i, \tag{3.3}$$

where β is the effect of the teaching resource, C_i is the classroom which individual i belongs, and R_{c_i} is an indicator variable for whether i's classroom uses the teaching resource or not.

Crucially while data are sampled at the student-level, there are unobserved effects invariant within students of the same classroom but which vary across classroom assignments. For example, the qualities of individual classroom teachers (their enthusiasm, experience, teaching style) will be shared across students from the same class, but will vary across students in different classrooms. Since these factors will plausibly affect the outcome, but are unobserved in the data, the residuals of our statistical estimation will exhibit a group-level structure. More specifically, if characteristics of the specific classrooms play any role in students' educational attainment, then some part of the residual error from the classroom size regressor will be correlated by individual. That is, we can decompose the residual term such that $\epsilon_i = \epsilon_{C_i} + \gamma_i$, where ϵ_{C_i} is the error associated with unobserved effects at the classroom level, and γ_i is the random error component centred around 0.[5]

[5] I shall assume in all the following examples that there are no relevant covariates to control for. This assumption is unrealistic, but the focus of this paper is on potential bias of the treatment regressor's variance,

Study 2. Now consider a different study where researchers randomly assign whether classrooms use the teaching resource or not. Researchers again measure the same educational performance metric at the student level (Y_i). Each individual's exposure to the resource treatment is determined by their membership of a specific classroom. As in Study 1, therefore, this predictor is invariant within classrooms. Unlike in Study 1, however, whether a *classroom* has the resource is determined by random assignment. To estimate the effect of the teaching resource, the researchers estimate a very similar model to that in Study 1:

$$Y_i = \beta \times \mathrm{D}_{C_i} + \epsilon_i, \tag{3.4}$$

where β is the effect of the treatment, C_i is the classroom to which i belongs, and D_{C_i} is the randomised treatment assignment of classroom C_i. As in Study 1, the same unobserved classroom-level effects like teacher quality may separately impact the outcome. So again it seems reasonable to disaggregate ϵ_i into separate classroom-level and random components.

Study 3. Finally, consider a different experiment where students take part in repeated rounds of an experimental game. Researchers randomly manipulate a binary feature of the game (the treatment) each round. Since there are multiple rounds students will, in expectation, be exposed to both treatment and control conditions. Data is collected at the individual-round level. Thus researchers estimate the following model:

$$Y_i = \beta \times D_{it} + \epsilon_{it}, \tag{3.5}$$

where i indexes individual students, t indexes rounds of the experiment, D_{it} is the round-specific treatment assignment of i and ϵ is the round-level error.

Let us assume that there are no learning or other order effects at play in this scenario. D_{it} is randomised at the individual round-level such that our estimate of the size of β will be

and not in the variance estimates of covariates. Therefore we can ignore these covariates for the sake of simplicity.

Study 1		*Study 2*		*Study 3*	
Observation	R	Observation	D	Observation	D
$Y_{c_1,i=1}$	1	$Y_{c_1,i=1}$	1	$Y_{i_1,t=1}$	1
$Y_{c_1,i=2}$	1	$Y_{c_1,i=2}$	1	$Y_{i_1,t=2}$	0
$Y_{c_2,i=3}$	0	$Y_{c_2,i=3}$	0	$Y_{i_2,t=1}$	0
$Y_{c_2,i=4}$	0	$Y_{c_2,i=4}$	0	$Y_{i_2,t=2}$	1
...		...		...	
$Y_{c_C,i=I-1}$	0	$Y_{c_C,i=I-1}$	0	$Y_{i_G,t=1}$	0
$Y_{c_C,i=I}$	0	$Y_{c_C,i=I}$	0	$Y_{i_G,t=2}$	1

Table 3.2: Comparison of the data structures for three hypothetical studies.

R is an observational indicator about whether the main predictor is present or not. D is similarly binary indicator, except it is randomly assigned at the group-level (Study 2) or at the observation level (Study 3).

unbiased both within and across individuals. But since we have multiple observations for each individual, unobserved individual-level characteristics will be invariant across portions of our dataset. For instance, suppose that the experimental study measures some form of cognitive-performance like multiplying two numbers together. The treatment involves some additional distraction during the task. Regardless of the presence of this distraction, it is likely that some portion of the performance for each individual will be driven by their (unobserved) pre-treatment cognitive ability. Since this ability is invariant within individuals, and individuals complete multiple rounds across both treatment conditions, the residuals for each individual ϵ_{it} will likely be correlated. Again, therefore, our intuition is that the error term of the model can be decomposed into a group level component (in this case the student) and a random component.

Despite their different designs, all three studies have very similar data structures as summarised in Table 3.2. In each case the data exhibits clear clustering – by classroom in Studies 1 and 2, and by individual in Study 3. In Studies 1 and 2 note that the group members are students within classrooms, whereas in Study 3 the units of analysis are repeat observations taken from the same students. Each study has a single a binary predictor, multiple observations per group, and an expectation that unobserved group-level factors will impact the outcome.

How should we handle the potential impact of group-level factors on the error term of our predictor in each study? In terms of observational studies like Study 1, the guidance in the statistical literature is clear: cluster your standard errors (Angrist and Pischke, 2009).[6] Otherwise, your estimates of the uncertainty around your point estimates will be biased.

Should the same logic apply to either of the two experimental studies? On the one hand, Study 2 and Study 3 both have similar data structures to Study 1 and share the propensity for group-level influences on the outcome. Unlike Study 1, however, the predictor is randomised in both Study 2 and Study 3.[7] Does the fact that the predictor is randomised preclude the need to cluster?

Studies 2 and 3, moreover, differ from each other in terms of the level at which treatment is randomised. In Study 2, treatment is assigned at the group level. All observations within a group share the same treatment assignment. In Study 3, on the other hand, treatment is randomly assigned at the unit of observation and, consequently, we do not expect treatment assignment to be correlated within groups. More formally, let ρ_x denote the relationship between treatment assignment for observations within groups. ρ_x is the intraclass correlation coefficient of treatment assignment. In Study 2, we expect $\rho_x = 1$ since treatment assignment within each group is perfectly correlated. In Study 3, $\rho_x = 0$ in expectation. Does this difference mean that clustering is appropriate in one case but not the other?

Conflicting advice for clustering experimental errors

The literature is not settled on the answer to these questions (e.g. see Blattman, 2015), and specific and detailed guidance for experimentalists remains limited. Indeed, two recent working papers on the correct application of clustered variance estimators come to seemingly conflicting conclusions about contexts with randomised predictors (Cameron and Miller,

[6]Indeed, the classroom example is a modification of the motivating example in Angrist and Pischke's (2009) textbook example of when cluster-robust estimation is useful.

[7]Note, in Study 2, the randomisation across classrooms should ensure that features like teacher quality are balanced across treatment and control, even though observations are drawn at the student-level. While the parameter estimate will be unbiased, therefore, group-level features will likely impact the error terms as discussed above, and do so in a way that is correlated by classroom membership.

2015; Abadie et al., 2017). Whereas Cameron and Miller (2015) argue that randomisation typically precludes the need to cluster one's variance estimates, Abadie et al. (2017) suggest clustering is necessary in some, but not all, experimental protocols.[8]

These arguments are seemingly incompatible because they implicitly consider different types of design. Neither paper is solely focussed on contexts with randomised predictors, and as a result their prescriptions do not adequately generalise to the diversity of experimental designs common across the social sciences. In the remainder of this section I highlight the key insights from this research to provide a comprehensive guide to variance estimation in experimental contexts.

Cameron and Miller (2015) argue that 'if we think that either the regressors or the errors are likely to be uncorrelated within a potential group, then there is no *need* to cluster within that group' (p.17, emphasis added). It is unclear, however, what makes clustering unnecessary from their discussion. On the one hand, this could be interpreted as suggesting that when a regressor is randomised there will be no difference in cluster-robust and conventional variance estimates (and therefore clustering is not needed). This claim can be tested using Monte Carlo simulations. In Section 3.3 I demonstrate that in fact randomisation does not preclude differences in variance estimates. Therefore, this strategy does not give us a principled reason to choose one variance estimator over another.

On the other hand, a more robust interpretation is that one *should* not cluster because in expectation the correlation will be zero in randomised contexts. In contexts where treatment is randomised at the individual-level, any non-zero correlation is a fluke of a random assignment procedure rather than bias caused by group-level clustering into treatment or control. This argument is more promising – it does provide a principled approach to variance estimation – but only holds if treatment is uncorrelated with groups in expectation. For instance, in Study 3 treatment is randomised at the observation-level, so not only is the

[8] Another tempting solution is to check whether clustering makes a difference and report whichever error is larger. As discussed in the introduction, however, this conservative approach is misguided to the extent it conflates hypothesis testing with variance estimation, and misconstrues the fundamental statistical properties of the variance estimate.

point estimate itself unconfounded but *in expectation* the within-group correlation of treatment (ρ_x) should be zero. If, as in Study 2, treatment is randomly assigned across groups but is correlated within groups, then by definition ρ_x is non-zero. Therefore, cluster-robust and conventional estimators may differ. Which estimator is correct in this case, and why? And what if, aside from randomisation, researchers have sampling-related concerns about the groups in the data?

Design-based approach to clustering

Abadie et al. (2017) provide a general solution that incorporates similar intuitions to Cameron and Miller (2015) but which can be extended to clustered treatment assignment designs common in experimental research. Their approach builds on the broader theory behind experimental uncertainty discussed in Section 3.1. When choosing the correct variance estimator in the presence of clustered data and randomisation, experimentalists should consider both how clustering affects the sampling and treatment assignment of observations respectively. In short, clustering is appropriate, they argue, if either the assignment of treatment or sampling are conducted at the group-level.

Group-level treatment assignment. Consider first the case of a completely randomised experiment like Study 3. The estimated variance of the effect is, by definition, unrelated to the treatment. Even if there are group-correlated errors, the intergroup correlation of treatment assignment is zero in expectation and therefore we should not expect any bias in the variance estimate (under the causal uncertainty interpretation).[9]

With a single random treatment assignment schedule, however, there is a non-zero probability that $\rho_x > 0$ even with completely randomised treatment. For instance, in Study 3 any given subject may get treated more often than not (or vice versa) precisely because assign-

[9]This logic is very similar to what Cameron and Miller (2015) appear to imply. The difference is that Abadie et al. (2017) make clear that the rationale behind clustering is not made with reference to any observed treatment correlation in the data.

ment is random.[10] The observed correlations in treatment at the group-level are a quirk of an (as-if) random procedure conducted at the observation-level. They do not reflect inherent correlations in treatment assignment within groups. If we were to repeat treatment assignment multiple times, the correlations would average to zero. It is therefore not appropriate to cluster on the basis of these correlations because they exist purely by chance.[11] Note this claim is made without reference to the data. We can know whether $\rho_x = 0$ in expectation purely from the design of the treatment assignment.

When treatment is assigned at the group-level, however, error-correlations due to group-level unobserved factors will not be cancelled out (in expectation) by the randomisation of treatment across groups. In this context, the conventional variance estimator *is* inappropriate. Suppose you have two pre-existing groups – one is assigned to treatment and one to control. Suppose further that some unobserved feature uniformly affects the outcomes of individuals in the treatment group. Since all observations in that group receive the same treatment assignment, there is no way for the estimation strategy to disentangle this group-level influence on the outcome (despite treatment assignment over the groups being random). Our errors should therefore account for these deliberate error correlations, and we do so by using the clustered variance estimator.

Group-correlated random treatment assignment shares features with observational studies where clustering is used to correct variance estimates for unobserved correlations between groups, predictors and outcomes. An unobserved contextual variable may influence outcomes independent of the effect of the main predictor. If groups of observations share this contextual feature as well as values for the main predictor, then conventional estimates of variance

[10] Consider flipping a fair coin ten times. In expectation we should get five heads and five tails but, inherent to the randomness of the procedure, we may plausibly get six, seven, or even eight heads from the ten coin flips. Randomness does not, in and of itself, guarantee balance empirically. Random procedures are random precisely because the assignment to one value or another is determined by some probabilistic mechanism where the functional form is both known and independent of all other variables. See Titiunik (2020) for a good discussion on the various definitions of 'random'.

[11] In a similar sense, researchers would not re-run an experiment just because a certain individual received the treatment condition in six out of ten rounds (when the expected proportion would be five out of ten rounds).

will not account for these group-level error correlations. In a similar way, if experimental treatments are correlated within groups, and groups share some background unobserved features that impact the outcome, then we need to correct for that in our estimation of the variance. This occurs in the hypothetical Study 2 because treatment is constant within classrooms and students within classes will share unobserved features that influence the residuals of the statistical model. In short, one should only cluster one's errors *on the basis of treatment assignment* if treatment assignment is correlated within groups.

Group-level sampling. Researchers making claims about a broader population should ensure variance estimates account for the absence of those units not sampled in their data. In the case of clustered data, the variance estimator should take into account whether the number of observed *groups* is part of a wider population of groups. Imagine a completely randomised experiment that is conducted on two individuals. They complete a set number of rounds, the treatment is randomised at the round-level, and the experiment is well-powered. One of these individuals, however, refuses to engage in the experiment and no matter what treatment condition they receive they always report the same outcome. This experiment will have two likely features. First, the point-estimate of the treatment is likely to be close to zero. But moreover, since the number of observations per individual is large, we expect the conventional standard errors to be quite narrow.

Intuitively these errors do not reflect the fact that an obstinate subject is skewing the results. At first glance, the narrow confidence intervals suggest the point-estimate is very precise. The problem in this case is that the researcher wants to make a claim about the experimental effect on a wider population of individuals but by chance one of the subjects is seemingly skewing not only the point-estimate but also the uncertainty of that estimate. By swapping the obstinate individual for another well-behaved subject, we might expect the treatment effect to change. Therefore, the uncertainty around the original estimate should account for this variability.

In such situations, clustering is appropriate since it captures the fact that the variance estimate is contingent on which groups were sampled from a wider population of groups.[12] In the case of the obstinate subject, we would expect this to widen the confidence interval around the low treatment effect estimate, to capture the fact that the average treatment effect in the population will likely be higher than that recovered in this specific sample of individuals. In other words, when we sample at the group-level (selecting some groups but not others), our variance estimate should be adjusted to account for the fact that not all groups are included in the data. Note that in this case we cluster regardless of how treatment is assigned, to capture sampling rather than causal uncertainty related to which groups are observed. Again, however, once the sampling strategy is determined, whether the researcher should cluster standard errors is determined without reference to the observed data itself.

In addition to sampling being conducted at the group-level, two additional features must hold for the cluster-robust estimator to be unbiased. First, the number of clusters in the data needs to be reasonably large since the cluster-robust estimator developed by Liang and Zeger (LZ-estimator; 1986) operates asymptotically (Cameron and Miller, 2015).[13] Second, the number of *groups* in the data must be a small proportion of the groups in the population or, equivalently, the number of groups in the population must be assumed to tend to infinity. If they are not, then cluster robust variance estimates will be conservatively biased against the true variance (Cameron and Miller, 2015; Abadie et al., 2017).

This "vanishing proportion" feature of the typical cluster-robust variance estimator has substantial implications for experimental researchers. When designing their analyses, experimenters must decide whether they wish to make an inference about a population beyond the groups included in the data. There are many cases where this will likely be the case

[12] Where repeat observations are taken, individuals can constitute groups (as in Study 3).

[13] The LZ-estimator can be expressed as $V[\beta] = (X'X)\hat{B}_C(X'X)$ where $\hat{B}_C = \sum_{g=1}^{G} X_g'\hat{u}_g\hat{u}_g'X_g$ and $\hat{u}_g = y - X_g\hat{\beta}$ (Cameron and Miller, 2015, p.8). Hence it is clear that the estimator operates by averaging over the groups, and will therefore perform better when the number of groups in the data is large. The optimal number of clusters will vary dependent on the features of the clusters themselves (Carter, Schnepel and Steigerwald, 2017). Corrections to the cluster-robust estimator can be made to account for small absolute numbers of clusters in the sample.

– repeat-round games using a sample of individuals are typically used to infer the size of an effect on some wider human population (be it geographical or socio-economic). If generalisability is desired, then the requirement that the number of groups in the data is small relative to the population of groups must hold.

If the number of groups is not a vanishingly small proportion of possible groups, or researchers simply want to estimate the uncertainty of causal effects with respect to the specific sample, then cluster-robust estimation should not be used. In these cases, the data does not contain a vanishingly small sample of the clusters in the population – instead it contains the "population" of clusters itself. Consequently, the data's composition would violate the asymptotic assumptions needed for the cluster-robust estimator to be unbiased.

Summary. Table 3.3 distils the design-based approach and provides a simple reference for experimental researchers to use. If either treatment is assigned at the group level or the groups in the data are a small sample of groups from a larger population then clustering is appropriate (Abadie et al., 2017). If neither hold, then one should not use cluster-robust variance estimation. Two implications of this section are worth restating. First, researchers can determine the appropriateness of cluster-robust variance estimation before collecting data. Both the assignment and sampling strategies are known prior to data collection. Second, researchers should be sensitive to whether their analysis aims to make an inference about a population beyond their sample, and if so whether there are sufficiently many groups both in their sample and the wider population for the cluster-robust estimator to be unbiased. In all cases, researchers should make explicit the type(s) of uncertainty they quantify in their analysis.

What if experimenters choose the wrong estimation strategy? How big a difference does the choice of variance estimator make in practice, and therefore how biased can the standard errors be? In the next two sections I present simulation and replication evidence to show that the difference in variance estimation can be very substantial in experimental contexts.

	Sampled groups small proportion of all groups	*Sampled groups large proportion/all potential groups*
Treatment assignment at group-level	Cluster errors	Cluster errors
Treatment assignment at observation-level	Cluster errors (if making a population-level claim)	Do not cluster errors

Table 3.3: Should researchers cluster based on the design-based approach to variance estimation?

In turn, these results demonstrate the substantive importance of following the design-based approach outlined above.

3.3 Simulation evidence in favour of the design-based approach

This section expands upon the general problem of variance estimation in statistical terms. I focus in particular on the treatment assignment aspect of variance estimation. Sampling-related concerns are ubiquitous and covered in depth elsewhere (e.g. Huber, 1967; White, 1980; Liang and Zeger, 1986). Using Monte Carlo simulations, I demonstrate that even under observation-level randomisation it is probable that conventional and cluster-robust variance estimates will diverge substantially.

Conventional estimation of variance

Suppose some treatment D is linearly related to an outcome Y. The true model is specified as follows,

$$Y = D\beta + \epsilon, \tag{3.6}$$

where D is a $n \times k$ matrix, β is a $k \times 1$ vector, and ϵ is the randomly distributed residual. For the simplest experimental set up we consider, D can be considered a $n \times 1$ matrix of treatment assignments, and β the corresponding effect of treatment on the outcome.

With a finite sample of observations, the optimisation problem is finding $\hat{\beta}$ such that the squared residual ($\epsilon'\epsilon$) is minimised. In Appendix A1, I present the full derivation for completeness. Here it is sufficient to state that our estimate of variance $var(\hat{\beta})$ is the squared difference between the true average causal effect (β) and our best estimate ($\hat{\beta}$), such that:

$$Var(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'], \tag{3.7}$$

which is equivalent to:

$$Var(\hat{\beta}) = (D'D)^{-1}E[\epsilon\epsilon']D(D'D)^{-1}. \tag{3.8}$$

The variance of $\hat{\beta}$ in Equation B.6 is itself an estimate–in exactly the same sense that $\hat{\beta}$ is an estimate of the true β.[14] This feature is often neglected. Given a finite sample, we do not know the true variance of the treatment effect (and covariates). Instead, we observe a subset of the population, note the deviations from the predicted value (given the linear model above), and from this infer something about the true variance in the population. As such, the standard errors of the coefficients are themselves an estimate, and not a deductive by-product, of regression models.

Since X is known, from Equation B.6 our estimate of variance reduces to $E[\epsilon\epsilon']$, which is the variance-covariance matrix (Ω). Under conventional estimation, we assume that $\Omega = \sigma^2 \times I$. This specification is a direct result of the standard Gauss-Markov (GM) assumptions. First, for every value of D, the conditional expectation $E[\epsilon\epsilon'|D]$ is assumed to be constant (i.e. variance is homoskedastic). Therefore, the diagonal elements of the variance-covariance matrix are equal to some constant σ^2. Second, since we also assume error terms are uncorrelated, there is no covariance between any two distinct residuals (i.e. off-diagonal elements of the

[14]The Gauss-Markov theorem, which rests in part on the above assumptions, proves that $\hat{\beta}$ is the best linear estimator of β.

covariance matrix are zero). Hence the variance-covariance matrix is equivalent to multiplying the $n \times n$ identity matrix by σ^2. Imposing these restrictions on the structure of Ω, we only need to recover an estimate of σ^2, using the observed error in our sample: $\hat{\sigma}^2 = \frac{e'e}{n-k}$.

Any estimate of the true variance of $\hat{\beta}$ is entirely dependent on the simplifying assumptions we make about the nature of the variance-covariance matrix Ω. In particular, if either of the two GM assumptions are violated then our estimate of the variance of $\hat{\beta}$ will be incorrect (by definition). This latter feature matters hugely for statistical inference. Consider the parallel concern about the point-estimates $\hat{\beta}$. When we violate the GM assumptions – if for instance observations are not independent – our estimated coefficients are no longer unbiased linear estimators of the true β. In which case, our statistical inferences using conventional estimation are invalid. The same logic applies to variance estimation. If we do not satisfy the GM assumptions, our estimation of the variance of $\hat{\beta}$ are incorrect, and therefore any statistical inference about β is harmed because we have incorrectly estimated its variance.

Biased standard errors under the assumption of clustered data

Clustering in the data is a specific violation of the GM assumptions, whereby the covariance between off-diagonal elements of Ω are non-zero. This is true if there is group structure to residuals, such that $cov(\epsilon_i, \epsilon_j) > 0$ if $g_i = g_j$, else $cov(\epsilon_i, \epsilon_j) = 0$. In which case, $\Omega \neq \sigma^2 I$ (since multiplying the identity matrix means all off-diagonal elements are equal to zero) and so σ^2 is not an unbiased estimate of $\hat{\beta}$'s variance.

When Ω exhibits clustering, we cannot reduce $E[\epsilon\epsilon']$ to an estimation problem over σ^2 alone. Instead, assuming homoskedastic residuals, the correct specification of the variance of β is

$$\hat{\text{var}}(\hat{\beta}) = \sigma^2 (D'D)^{-1} D'V D (D'D)^{-1}, \tag{3.9}$$

where V is the block-diagonal matrix of inter-group error correlations ρ (Moulton, 1986).

In other words, we have substituted the identity matrix, which imposes strict conditions on the values of the variance-covariance matrix, with a matrix of the same dimension as I but which is capable of capturing the extent to which residuals may covary when observations belong to the same group.

V itself is unobservable since we cannot observe the entire population and therefore do not know the actual covariation between groups. Valid inference about coefficients' variance when data has group-level structures requires an estimate of both σ^2 and V. Incorrectly using the naive estimator of variance, when residuals exhibit within-group covariance, leads to estimation bias (Moulton, 1986; Liang and Zeger, 1986). The extent to which the conventional variance estimator is biased (expressed as the ratio between the true and naive variance estimate of $\hat{\beta}$) can be expressed as,

$$\text{Bias}_{\text{var}(\hat{\beta})} = 1 + [(\text{var}(m_i)/\bar{m}) + \bar{m} - 1]\rho_x\rho, \tag{3.10}$$

where ρ_x is the intraclass correlation coefficient (ICC) of x_{gi}. The extent to which variance is inflated or deflated by is known as the "Moulton Factor" (Angrist and Pischke, 2009) after Brent Moulton who first outlined this relationship (Moulton, 1986).

As either the ICC or correlation of errors within groups increase (decrease), the standard error is inflated (deflated). Moreover, since both ρ and ρ_x have a possible range $[-1, 1]$ (in Moulton's specification) it is possible the conventional estimator can be both overly-conservative or not conservative enough of the true variance of the predictor.[15] As the mean group size $\bar{m}$ decreases, or observations become more unevenly distributed across groups, then this would also inflate the conventional estimate of the predictor's variance.[16]

[15]This range is true of the ICC estimator as detailed in Moulton (1986) (see Griffin and Gonzalez, 1995). More recent ICC specifications have a strictly positive range: given a random effects model $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ for individuals i and groups j, the ICC can be expressed as $\frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\epsilon}$, i.e. the proportion of variance accounted for by the group over the total variance (Rodríguez and Elo, 2003). But note that even under the assumption that $\rho_x > 0$, if $\rho < 0$ then $\rho \cdot \rho_x < 0$.

[16]In a non-experimental setting, it is clear how a group-level regressor repeated in an individual-level dataset skews standard errors. Where the regressor is invariant within groups but variant across groups, $\rho_x = 1$ and so even with some small error-correlation within groups, the conventional estimator will be biased.

Clustering under complete randomised assignment

The Moulton Factor captures the extent to which the conventional estimator is biased if, in fact, one should cluster standard errors in a given design. More broadly the Moulton Factor is a scalar that reflects the extent to which a variance estimate changes as one switches between using conventional and clustered variance estimators. Usefully, therefore, we can use the Moulton Factor to understand whether and why clustering alters variance estimation under the assumption of randomisation.

Let $g \in G$ index groups of observations that each contain j observations (for the sake of simplicity).[17] D_{ig} is the treatment assignment for the ith observation within group g and y_{ig} denotes the corresponding observed outcome. We estimate the model $y^*_{ig} = \beta_0 + \hat{\beta}_1 D_i g + \epsilon_{ig}$ and focus on the potential bias of $var(\hat{\beta}_1)$ which is given by Equation 3.10. Under these assumptions, the Moulton factor collapses to,

$$\text{Bias}_{\text{var}(\hat{\beta}_1)} = 1 + [j-1]\rho_x\rho. \tag{3.11}$$

Hence, differences in the estimated variance rely on both the ICC and correlation of within-group residuals being $\neq 0$ and, trivially, that the number of observations per group must be > 1. Less trivially, as the number of observations per group increases, the potential bias of the conventional variance estimator increases (assuming some non-zero $\rho_x\rho$). Since both ρ and ρ_x must be non-zero for the estimated variance to be biased, I consider each inflation term in turn.

Correlated errors. Despite randomisation, it is likely that errors will be correlated within groups. Observation-level assignment will randomise treatment across observations within groups but that does not mean that inherent characteristics of that group are also randomised

[17] In fact, what follows below is not contingent on equal group sizes. Unequal group sizes affect the size of the variance bias factor, but not the two correlation terms ρ_x and ρ. To the extent that one's dataset may contain missing data or uneven group sizes, the actual impact of group-related errors is mediated by the square-bracketed term in Equation 3.10, instead of that in Equation 3.11.

across observations. Suppose we run a repeat-observation experiment on a set of individuals with a binary treatment and some outcome like giving in a dictator game. The treatment may affect one's willingness to distribute, but each individual will have some unobserved qualities that affect their latent attitudes towards giving – altruistic preferences for example. These features do not affect the estimate of the treatment coefficient, but they may affect the estimate of the coefficient's variance. The same logic applies in contexts where randomisation occurs at the group-level. Unobserved features of the group will be invariant across these observations (and invariant with the treatment indicator too).

Since unobserved qualities are constant across rounds, irrelevant of treatment, there may be some constant unobserved effect of each group's profile. In which case, $\rho \neq 0$.

Intraclass correlation of the treatment regressor. $\rho \neq 0$ alone is insufficient to lead to a difference between conventional and cluster-robust variance estimates. The total inflation factor is a product of both the error correlation term and the intraclass correlation of the treatment regressor (i.e. $\rho \times \rho_x$). If $\rho_x = 0$, then irrespective of any correlation in error terms across groups, there will be no effect on the estimated variance of the treatment effect. This feature appears to motivate the claim in Cameron and Miller (2015) that randomisation precludes the need for clustered variance estimation.

In a fully randomised experimental context the expectation is that $\rho_x = 0$ since treatment is randomised within groups. Since treatment is randomly assigned at the observation level, $P(D_{ig}) \perp P(D_{ig'}), \forall g \neq g'$. Hence the ICC of the treatment regressor will equal 0 in expectation. That is $\rho_x = 0$ when $P(D_{ig} = z) = P(D_{ig} = z'), \forall z, z' \in Z$ since the value of the regressor is not contingent on any feature of the group g, and so the proportion of variance explained by the grouping variable will be equal to zero.

In short, therefore, our estimates of randomised treatment effects *should not* be affected by repeat observations so long as the treatment is randomised within and across groups. Even if there is some group-correlated unobserved features in the data that are related to the

outcome, when $\rho_x = 0$ the scaling factor collapses to 1 and the clustered and non-clustered errors are identical.

Despite this expectation, given the stochastic nature of random assignment, any particular randomisation may result in $|\rho_x| > 0$. Even a small observed ρ_x has the potential to bias one's variance estimate if either the group size or within-group error correlation is sufficiently large. In fact, the probability of observing a non-zero ρ_x is dependent on the number of rounds. When j is small the probability of achieving truly randomised allocations of treatments decreases and thus the likelihood that $\rho_x \neq 0$ increases.[18]

Under randomisation for a binary treatment,

$$\lim_{j \to \infty} P(D_{ig} = x) = 0.5, x \in \{0, 1\}.$$

But when the number of observations per individual j is relatively small, the within-individual proportion of treated and non-treated rounds may not be perfectly balanced. In a two-round experiment, for example, the probability of a completely imbalanced treatment profile is the same as a completely balanced profile: $P(D_1 = D_2) = P(D_1 \neq D_2) = 0.5$. And across the full sample, even if $P(D_{ij} = 1) = 0.5$, there may be sufficient imbalance within individuals' own set of assignments such that $\rho_x \neq 0$.

Figure 3.1 demonstrates this result using Monte Carlo simulations of ρ_x where treatment assignment is completely randomly assigned, for varying numbers of rounds per individual. The expected value across all round sizes is equal to zero, but with fewer rounds, there is clear potential for the observed ICC value of the regressor to be non-zero. At the lower bound of group sizes ($j = 2$) the ICC distribution is notably shallow but as group size increases to five, a 95 percent interval covers [-0.03,0.03].

Given the results in Figure 3.1, imbalanced treatment assignment within individuals as a result of random assignment will not always preclude within-group correlation of treatment.

[18] Though this will be somewhat offset by reducing the potential bias of the standard variance estimator relative to larger group sizes (see Equation 3.11).

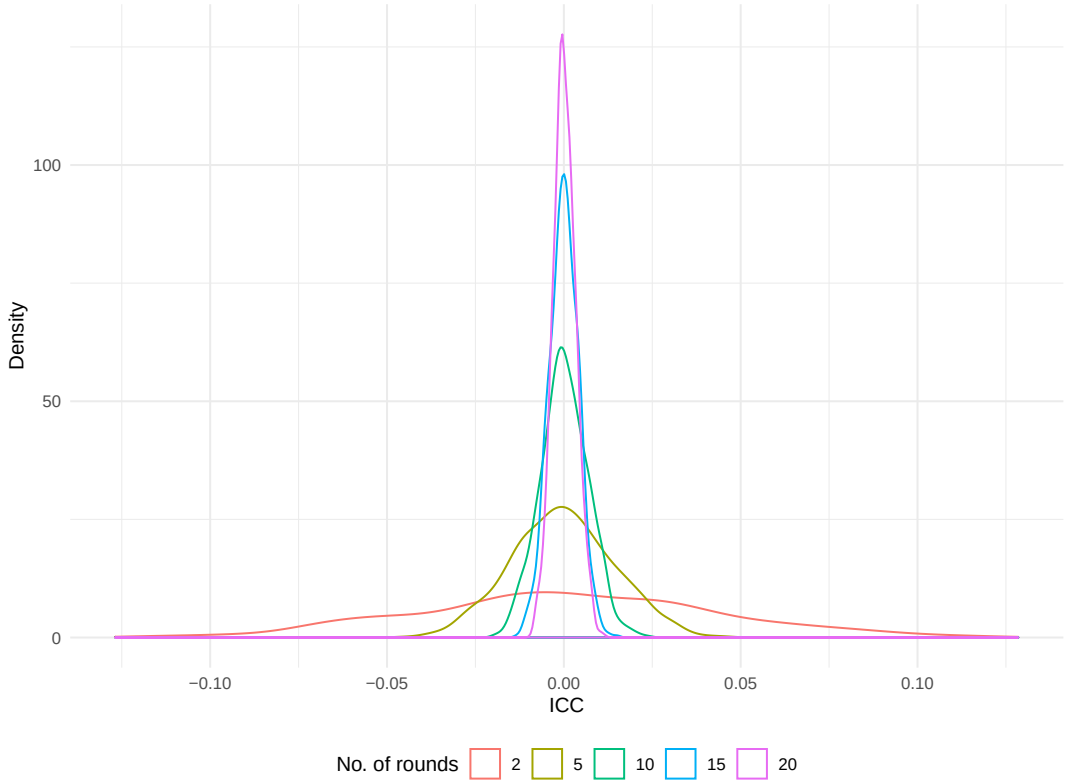


Figure 3.1: Distribution of Monte Carlo simulated intra-class correlations of the treatment regressor (ρ_x) with fully randomised treatment assignment over units. The coloured lines correspond to different numbers of observations per group j. Each simulation randomly assigns a binary treatment to $i \times j$ units, where $i = 500$. The distribution of values for each $j \in \{2, 5, 10, 15, 20\}$ is obtained by repeating the randomisation procedure 1000 times, calculating the ICC separately each iteration.

Thus, while the procedure of binary randomisation may itself be a stochastic process, there is no guarantee that the resultant treatment assignments *within* groups approximates a random distribution, such that there is some potential for $\rho_x \neq 0$.

In the case of group-correlated treatment assignment, $\rho_x \neq 0$ by definition. Therefore any group-level error structure as a result of unobserved features shared across members of a group will alter the estimated variance as determined by Equation 3.10.

Total change in variance. Non-zero values for both ρ and ρ_x are possible regardless of the level at which treatment is assigned. Multiple observations from the same group will likely

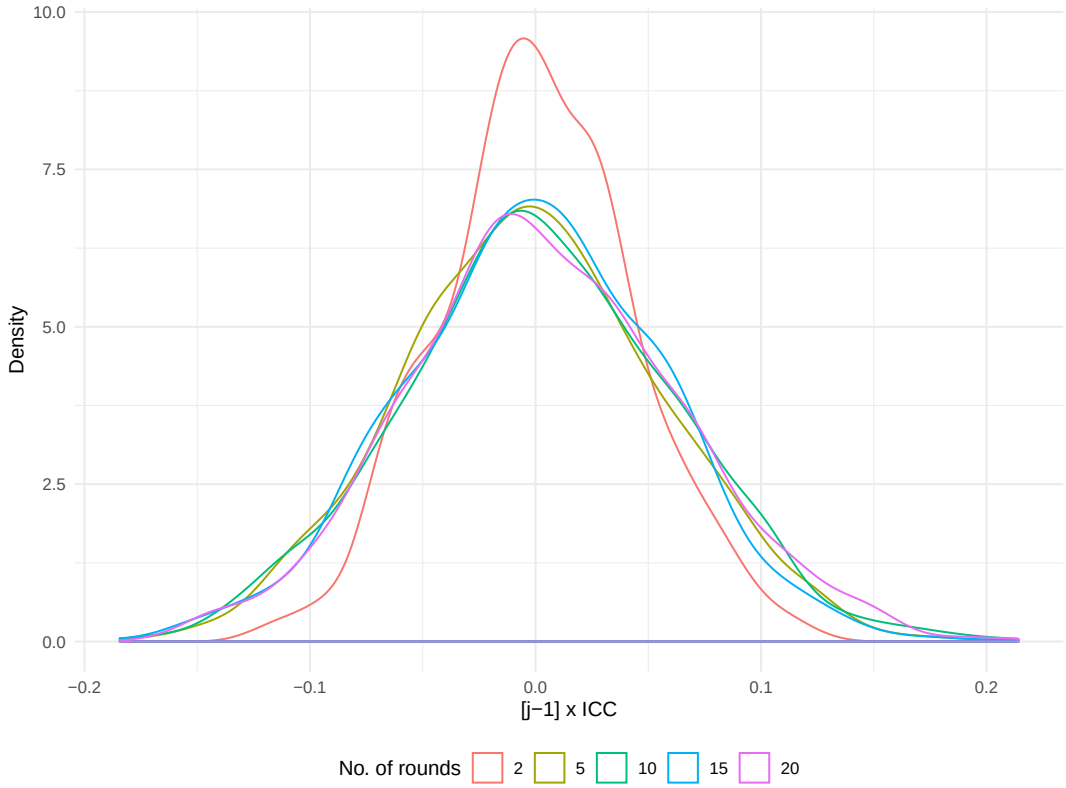


Figure 3.2: Distribution of the ICC estimates in Figure 3.1 multiplied by $[j-1]$. The effect of [j-1] is to offset any deflation in the variance of the ICC distribution as round size increases. Coloured lines again indicate different round sizes, holding fixed the number of individuals.

mean errors are correlated within groups. Even when randomisation is conducted at the observation-level, this does not preclude non-zero within-group correlations for any given random treatment assignment. Clustering will therefore likely make a difference to one's estimate of the regressor's standard error. But by how much?

In Equation 3.11 j features as both an inflater of the entire scaling factor, but also deflates the variance of ρ_x's random distribution (see Figure 3.1). When these two forces are combined, even though an increase in the number of rounds lowers the probability of a non-zero ρ_x, this is undone by the larger $[j-1]$ inflationary term that acts separately to ρ_x in the scaling formula. Figure 3.2 displays the results of multiplying the computed ICCs in Figure 3.1 by the corresponding $[j-1]$. As a result, even at relatively large number of rounds per subject, the variance of the random ICC distribution around the mean of 0 is considerable.

Second, given ρ has a potential range $[-1, 1]$, the total distortion of the standard error is moderated by the extent of the error correlation. At its maximum $\rho = 1$, errors are entirely correlated within individuals, suggesting that all residual variance in the outcome is explained by invariant features of the subjects themselves. As $|\rho| \to 0$, this mediates the entire scaling factor, and thus the potential divergence from the conventional estimate decreases.

The results in this section demonstrate that conventional and cluster-robust variance estimates can vary substantially. These simulations also offer direct evidence against the argument that randomisation precludes any difference between cluster-robust and conventional variance estimates. We are likely to observe differences in estimated standard errors between conventional and cluster-robust estimation, as a result of stochastic variation of treatment assignment within groups. These differences can be substantial, even for relatively simple experimental scenarios like the one explored in this section. Researchers should therefore pay careful attention to ensure they use the correct variance estimator, dependent on experimental design, as discussed in Section 3.2.

3.4 Review of Experimental Papers

To illustrate the importance of this paper's argument in substantive terms, this section presents a review of all papers published in the *Journal of Experimental Political Science* (*JEPS*) between 2017 and 2019 (n = 49). *JEPS* publishes short (4000 word) studies using experimental methods, making it convenient to analyse the variance estimation methods used across a wide corpus of research involving randomisation.[19]

In total, all but two of the articles published between 2017 and 2019 in *JEPS* contain original experimental studies. For each article, I record the research domain, the variance estimation strategy, and any justification given for the method of variance estimation. I also

[19] *JEPS* also states explicitly on its website its willingness to consider null results. Therefore, there is little *a priori* reason to suspect researchers would systematically modify their variance estimation in order to inflate the chance of recovering a statistically significant coefficient.

note whether the articles satisfy either of the two clustering criteria given by Abadie et al. (2017) and the analysis in this paper. Table A1 in the Appendix presents a comprehensive list of all articles reviewed in this exercise.

Table 3.4 reports all articles that reported using clustered standard errors. In almost all cases, authors are correct to use the cluster-robust estimator. For instance, Zhang (2018) cluster their results by experimental session since treatment is assigned at the experimental session level (all individuals in one session receive the same treatment). A majority of the papers in Table 3.4 cluster on the basis of cluster-correlated treatments.

Notably, authors rarely correctly justify their use of these procedures (if they justify them at all). DeScioli and Kimbrough (2019), for example, note that "to control for repeated measures and group effects, we include random effects for participant, and we cluster standard errors at the group level." While random effects control for between subject differences, the repeated measures and group effects in this design alone are insufficient justification for clustering. Clustering makes sense both because the treatment assignment is correlated by group and because the individuals are plausibly drawn from a much larger population. In Demel et al. (2019), the authors note that "standard errors are adjusted parametrically to account for clustering within decision-making participants," which again is insufficient reasons to use the cluster-robust estimator without noting either the clustered sampling or assignment (both of which the authors' design includes).

Two papers use cluster-robust variance estimation when it appears they should not. In Green, Davenport and Hanson (2019) the authors explore the long-term consequences of the Vietnam war draft on political behaviour. They leverage the natural randomisation of draft position being the result of lottery draws of birthdates within the year. The authors then cluster their standard errors by year since three separate random draws occurred between 1969 and 1971.[20] However, since the authors sample from across *all* draft lottery years, and the probability of assignment to "treatment" is equal between drafts, then neither the

[20] In an abstract sense, these draws can be though of as experimental sessions.

Table 3.4: *JEPS* articles (2017-19) using cluster-robust variance estimation

Authors	Description	Sample	Cluster-level	N per cluster	No. of clusters	Clustered sampling?	Clustered treatment?	Should cluster?
DeScioli and Kimbrough (2019)	Effects of information and communication on side-taking	*Not stated*	Group assignment	25	160	Yes	Yes	Yes
Demel et al. (2019)	Effects of employment status and ideology on redistributive preferences	Employed, unemployed, and students	Individuals	364	3	Yes	Yes	Yes
Loewen and MacKenzie (2019)	Effect of federalism on service representation	Political elites	Individuals	202	2	Yes	No	Yes
Zhang (2018)	Effect of enforcement institutions on likelihood of reporting corruption	Students	Session	20	18.5	No	Yes	Yes
Kingsley and Muise (2018)	Effect of communication in a public goods game	*Not stated*	Group assignment	21	25	Yes	Yes	Yes
Clayton (2018)	Effect of gender quotas on bias	Adults	Electoral district	135	7	Yes	Yes	Yes
Li and Zeng (2017)	Effect of facets of foreign direct investment on support in the recipient country	Adults	Individuals	1030	8	Yes	No	Yes
Klofstad (2017)	Effect of candidates' facial and vocal features on perceptions of competence	Adults	Candidate-pair	200	2	Yes	Yes	Yes
Green, Davenport and Hanson (2019)	Effects of Vietnam draft on behavour in the long term	Men	Birth year	3	300	No	No	**No**
Enos and Celaya (2018)	Effects of segregation on intergroup relations	Mturk/Student	Sessions	5	216	No	No	**No**

sampling nor treatment requirements of the design-based approach are satisfied.[21]

Similarly, in Enos and Celaya (2018), five separate experimental sessions are run on different subjects. In the pooled results reported in their Table 1, the results are clustered by session. Randomisation occurs *within* each experimental session. Moreover, it is unclear why or whether experimental sessions should be treated as sampling-based strategy. In the same sense that the draft lotteries constitute a population of potential clusters, the experimental sessions are the population of all sessions run. The potential "super-population" of all realised and unrealised experimental sessions seems, on the face of it, an unsatisfactory justification for clustering. The researchers did not, after all, sample from this set of potential groups prior to assigning treatment.

Among the set of papers that did not use cluster-robust variance estimation, four papers have designs where clustering would have been appropriate, and in two further studies a straightforward judgement is not possible. These papers are reported in Table 3.5.

Of the first four papers, half include cases where clusters were sampled from a wider population of clusters (Morton and Ou, 2019; Bassi, 2019; Barber et al., 2017). Morton and Ou (2019) note that clustering is appropriate for their design, on the basis of clustered sampling, but that privacy concerns precluded individuals from being identified in the data. Barber et al. (2017) similarly should cluster on the basis of clustered sampling. In this study, each subject considers multiple minority protections and is asked their opinion about the status quo. The researchers randomly vary whether a given minority right is presented as the status quo. Their analysis pools individuals' choices across the different minority protections. Since the unit of analysis is at the choice-level and individual subjects are plausibly a subset of a larger population clustering would have been appropriate.

The other two papers include treatments that are correlated at the cluster-level (Lorenz, Paetzel and Tepe, 2017; Jeon, Johnson and Robinson, 2017). In Lorenz, Paetzel and Tepe

[21]The treatment is operationalised as having either a low or high draft number. Since lottery draws were exhaustive (all days in the year were drawn), the probability of assignment to a low draft number is approximately 0.5 in each year the lottery was conducted.

Authors	Description	Sample	Clustered sampling?	Clustered treatment?	Should cluster?
Morton and Ou (2019)	Effect of public voting on prosocial behaviour	Students	Yes	-	**Yes**
Bassi (2019)	Effect of weather on vote choice	Students & university employees	Yes	Yes	**Yes**
Lorenz, Paetzel and Tepe (2017)	Effect of framing taxes on redistribution preferences	*Not stated*	-	Yes	**Yes**
Jeon, Johnson and Robinson (2017)	Effect of ethnic diversity on social sanctioning	Two ethnic groups	-	Yes	**Yes**
Barber et al. (2017)	Effect of status quo bias in ballot wording on minority rights	Mturk	Yes	-	**Yes**
Rogers (2017)	Effect of risk elicitation on charitable giving	Students	Yes	-	Maybe

Table 3.5: *JEPS* articles (2017-19) with group-level structure that do not using cluster robust estimation

(2017), the authors conduct a multi-round group experiment in which participants decide on levels of redistribution. Two features are completely randomised per round: the endowment one receives, and the group to which you are assigned. On this basis, there is no clustered treatment assignment. However, the experimenters are interested in how framing redistributive choices affects individuals' choices. They therefore vary whether participants are told that the game is about taxation or a minimum income, and this framing occurs at the session-level. Individuals are assigned to only one session. Therefore, there are session-level groups where treatment is assigned. Thus, in this analysis, cluster-robust variance estimation is appropriate. Session-level clustering issues occur in Jeon, Johnson and Robinson (2017) study of ethnic diversity and social sanctioning. Individuals from two ethnic groups

on the border of Kenya and Tanzania play a modified dictator game where a third-player can punish inequitable distributions by the dictator. The experimenters randomise the ethnicity of the different roles, leading to four separate treatment arms. As they note, when the ethnic composition of the session is balanced the four treatments are equally likely. But the composition of ethnicities within each session varies substantially (one session had a 71%-29% split). Since different ethnic compositions of sessions will affect treatment assignment probabilities, there is essentially clustered treatment assignment. The authors should therefore cluster at the session-level.[22]

Two studies proved more difficult to categorise. Rogers (2017) tests whether there are differences in risk elicitation experiments when individuals' earnings are given as charitable donations instead of direct payments to the participant. They use two separate risk elicitation experiments to test for these effects. Some, but not all, subjects play one of these games – the Bomb Risk Elicitation Task (BRET; Crosetto and Filippin, 2013) – twice. Therefore, in the analysis of this game there is a mixture of once- and twice-observed subjects. Later analyses focus just on those who completed BRET twice. The treatment over whether money is earned versus donated is orthogonal to assignment to the 2-round BRET game. Therefore there is no clustered treatment assignment in this experiment. However, since individuals are a smaller sample of a large population and there are multiple rounds (for 40 participants) there is a cluster-based sampling component to the design. How one should handle this situation, given that the full analysis contains both single- and double-observations is not clear. One pragmatic solution would be to run the analyses entirely separately for the two groups, and cluster variance estimates for those observed twice in the BRET game.

Bassi (2019) explores the effect of weather on vote choice between candidates who are identical except for their "risk". Participants are randomly assigned to sessions held on

[22]In this case the treatment probabilities are, in expectation, equal. To the extent this is a pre-data, design-related issue is debatable. If the randomisation procedure itself had produced imbalances purely by chance then clustering wouldn't be necessary. But since participation by members of ethnic groups is non-random, and the probabilities vary between sessions, it does seem pertinent to apply cluster-robust estimation.

days that had the largest difference in predicted precipitation, and complete ten rounds each of positively and negatively framed choices.[23] The payoff to choosing the risky candidate is sequentially altered over rounds such that rational voters should switch to voting for the non-risky option midway through each set of ten rounds. Bassi checks differences in preferences between positive and negative framing descriptively by comparing round-level proportions visually. To assess the statistical significance, however, Bassi sums the number of choices for the non-risky candidate per subject and conducts a Welch difference-in-means test. Similarly, to inspect the effect of weather, Bassi compares the total number of risky choices across all 20 rounds and compares again compares mean differences between good and bad weather.

This estimation strategy – summarising behaviour at the individual level – precludes clustering. However, it is not clear that this strategy is optimal. Some subjects may exhibit inconsistent behaviour, switching between candidates more than once. Rather than simply averaging away these issues, why not model the measurement error explicitly within the estimation strategy? A more robust strategy is to pool observations at the choice level, and model whether the risk-free incumbent is chosen on the basis of the difference in expected payoff, framing, and weather. In this proposed test, since the weather is invariant at the session-level cluster-robust variance estimation is appropriate on the basis of clustered treatment assignment. Bassi (2019) also implicitly draws individuals from a wider population, and therefore clustering would similarly be appropriate on the basis of clustered sampling.

Replications

How does the reported variance change, and thus how are the results affected, when we correct the variance estimator used in the papers reported above? I re-estimate the two incorrectly clustered papers in Table 3.4 and provide revised analysis of the experimental

[23] This framing is with respect to the earnings of a comparison country – either lower (positive) or higher (negative) than the expected payoff in the experiment.

results in Bassi (2019) from Table 3.5 using a pooled choice-level model.[24]

For Green, Davenport and Hanson (2019) and Enos and Celaya (2018), I compare an exact replication of their main results with the same estimation without using cluster-robust standard errors. Since Bassi (2019) does not pool results in the study itself, I construct a linear probability model for each of the four alternative weather measures used in the study.[26] I then compare the standard errors of the weather terms based on whether cluster-robust or conventional variance estimation is used.

Table 3.6 reports the results of this exercise. Each row is a separate model, and β refers to the main treatment coefficient. Overwhelmingly, the results show that correcting the variance estimation procedure increases the standard error of the treatment estimates – sometimes by a considerable amount. In both Green, Davenport and Hanson (2019) and Enos and Celaya (2018), conventional standard errors are *larger* than the original clustered standard errors. In neither paper does this change the substantive results,[27] but the percentage increase in the standard error is often over 10 percent and in the case of Enos and Celaya (2018) the standard error is almost doubled.

In Bassi (2019), the change in standard errors are even more stark. To be clear, the author does not report pooled models and so this is not a direct replication. That said, the sign of each model is in line with Bassi's findings – good weather decreases the probability of voting for the safe candidate (holding the expected payoff constant). Rainfall (in absolute and relative terms) increases the likelihood of risk-averse choices. The difference between cluster-robust and conventional standard errors is very large. In three of the four models, the clustered standard error is over twice the size of the conventional estimate. Moreover,

[24][25] Replication files are also available for Morton and Ou (2019) but, as noted previously, without subject identifiers it is not possible to compare clustered to non-clustered results.

[26] Model Specification:

$$P(\text{Vote for Incumbent})_{ij} = \alpha + \beta_1 \text{Difference in Payoff}_{ij} + \beta_2 \text{Framing}_{iij} + \beta_3 \text{Weather}_j + \epsilon_{ij},$$

where i indexes individuals and j experimental sessions.

[27] Indeed Green, Davenport and Hanson (2019) find null results for all estimations in their study. I focus on the partisanship measures, but the full paper considers other types of long-term effects too.

Coef.	N	β	SE	$SE_{Cluster}$	ΔSE	ΔSE (%)
Bassi (2019)						
Objective Weather	3320	-0.08	0.013***	0.033*	+0.02	+152.496
Subjective Weather	2660	-0.09	0.014***	0.032**	+0.017	+122.518
Absolute Rainfall	3320	0.06	0.013***	0.029	+0.016	+127.572
Relative Rainfall	3320	0.07	0.016***	0.025**	+0.009	+58.466
Enos and Celaya (2018): Original model clustered						
Black	1081	-0.12	0.032***	0.024***	+0.008	+34.014
White	1081	0.10	0.031**	0.016***	+0.016	+98.772
Green, Davenport, and Hanson (2019): Original model clustered						
Vote '12: Obama	899	0.82	3.339	2.915	+0.424	+14.5
Primary '16: Any Dem	648	-0.66	3.932	3.964	–0.032	–0.8
Obama Approval	898	-0.42	3.313	2.87	+0.443	+15.4
Party ID: Democrat	897	-3.82	3.221	3.032	+0.189	+6.2
Liberal / Very Liberal	912	-0.68	2.976	2.905	+0.071	+2.4
Don't Support Tea Party	907	1.03	2.525	2.232	+0.293	+13.1

Table 3.6: Replication of *JEPS* articles with incorrect variance estimates

the absolute rainfall model coefficient becomes insignificant at conventional levels, which is notable given the respective means difference in Bassi (2019) is significant.

In summary, this section has applied the design-based criteria to assess how well published experiments accord with the design-based approach advocated in this paper and in Abadie et al. (2017). I find both cases where researchers incorrectly use clustered variance estimators and cases where cluster-robust estimation should have been used. In addition, for the three studies replicated in this section correcting the variance estimator has large inflationary effects on the reported variance – in one case, undermining a published finding. Irrespective of whether the results affect the degree of confidence in the (null) results, researchers should be concerned whether reported standard errors accurately reflect the uncertainty in their

point estimates as an inherent estimation problem.

3.5 Practical guidance for experimentalists

Section 3.4 demonstrates that group-level data structures exist in a variety of experimental settings. Researchers should be clear about what type(s) of uncertainty they wish to capture and how this correlates with any group-level features of their data. If groups are sampled from a wider population of groups, or treatment is assigned at the group-level, cluster robust estimation is appropriate. Before concluding, it is worth briefly noting some concrete and practical implications for common types of experimental design.

Randomised controlled-trials. There are many reasons why researchers may assign treatments that are correlated at the group level. One particular instance is when treatment affects multiple observed units simultaneously (as in the hypothetical Study 2). This is most obviously the case in randomised controlled-trials (RCTs), which frequently involve group-level treatments (Peters, Langbein and Roberts, 2016; Kendall, 2003). For instance, Banerjee et al. (2007) conduct an RCT in which treatment is a standardized teaching programme assigned at the school year-group level. Maths and language scores at the student level are then compared across control and treatment schools for that year-group. Each student belongs to a school year-group in either treatment of control. Students of the same classroom, therefore, have perfectly correlated treatment assignment.

Clearly, in cases such as these where treatment is delivered at the group-level, $\rho_x \neq 0$ by design. All students in the same class receive the same treatment assignment, and therefore clustering standard errors is appropriate. Moreover, the magnitude of difference in standard errors between conventional and cluster-robust estimations will likely be much greater than in the fully-randomised context since ρ_x will equal 1 in expectation.

The researcher knows whether treatment assignment is assigned at the group-level prior to experimentation. How treatment is assigned is not a feature of the data, but of the

experimental protocol. Researchers can ascertain whether clustering is appropriate on the basis of treatment assignment prior to estimation (and even pre-register it).

This type of experimental design has a lab-based parallel. In some lab experiments, treatment may be only feasible at the experimental-session level such that all participants for a given session receive the same treatment.[28] This might be the case because subjects interact in a common experimental game, or because some feature of the lab must be altered in order for the treatment to be delivered. In which case, as in the RCT design, $\rho_x \neq 0$ in expectation. Moreover, while experimental sessions are not naturally-occurring groups like villages (groups where we would expect unobserved group-invariant features that affect the error term), experimental session groupings are know to influence outcomes (see Fréchette (2012) for a full discussion of session-effects). In which case, not only *should* researchers cluster based on clustered assignment, but also we would expect the difference in variance estimations to be substantial.

Conjoint experiments. Conjoint experiments typify designs that have both grouped data and observation-level treatment (as in the hypothetical Study 3). Conjoint experiments involve presenting subjects with a combination of two profiles, and forcing subjects to choose which profile they prefer. The value for each attribute (the attribute level) for each profile is randomised (Hainmueller, Hopkins and Yamamoto, 2014). To increase the efficiency of these designs, experimenters typically ask subjects to make choices across multiple sets of candidates. For instance, in a two-profile design in which n subjects make k choices, the dataset is $n \times k \times 2$ rows long. Since the attributes are randomised at the profile-level, conjoint experiments do not satisfy the assignment-based justification for cluster robust estimation.[29]

Group-level error structures will arise in conjoint designs. Each $k \times 2$ subset of the data

[28] Session-level treatment inherently captures some form of causal uncertainty. Given each group is assigned to one treatment arm or another, how different would our effect(s) be if the random treatment assignment differed?

[29] Unlike experimental designs discussed thus far, in the conjoint experiment there are multiple treatment coefficients. If each profile has A attributes and an average of L potential values per attribute, there will be $(A-1) \times L$ treatment coefficients (since each attribute has one omitted reference category). This does not substantively alter the variance explanation.

corresponds to a single individual. Therefore, any idiosyncratic features of the individual that affect their responses, or attitudes, will affect multiple observations in the dataset. For instance, while the potential values of each attribute are constant across individuals, how individuals interpret these values may differ (Dafoe, Zhang and Caughey, 2018). In which case, individual-level cognitive or semantic-processing differences may lead to error structures that are similar within an individual's set of responses, but differ between individuals.

The typical conjoint design does, however, employ cluster-based sampling. From the population of possible subjects, we randomly sample a vanishingly small number of individuals. Each of these individuals then completes multiple rounds of the experiment. Therefore, if we wish to make a generalised claim about the variance in the population, we *should* cluster standard errors. And since the population is likely to be much larger than the number of individuals sampled, we satisfy the asymptotic assumptions of the LZ estimator. If, however, our variance estimation is based on estimating the causal effect conditional on those chosen, then clustering is not required.

Researchers may wish to refrain from using cluster-robust variance estimators within conjoint designs if they either 1) want to assess the causal effect on this specific sample, or 2) field the experiment on a sample that likely contains a high proportion of clusters within the relevant population. Suppose we run a conjoint on politicians in a state legislature. Instead of having a small sample of subjects from a large population, with a sufficient response rate one's data would include a substantial proportion of all state legislators. In which case, it is inappropriate to cluster standard errors because the asymptotic assumption that $\frac{g}{G} \to 0$ is not satisfied. This guidance holds even if there are idiosyncratic features of individual legislators that lead to correlated errors across each respondent's choices.

3.6 Discussion

In this paper, I present a comprehensive guide to cluster-robust estimation in experimental settings. I build on recent debates about variance estimation in the presence of randomisation (Blattman, 2015; Cameron and Miller, 2015; Abadie et al., 2017) and simplify the correct logic for applied researchers. In particular, I motivate the design-based logic developed by Abadie et al. (2017) and provide concrete estimation advice for applied experimentalists.

In short, if either treatment is assigned at the group-level or the groups in the data are a small sample from a wider population then clustering is appropriate. Conversely, experiments with large proportions of the group population and observation-level randomisation should not use cluster-robust variance estimators, irrespective of whether group-level features influence the outcome.

Using Monte Carlo simulations and replication data, I demonstrate that randomisation does not eliminate differences between conventional and cluster-robust standard errors. Researchers will likely observe differences between these two variance estimates, and have to make a substantive decision about which is the unbiased estimate. The principled guidance of this paper aims to encourage best practice, and debunk assumptions about randomisation's impact on variance estimates.

More generally, this paper aims to highlight that point-estimate uncertainty is itself an estimation problem. Given its fundamental importance to hypothesis testing, variance estimation should be treated as seriously as the corresponding point-estimates. Since the actual variation of the population is unknown, being aware of how the design of one's experiment influences the appropriateness of different variance estimators is key for valid inference about experimental treatment effects.

References

Abadie, Alberto, Susan Athey, Guido Imbens and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" *arXiv:1710.02926 [econ, math, stat]* .

Abadie, Alberto, Susan Athey, Guido W Imbens and Jeffrey M Wooldridge. 2014. Finite Population Causal Standard Errors. Working Paper 20325 National Bureau of Economic Research.

Abadie, Alberto, Susan Athey, Guido W. Imbens and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88(1):265–296.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Banerjee, Abhijit V, Shawn Cole, Esther Duflo and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics* 122(3):1235–1264.

Barber, Michael, David Gordon, Ryan Hill and Joseph Price. 2017. "Status Quo Bias in Ballot Wording." *Journal of Experimental Political Science* 4(2):151–160.

Bassi, Anna. 2019. "Weather, Risk, and Voting: An Experimental Analysis of the Effect of Weather on Vote Choice." *Journal of Experimental Political Science* 6(1):17–32.

Blattman, Chris. 2015. "Clusterjerk, the Much Anticipated Sequel." https://chrisblattman.com/2015/12/11/clusterjerk-the-much-anticipated-sequel/.

Cameron, Colin A. and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50(2):317–372.

Carter, Andrew V., Kevin T. Schnepel and Douglas G. Steigerwald. 2017. "Asymptotic Behavior of a t -Test Robust to Cluster Heterogeneity." *The Review of Economics and Statistics* 99(4):698–709.

Clayton, Amanda. 2018. "Do Gender Quotas Really Reduce Bias? Evidence from a Policy Experiment in Southern Africa." *Journal of Experimental Political Science* 5(3):182–194.

Crosetto, Paolo and Antonio Filippin. 2013. "The "Bomb" Risk Elicitation Task." *Journal of Risk and Uncertainty* 47(1):31–65.

D'Adamo, Riccardo. 2019. "Cluster-Robust Standard Errors for Linear Regression Models with Many Controls." *arXiv:1806.07314 [econ]* .

Dafoe, Allan, Baobao Zhang and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26(04):399–416.

Demel, Simona, Abigail Barr, Luis Miller and Paloma Ubeda. 2019. "Commitment to Political Ideology Is a Luxury Only Students Can Afford: A Distributive Justice Experiment." *Journal of Experimental Political Science* 6(1):33–42.

DeScioli, Peter and Erik O. Kimbrough. 2019. "Alliance Formation in a Side-Taking Experiment." *Journal of Experimental Political Science* 6(1):53–70.

Duch, Raymond, Denise Laroze, Thomas Robinson and Pablo Beramendi. 2020. "Multi-Modes for Detecting Experimental Measurement Error." *Political Analysis* 28(2):263–283.

Eicker, Friedhelm. 1967. "Limit Theorems for Regressions with Unequal and Dependent Errors." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1(Statistics):59–82.

Enos, Ryan D. and Christopher Celaya. 2018. "The Effect of Segregation on Intergroup Relations." *Journal of Experimental Political Science* 5(1):26–38.

Esarey, Justin and Andrew Menger. 2019. "Practical and Effective Approaches to Dealing With Clustered Data." *Political Science Research and Methods* 7(3):541–559.

Fréchette, Guillaume R. 2012. "Session-Effects in the Laboratory." *Experimental Economics* 15(3):485–498.

Green, Donald P. and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* 16(2):138–152.

Green, Donald P., Tiffany C. Davenport and Kolby Hanson. 2019. "Are There Long-Term Effects of the Vietnam Draft on Political Attitudes or Behavior? Apparently Not." *Journal of Experimental Political Science* 6(2):71–80.

Griffin, Dale and Richard Gonzalez. 1995. "Correlational Analysis of Dyad-Level Data in the Exchangeable Case." *Psychological Bulletin* 118(3):430–439.

Grimmer, Justin, Solomon Messing and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4):413–434.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(01):1–30.

Huber, Peter J. 1967. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1(Statistics):221–233.

Imbens, Guido W and Michal Kolesar. N.d. "Robust Standard Errors in Small Samples: Some Practical Advice." . Forthcoming.

Jackson, John E. 2019. "Corrected Standard Errors with Clustered Data." *Political Analysis* pp. 1–22.

Jeon, Sangick, Tim Johnson and Amanda Lea Robinson. 2017. "Nationalism and Social Sanctioning Across Ethnic Lines: Experimental Evidence from the Kenya–Tanzania Border." *Journal of Experimental Political Science* 4(1):1–20.

Kendall, J M. 2003. "Designing a Research Project: Randomised Controlled Trials and Their Principles." *Emergency Medicine Journal* 20(2):164–168.

Kingsley, David C. and Daniel Muise. 2018. "More Talk, Less Need for Monitoring: Communication and Deterrence in a Public Good Game." *Journal of Experimental Political Science* 5(2):88–106.

Klofstad, Casey A. 2017. "Looks and Sounds Like a Winner: Perceptions of Competence in Candidates' Faces and Voices Influences Vote Choice." *Journal of Experimental Political Science* 4(3):229–240.

Li, Xiaojun and Ka Zeng. 2017. "Individual Preferences for FDI in Developing Countries: Experimental Evidence from China." *Journal of Experimental Political Science* 4(3):195–205.

Liang, Kung-Yee and Scott L Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73(1):13–22.

Loewen, Peter John and Michael Kenneth MacKenzie. 2019. "Service Representation in a Federal System: A Field Experiment." *Journal of Experimental Political Science* 6(2):93–107.

Lorenz, Jan, Fabian Paetzel and Markus Tepe. 2017. "Just Don't Call It a Tax! Framing in an Experiment on Voting and Redistribution." *Journal of Experimental Political Science* 4(3):183–194.

Morton, Rebecca B. and Kai Ou. 2019. "Public Voting and Prosocial Behavior." *Journal of Experimental Political Science* 6(3):141–158.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32(3):385–397.

Neyman, Jerzy. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5(4):465–480.

Peters, Jörg, Jörg Langbein and Gareth Roberts. 2016. "Policy Evaluation, Randomized Controlled Trials, and External Validity—A Systematic Review." *Economics Letters* 147:51–54.

Rodríguez, Germán and Irma T. Elo. 2003. "Intra-Class Correlation in Random-Effects Models for Binary Data." *The Stata Journal* 3(1):32–46.

Rogers, Jonathan. 2017. "Nothing to Lose: Charitable Donations as Incentives in Risk Preference Measurement." *Journal of Experimental Political Science* 4(1):34–56.

Schochet, Peter Z. 2013. "Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference." *Journal of Educational and Behavioral Statistics* 38(3):219–238.

Titiunik, Rocio. 2020. "Natural Experiments." *arXiv:2002.00202 [econ, stat]* .

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4):817.

Wooldridge, Jeffrey M. 2003. "Cluster-Sample Methods in Applied Econometrics." *The American Economic Review* 93(2,):133–138.

Zhang, Nan. 2018. "Institutions, Norms, and Accountability: A Corruption Experiment with Northern and Southern Italians." *Journal of Experimental Political Science* 5(1):11–25.

4 Knowledge or ignorance? Assessing the causal effects of campaign finance disclosure on vote choice

Fine-grained information about campaign finance is abundant in the American political system. But do these disclosures affect vote choice? Existing studies on the effects of disclosure tend to focus on voter perceptions, few explicitly test the robustness of disclosure to partisan signals, and none have compared the effects of disclosure across representative and direct democratic elections. This paper addresses these outstanding issues by implementing a series of conjoint experiments that jointly assess the effects of disclosing campaign finance information across types of election and varying levels of political information. I find that when partisan cues are not primed, participants do use various features of a candidate's financial profile to influence their vote choice. When these partisan signals are explicit, however, the effects of describing a campaign's financial profile on vote choice drop away. In ballot initiative contexts, I find similar results: in the presence of explicit policy choices, vote choice is not affected by disclosure.

Information about campaign finance is abundant in the United States. At both the federal and state level, most donors are required to disclose publicly both the amount and to whom they give money. Since *Buckley v. Valeo (1975)*,[1] the Supreme Court has continuously reaffirmed its position that mandatory disclosure provides voters with an "informational benefit" (Briffault, 2010). Voters, the Court argues, can infer relevant information about political campaigns. In the formal theory literature, too, disclosure mechanisms are argued

[1] *Buckley v. Valeo 424 U.S. 1* (1976)

to be Pareto improving features of electoral systems (Coate, 2004). Campaign spending gives voters greater information about candidates, but comes at the expense of favours granted to those interest groups who fund that spending. Voters therefore face a trade-off between knowledge about candidates and the compromises made in favour of narrow but well-funded interests (Ashworth, 2006).

Theoretical work has discussed potential mechanisms through which voters might use information from campaign finance disclosures (Prat, Puglisi and Jr, 2010), but what do voters do in practice?[2] Recent empirical research on the impacts of disclosure is mixed (Primo, 2013; Sances, 2013; Dowling and Wichowsky, 2013; Ridout, Franz and Fowler, 2015; Dowling and Wichowsky, 2015; Dowling and Miller, 2016; Wood, 2017, 2018; Rhodes et al., 2019; Spencer and Theodoridis, 2020). Estimating the independent effect of disclosure on observed vote choice is difficult given the endogeneity of factors like electoral viability and popularity to the actions of donors (Arceneaux, 2010; Brown, 2013). Hence the vast majority of research in this area has focussed on experimental manipulations in survey and lab environments. This research shows that while voters' attitudes towards candidates are not immune to the effects of disclosing donor information (Wood, 2017; Spencer and Theodoridis, 2020), the marginal informational benefits of doing so may be limited (Primo, 2013).

Several issues within this experimental field remain understudied. First, few studies have examined how disclosure affects vote choice in a forced choice context (Dowling and Wichowsky, 2013). Studies typically focus on separate assessments of each candidate (Ridout, Franz and Fowler, 2015; Rhodes et al., 2019; Dowling and Wichowsky, 2015), perceptions of corruption (Spencer and Theodoridis, 2020), or perceptions of a candidate's positions (Sances, 2013) or those of interest group's (Primo, 2013). While disclosure may shift voters' perceptions about candidates, are these shifts sufficient to induce a change in vote choice?

[2]This paper does not address the related issue of whether voters, of their own volition, access disclosure information. Some evidence suggests that voters exhibit limited proclivity to do so (Primo, 2013). Indeed, the sheer volume of disclosure now made available as a result of online reporting may make it harder for voters to access and infer relevant snapshots of campaigns (Briffault, 2010). Organisations like OpenSecrets.org and the National Institute of Money in State Politics do seek to simplify and summarise these disclosures, and in this paper I focus primarily on aggregated statements about campaigns' funding.

Second, it is unclear how robust disclosure's effects are to other relevant features of the electoral context. Do the observed effects of disclosure persist when partisan signals voters are likely to receive during an election are revealed? While some studies explicitly control for partisanship (Dowling and Wichowsky, 2015; Dowling and Miller, 2016; Rhodes et al., 2019), the extent to which partisanship intervenes on any effect of disclosure remains understudied. Third, previous experiments focus on candidate elections, but the effects of disclosure may be different in initiative campaigns. The lack of explicit partisan signals, and the single-issue focus, may mean voters attach different weights to various dimensions of an initiative campaign's funding compared to candidate campaigns. Initiative campaigns, moreover, are prominent policymaking venues in the United States, with vast sums of money spent for and against propositions (Stratmann, 2010). Despite this, we know little about the comparative differences on the effects of disclosure across types of election.

In this paper, I address these outstanding concerns using a series of experimental conjoint surveys on a subject pool of US citizens. I test the effects of disclosure on vote choice across two types of election – gubernatorial elections and ballot initiative races. I focus on five different facets of disclosure that may affect vote choice: the financial size of the political campaign, the average size of a donation to the campaign, the type of largest donor, proportion of funds provided by that donor, and the geographic origin of donations. Isolating effects of each of these facets refines our understanding of *how* voters use disclosure information to inform their vote choice, if at all. Conjoint experiments efficiently isolate the causal impact of disclosure when delivered in an aggregated and simplified manner, at the point at which individuals cast their "vote". Since respondents are posed with a forced choice between two candidates, or the choice to endorse or reject a policy proposal, these experiments approximate voting scenarios more closely than some previous work.

I modify the typical conjoint design by randomly varying not only the content of various campaigns attributes, but also the number of attributes displayed to respondents (Sen, 2017). This strategy helps address recent concerns about informational equivalences – ex-

plicit cues that affect beliefs about unobserved factors upon which a respondent acts – within experimental research (Dafoe, Zhang and Caughey, 2018). Randomly varying the number of attributes tests whether observed causal effects are resilient to richer information environments that more closely resemble the electoral context voters typically face.

I use this approach to help disentangle the "possible" effect of disclosure – whether voters are ever receptive to campaign finance cues – from the "actual" effect of disclosure – what voters are likely to do given other relevant cues that are common in more realistic electoral settings. I use the additional layer of randomisation to test whether the effects of disclosure are resilient to respondents' partisanship (and that of the candidate, campaign or donor in question) (Dowling and Miller, 2016; Rhodes et al., 2019; Spencer and Theodoridis, 2020). Half of all subjects receive additional randomised information about candidates' ideology, partisanship and political experience. This strategy more robustly tests whether any observed effects of disclosure operate only by priming individuals perceptions of relevant background characteristics.

I then compare these results to the same subjects' behaviour in hypothetical ballot initiative elections. By comparing across types of election, I can assess whether the impact of disclosure differs when voters consider specific policy issues as opposed to candidates for public office. This paper provides the first test of whether the extent of informational benefit differs between direct and representative forms of democracy (Briffault, 2010).

I argue that while disclosure can impact vote choice its total effect is negligible once overt but common partisan signals are included. When subjects are presented with aggregate disclosure information alone, voters do show some propensity to move away from candidates with high average donations and relatively concentrated groups of donors. But when the candidates' previous experience, ideology and partisanship are revealed, these effects are substantively reduced and statistically indistinguishable from zero. This paper finds little support, therefore, that in real world contexts disclosure has a distinct impact on vote choice.

With respect to initiative elections, moreover, I do not find any significant effect of disclo-

sure on vote choice. The similarity between these results and the candidate condition with partisan signals suggests specific policy information plays a similar role to overt partisan cues in the case of candidate elections. Subjects in the experiment appear to have relatively fixed political views on issues, which disclosure does not seem to alter.

Taken together, these results refine our understanding of what information voters find relevant when voting in different electoral contexts. Given the null findings, these results help explain why voters exhibit limited proclivity to access disclosure information of their own volition (Primo, 2013). More broadly, contributing to a longstanding field of work surrounding how relatively uninformed voters use new information to make political decisions (Lupia, 1994; Carpini and Keeter, 1996), these results suggest campaign finance disclosures do not provide substantial heuristic cues.

From a policy perspective, my results suggest that consolidating campaign finance data into a clear and easy-to-digest format has little effect on voters' evaluations of candidates and propositions. The Supreme Court's justification for disclosure, which increasingly rests on the informational benefit assumption, appears unfounded – at least in the aggregate form it is presented in this experimental context. That is not to say there are no benefits to disclosure at all. Campaigns may act differently, soliciting unsavory donations, if their actions are not visible. In that sense, disclosure (irrespective of who then accesses it) may dissuade unethical behaviour. And other forms of aggregated disclosure – news bulletins, investigative reporting, criticisms made by competing candidates – may still affect vote choice. More narrowly, democratic values do not appear to be safeguarded by voters actively refining or altering their preferences based on campaigns' financial profiles.

4.1 The informational benefit of disclosure

In *First National Bank of Boston v. Bellotti (1978)*,[3] which struck down expenditure limits in ballot initiative races, the Supreme Court argued that disclosure allows voters to evaluate

[3] *First National Bank of Boston v. Bellotti, 435 U.S. 765* (1978)

the arguments presented for and against proposed legislation, and thus bolsters a voter's ability to make informed decisions on both issue and candidate elections (Jiang, 2018). More recently, in *Citizens United v. FEC (2010)*[4], the Court opined that, since contributions are effectively a form of speech, voters should have the right to know who is speaking. The Court's position is that voters can productively use this information to inform their voting behaviour.[5]

These justifications, however, leave it unclear *how* voters use disclosure to evaluate campaigns, and whether in reality this has a substantial effect on vote choice (Jiang, 2018). Moreover, despite the vast amount of money spent in initiative campaigns in the US, it is unclear whether we should expect the informational benefit to be similar across types of election. Does disclosure make up for informational deficiencies when other sources of information are not forthcoming? Or does disclosure have a separable, independent effect on voter evaluations irrespective of the presence of other information?

Assume, in the most abstract terms, that voters make a choice (for a single candidate, or an ordering of preferences etc.) with the goal of maximising their own utility. This decision is multi-dimensional: voters may care about a candidate's or campaign's ideology, specific issue positions, valence characteristics, and so on. Assume further that the problem the voter faces is about estimation. Given a set of signals (across these multiple dimensions) can voters adequately estimate the relevant quantities needed to make the most appropriate choice, that is, the choice they would make under perfect information (Primo, 2013)?

Tasks like researching candidates' biographies or checking federal campaign submissions are taxing (Primo, 2013). Moreover, some plausibly relevant factors such as a candidate or campaign's competence, trustworthiness, and viability are typically hard to observe. Candidates and campaigns may well suppress information or qualities that are deemed harmful to their electoral prospects. Individuals overcome these overly-taxing cognitive demands by

[4] *Citizens United v. Federal Election Commission, 558 U.S. 310* (2010)

[5] In Appendix Section C.2 I discuss the implications of the *Citizens United* decision for disclosure-based regulation in more detail.

using information signals that enable less costly estimates of the position or valence of a campaign. These heuristics, "efficient cognitive processes... that ignore part of the information", are not necessarily more error-prone than other forms of knowledge acquisition (Gigerenzer and Gaissmaier, 2011, p.451). With respect to political decisionmaking, heuristics can be very effective for evaluating the content and viability of various policies (Lupia, 1994).

In theory, following the logic of the US Supreme Court, campaign finance disclosure may play the role of a heuristic device.[6] It seems unlikely that voters would favour candidates with higher donation totals, in and of itself. Rather voters may prefer campaigns with higher total donations because it signals something about hard to observe but relevant characteristics of candidates (Wood, 2017). The "informational benefit" of disclosure, therefore, is the extent to which this information enables voters to refine (i.e. reduce uncertainty about) their estimates of candidates and campaigns along the relevant dimensions.

To make this logic clearer, suppose voters' estimations of each candidate depend on only two factors: the candidate's ideology and their valence.[7] Voters have some utility function to discriminate between candidates across these two dimensions. A candidate may trade-off some level of competence for closer ideological proximity, and *vice versa*.

Panel (i) of Figure 4.1 represents a configuration of a voter's ideal preference, indifference curves, and uncertainty surrounding the positions of two candidates (A and B). V knows that A and B lie somewhere within the two grey circles respectively, with the grey crosses indicating their midpoint. Since the midpoint of B is just closer to V than the midpoint of

[6]Disclosure is, more accurately, two separate processes. First, electoral actors must disclose to government agencies their financial interests (contributions and expenditures) within electoral campaigns. Dependent on the size of the donation, disclosure is required both of the donor and the recipient. A separate stream of disclosure research focusses on the potential "chilling effect" of requiring individuals to declare their donations in the public sphere. The extent of this effect is contested, with some evidence suggesting disclosure has a negative effect on the likelihood of contribution (La Raja, 2014), while others argue the effects are, in reality, negligible (Wood and Spencer, 2016). Second, these financial submissions are disseminated to the public either through public records or via endorsement statements in political advertising.

[7]For the sake of abstraction, valence here is a bundled dimension that includes concerns about trustworthiness, effectiveness in office, and previous experience in office. In practice, these are likely to be separable dimensions with varying levels of uncertainty in electoral contexts.

A, V will choose B.

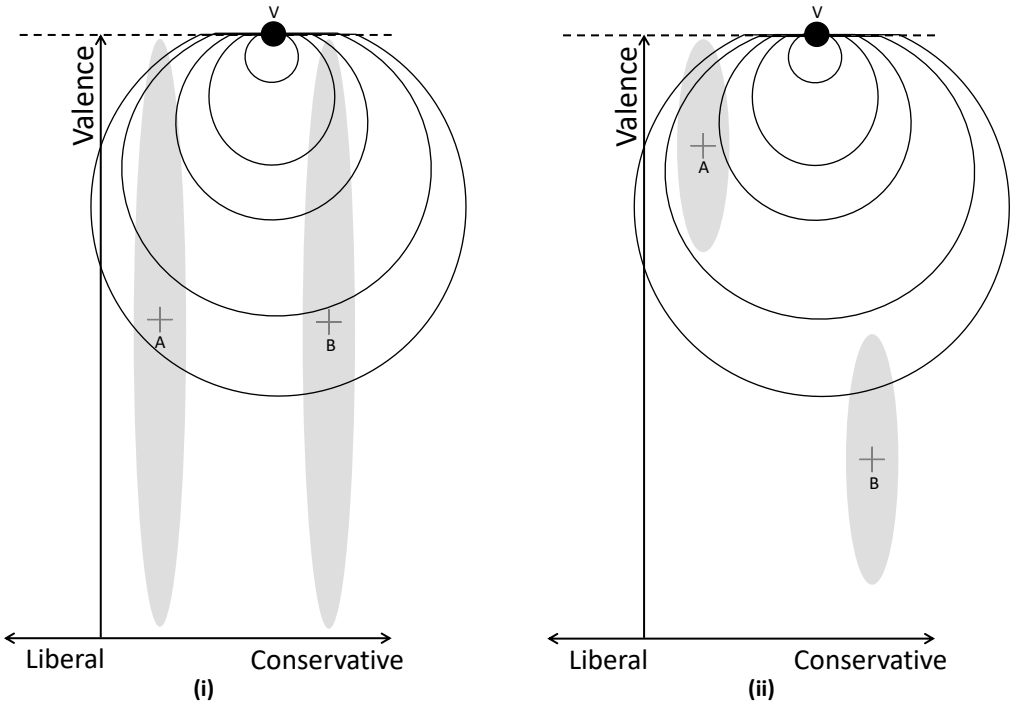


Figure 4.1: Hypothetical placement of a single voter (V) and two candidates (A and B) along ideological and valence dimensions. V is the voters ideal point. The capped oval lines are V's indifference curves. Grey circles represent the uncertainty of V about the two candidate's respective positions, with midpoints marked by crosses. Panel (i) indicates a situation of high uncertainty; panel (ii) indicates a situation of lower uncertainty after V receives additional cues.

Now suppose that V receives (effective) additional cues from campaign finance disclosure.[8] These cues alter the shape and size of the uncertainty surrounding the positions of A and B. Panel (ii) demonstrates this logic. The reduced uncertainty, as a consequence of signals gleaned from financial disclosure, mean it is much easier for V to discern which candidate is the optimal choice. A is now closer to V than the midpoint of B, so V will change their choice to A.

In summary, this model shows how disclosure can impact vote choice by refining uncertainty over candidate characteristics, making it easier to discriminate between multiple

[8]Different aspects of a campaign's financial profile may inform voters about valence and/or ideology. A campaign funded by business interests is likely to act as a cue about the conservative-leaning nature of a candidate. Large total donations, however, may signal that the candidate has high valence characteristics (Prat, Puglisi and Jr, 2010). I discuss these specific expectations at the end of this section.

candidates. Campaign finance cues improve the precision of voters' estimates regarding the position of a candidate in multi-dimensional space. But what about when relevant information is made explicit to voters? And does this logic also apply in non-candidate elections? I next discuss each of these issues in turn, before setting out my expectations regarding how specific facets of disclosure may affect voter evaluations.

4.1.1 Effects of disclosure with known information

The theory above suggests that irrespective of other visible characteristics of the candidate or campaign, information about funding reduces the uncertainty surrounding voters' estimations of valence. Disclosure, in other words, provides *additional* information for voters to use when assessing candidates. But not all characteristics are hard to observe. Candidates' partisanship is typically very visible and not subject to uncertainty. While a candidate's ideological position may be more uncertain, we would expect voters to typically know if candidates are "liberal" or "conservative" (or infer it from their partisanship). And voters are likely to know whether a candidate is an incumbent or not, even if they are uncertain over their trustworthiness or effectiveness.

Whether disclosure provides an additional informational benefit will depend on how sensitive voters' decisionmaking calculus is to the different dimensions at play. More concretely, if voters place more emphasis on the ideological dimension (compared to valence) and a candidate's ideology is not subject to much uncertainty, then disclosure may not in fact have an impact on voters' choices at the ballot. In which case, there will be no additional informational benefit of disclosure.

Consider Figure 4.2, which alters the weighting a voter places on the ideological and valence dimensions compared to Figure 4.1. V is now more sensitive to the ideological position of candidates than their valence, as indicated by the compressed indifference curves that are much more sensitive to shifts along the ideological dimension. The arrangement of positions is identical to Figure 4.1, but V now prefers B in panel (i). As before, panel (ii) depicts

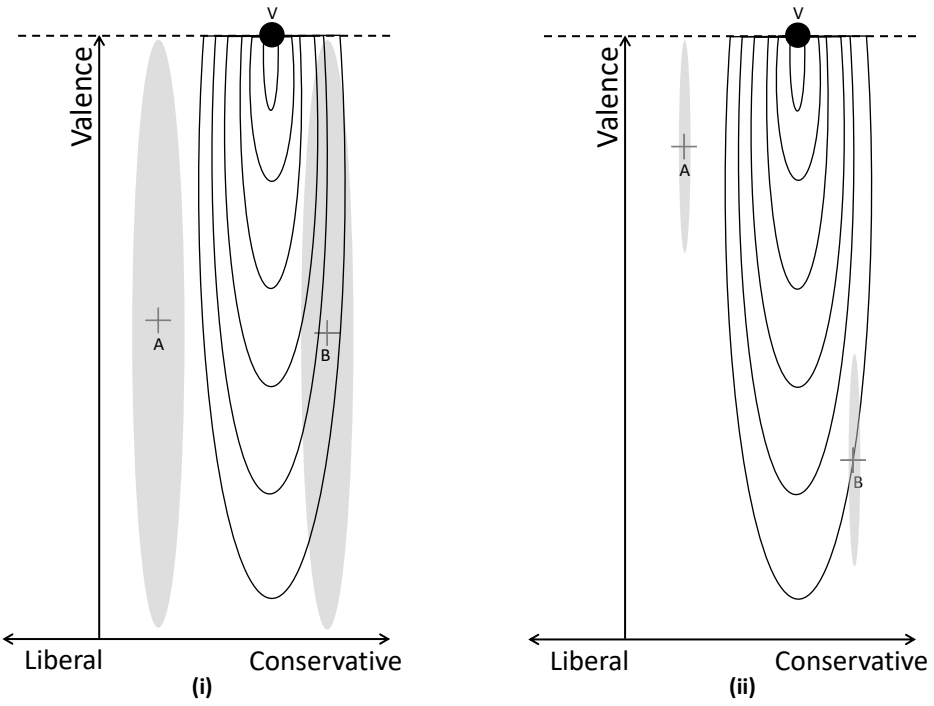


Figure 4.2: Hypothetical placement of a single voter (V) and two candidates (A and B) along ideological and valence dimensions. V's preferences are now weighted more heavily towards ideological concerns.

the refined estimates of V once disclosure and other relevant cues are taken into account. V's indifference curves remain constant. Since some background features about ideology are now explicit, V is almost certain of the ideological position of the two candidates. While disclosure has also reduced uncertainty on the valence dimension, V still continues to prefer B. Disclosure here, then, would not affect vote choice.

Put simply, if voters' preferences are dominated by ideological (or partisan) concerns, explicit primes may simply swamp any effect of disclosure. This result matters because such information is typically observed and highly salient in candidate elections. Existing evidence on the effect of partisanship on disclosure information is mixed. While some argue the effects of disclosure are mediated by copartisanship (Dowling and Miller, 2016), others have recently argued that the effects of disclosing group support can in fact be robust to the inclusion of partisan controls (Rhodes et al., 2019).

For campaign finance to provide an informational benefit, it must refine voters' choices even

in the presence of other explicit signals. Put another way, the effects of disclosure (if there are any) should be robust to the inclusion of other information, like party affiliation, that voters would expect to directly observe in an election and which increase voters' certainty over some, but not all, electoral dimensions.

4.1.2 Effects of disclosure across types of electoral campaign

The type of electoral contest may also mediate the effect of disclosure. Since ballot initiative elections are devoid of explicit party labels and other cues that typically guide vote choice, the benefit of disclosure could be *greater* (Garrett and Smith, 2005; Briffault, 2010; Primo, 2013). The US federal Courts itself is inconsistent on this issue. The opinion in *Sampson v. Buescher (2010)* for instance, argues the informational benefit of campaign finance is limited in direct democratic contexts. But other rulings have upheld the notion that the informational benefit is at least as important in initiative campaigns (Jiang, 2018).

Theoretically, there do seem to be "valence" concerns that are hard to observe directly, but which might influence vote choice. Voters may be concerned about the likelihood that the drafted legislation will in fact achieve its stated aims. They might also be concerned that the given legislation is anticompetitive, favouring a narrow set of out-of-state interests, even if they are broadly supportive of the policy in general. And voters may still want to estimate the ideological position of the policy to compare its compatibility with their other political beliefs.

On the other hand, while initiative campaigns lack overt partisan signals, they do focus on specific policy proposals – whether to increase the minimum wage, lower prescription drug prices or curtail state governments' taxation powers, for instance.[9] The specificity of initiative campaigns could, in fact, override other cues in a similar way to how partisanship signals may swamp the effects of disclosure in candidate elections. If voters utility functions

[9]It is worth noting that parties or their members can come out in favour of initiatives. These races are not necessarily totally devoid of partisan cues. The point is simply that initiatives are not labelled as 'Democrat' or 'Republican', and many issues are not explicitly supported by a party.

are weighted heavily towards the policy dimension (irrespective of other concerns), disclosure is unlikely to make a difference.

What remains unanswered is whether voters' behaviour is influenced differently across representative and direct democratic elections. To the best of my knowledge, this comparative aspect of disclosure (between election types) has not been tested before.

4.1.3 Relevant facets of disclosure

Measuring the effects of financial profiles on vote choice using observational data is difficult (Arceneaux, 2010; Brown, 2013). A campaign's financial profile is endogenous to the political context itself. Whether a candidate receives funding is a function of that candidate's prior likelihood of victory and assessments of electoral viability. Differences between safe and marginal seats will also alter the financial profile of candidate's campaigns. As a result, most works on the effects of disclosure have used experimental designs - predominantly with respect to political advertising. This research suggests that, absent other partisan signals, declaring the sponsor of an advertisement increases voters' abilities to place candidates ideologically (Sances, 2013). And while "attack ads" are less effective when disclosure reveals the identity of the sponsoring group (Dowling and Wichowsky, 2013; Ridout, Franz and Fowler, 2015), candidates can avoid a voter backlash if the attacking ad is sponsored by someone external to the campaign (Dowling and Wichowsky, 2015).

The focus on advertisements provides specific and compelling evidence that disclosure can affect voter decisionmaking, specifically by bolstering the inferences individuals make about politically relevant characteristics of candidates and campaigns. Voters do seem to infer (rightly or wrongly) different intentions by different donor sources (Dowling and Wichowsky, 2013). Yet advertisements are just one form of political expenditure, and revelations about specific ads tell voters little about the overall funding of campaigns. Attack ad sponsors may not be representative of a campaign's full financial profile (even if they are often visceral examples), and certain donors may select into or out of this form of highly visible political

activity.

To understand the broader effects of disclosure, therefore, I consider aggregate level disclosure – summaries of a campaign's full financial profile. Existing studies on aggregate disclosure demonstrate that voters value transparent profiles (Wood, 2017), and that aggregate disclosures can influence voters' perceptions of candidates' corrupt tendencies (Spencer and Theodoridis, 2020). On the other hand, Primo (2013) finds that the marginal benefit to voters of disclosure contained in news reports is 'trivial' – measured with respect to correctly identifying the positions of interest groups on ballot measures.[10]

These studies on voter perceptions are important, but leave open the question as to whether changes in perceptions based on disclosure affect voting itself. One paper (to the best of my knowledge) directly assesses the effect of aggregate disclosure on the likelihood of voting for a candidate, finding some support that this additional information alters vote choice (Dowling and Miller, 2016). Partisan information about the candidate is given to all respondents, rather than randomised between control and partisan-treated subsets.

To refine our understanding of these mechanisms, across electoral contexts, this paper focuses on five fundamental factors that describe different aspects of a campaign's donor profile - the total dollar-amount of donations, the average donation size, the proportion of funds from the largest donor, the type of largest donor, and the origin of donations. These factors are drawn from a combination of previous studies (Prat, Puglisi and Jr, 2010; Dowling and Miller, 2016; Wood, 2017; Spencer and Theodoridis, 2020) and further considerations about the possible facets of disclosure that might influence vote choice.

In comparison to previous studies, these facets seek to characterise financial profiles more comprehensively by capturing different types of inferences that voters might make. Below I discuss each of these factors in turn. Given the findings outlined above, there are competing intuitions as to both the size and direction of these effects. The purpose of the remainder

[10]While access to information is randomised in Primo's (2013) experiment, the option to actually view that information is left up to the subject. Primo finds both a low proclivity to access those disclosure-related articles, and insignificant differences in the number of correctly identified interest groups once other information accessed by respondents is taken into account.

of this section therefore is to motivate why these factors *might* matter. Where applicable, I highlight pre-experimental uncertainty about whether a given feature will have a positive or negative effect.

Total donations. The total size of donations is an indication of a campaign's scale. A relatively under-funded campaign, for example, is more restricted in its ability to carry out the political functions often seen as necessary for electoral success. A larger campaign is relatively unrestricted in the number of research, advertising, and get-out-the-vote operations it conducts. On the one hand, therefore, the total amount of campaign funding (holding constant its composition) is a potential indicator of *viability*. Smaller campaigns, unable to attract significant financial capital (relative to other campaigns), may be viewed as less viable by voters. Voters may use the size of the campaign as a signal of how donors, who may be more politically informed, have "pre-screened" the viability of the campaign in question.

On the other hand, voters may be distrustful of campaigns with very large donation totals (again, holding constant the composition of the campaign). If a campaign raises very large amounts of money, then voters may question the motivations of that campaign, since it might appear as if the campaign is trying to "buy" the election. Voters may (rightly or wrongly) perceive that large amounts of money mean a campaign has as an unfair electoral advantage, and react by tempering their support for the campaign.

Average donation. Where the total size of donations gives voters an indication of a campaign's viability and/or electoral capacity, the average donation seems likely to tell voters more about what *sort* of donor typically contributes to the campaign. This heuristic is already utilised by candidates in political campaigns. Both Bernie Sanders (Bump, 2016*a*) and Donald Trump (Bump, 2016*b*), for instance, touted their low average donation size (despite the relatively large total donations) as an indicator of their popularity in the 2016 electoral cycle. This trend continued into the midterms, particularly among candidates claiming strong ties to grassroots communities (Finkelstein, 2018).

Low average donations would suggest broader political support (holding constant the to-

tal donations), or at least that the typical donor comes from comparatively limited means. Conversely, a very high average donation might indicate that narrow but well-funded interests (the very wealthy or corporate groups) are the predominant base of support for that campaign. In all likelihood, if voters respond to this cue, we would expect them to be averse to candidates that are seen to serve narrow interests. It seems unlikely that the opposite effect would be true, that voters infer some positive quality from candidates whose average donation is very high.

Type of largest donor. The *type* of donor may give voters additional information about whether a candidate or proposal is broadly aligned with the sections of society they view favourably.[11] Broadly, donations can come from individuals or organisations. These organisations, for simplicity, can be divided into corporate, labor union and political advocacy types. Voters may associate any particular group's involvement in the campaign as a signal of the section of society that candidate or campaign is likely to serve. These signals may have differential effects dependent on one's pre-existing political leanings.

Moreover, a growing concern in the US system is the ability of certain donors to obscure their contribution activity through nonprofit "501" entities that, as charitable organisations, are not obliged to reveal their donors (Wood, 2018; Mayer, 2016; Rhodes et al., 2019). These "dark money" entities, while existing in a legal grey zone as 501(c)(4) entities, are especially problematic since their use is not uniformly distributed across the donor-spectrum. Dark money vehicles are useful primarily to those exceptionally wealthy individuals and groups who wish to obscure their involvement in the political process. This obscurity makes it very difficult to report such entities to voters through disclosure. While a 501(c)(4) organisation will be named, that name is typically mundane and uninformative to voters – for instance, "Americans for Prosperity".

If voters are unable to infer the source of donations – either because the name is withheld

[11]The Supreme Court itself has ruled on similar issues related to the type of donor. In *McIntyre v. Ohio Elections Commission (1995)* the Court argued there was limited informational benefit to disclosing information about private individuals.

or nondescript – they may shift their support away from that campaign. Or, as perhaps these groups hope, the reported name's obscurity may cancel out any potential cue to voters based on group or name recognition.

Proportion of funds by largest donor. Alongside the largest donor's identity, the proportion of a campaign's funds that are donated by a single donor may also matter to voters. This proportion reflects the degree of "capture" by any one particular donor/interest. A campaign solely funded by one corporate Political Action Committee (PAC) may cue voters differently to one where the same PAC is the largest donor, but contributes less than half of all funding. Different to average donations, this feature explicitly captures the concentration of financial support, rather than providing a signal about the base of support.

Intuitively, it seems likely that if voters use this cue, they would be adverse to campaigns funded by very few donors (i.e. where the largest donor donates a high proportion of funds). Campaigns with a high concentration of interested parties are likely to be those that most represent narrow interests.

Origin of donations. Finally, voters may care about where financial support comes from. Given the federal structure of the United States, voters may be concerned about whether campaigns are funded locally or not (especially for state level races). Whether a majority of donations come from within or outside of the state in question may provide some cue about "capture" by external interests.

This feature would have greater impact, hypothetically, if voters perceive some integrity to the *state* polity. Large numbers of donations from out-of-state actors could represent apparent interference in a state's affairs and thus may diminish voters' willingness to support a campaign or candidate. It may also be a (weak) signal of whether candidates care about the concerns of their constituency. This is particularly relevant for the sorts of elections considered in this article: state level races – for governor or ballot initiatives – that deal with state-specific political questions.

The above five factors are not an exhaustive list of features at the aggregate level. Taken

together, however, they are salient and aggregate-level indicators that characterise the overarching structure and composition of campaigns' financial profiles. These attributes can also be compared across different types of electoral races, and accord with similar features explored in previous research (Prat, Puglisi and Jr, 2010; Dowling and Miller, 2016; Spencer and Theodoridis, 2020).

4.2 Conjoint experiment to assess impact of disclosure on vote choice

To assess the causal effects of disclosure mechanisms across election types and controlling explicitly for potential informational equivalence, I conduct a series of conjoint experiments examining vote choice in the presence of disclosure. Subjects are presented with a forced choice between two campaigns – either two candidates or the support and opposition groups for an initiative proposal.[12]

Conjoint survey experiments are an efficient way to test the extent to which different attributes affect voters' choices in a robust, inferential manner (Hainmueller, Hopkins and Yamamoto, 2014). By randomising multiple attributes simultaneously, conjoint analyses enable researchers to distinguish the effects of each facet of disclosure separately. Concurrent randomisation of each attribute gives reasonable statistical power even for relatively small numbers of participants.

Moreover, since conjoint survey experiments typically ask respondents to choose between two profiles – in this case, to support or oppose an initiative campaign, or to vote for candidate A or B – this design is a natural analogue for the sorts of decisions voters make in American elections. Thus, we can assess subject behaviour within an experimental environment that more closely replicates the actual environment we wish to generalise our claims

[12] All replication materials – including R code and experimental data – is available at `https://github.com/tsrobinson/cf_conjoint`.

to. While this point is reasonably obvious with regards to two-party candidate elections, it also has benefits for the experimental study of ballot initiatives too. The decision a voter makes with respect to initiative policy is binary – whether or not to endorse a proposal – but there are typically separate "Yes" and "No" campaigns. Opposition groups raise their own funding sources and motivations, and play a key role in advocating for the status quo. Given the important role opposition groups play in initiative races (Gerber, 1999), there is good reason to include them in the study of disclosure mechanisms.

4.2.1 Randomising the number of attributes

Typical conjoint designs randomly vary the content of each attribute (the levels). This allows researchers to estimate the marginal effects of different features on respondents' choices within the experiment. The causal interpretation of these effects is defined with respect to the experimental context. That is, if an attribute-level has a significant positive effect on the likelihood of choosing a candidate, for example, the causal validity holds only given the context in which the decision was made. Such claims are very useful when we want to test whether a given feature or level *can* have an effect on a voter.

Claims about generalisability of the observed effect beyond the experimental context, however, require more stringent assumptions. Among other things, researchers must assume that the given set of attributes fully describe the pertinent features over which respondents make a choice.[13] The observed causal effects may not hold up in other contexts where some feature not included in the conjoint experiment acts on individuals' behaviour. For instance, voters may react, in abstract, to financial disclosures within the confines of an online survey but ignore them in reality when they make actual, electoral decisions.

Implicitly in this discussion we are referring to two different causal concepts. On the one hand, there is the purely experimental causal effect of shifting voter behaviour given the experimental conditions alone. This is the "possible" effect, in the sense that researchers

[13]Note this is a separate issue to the generalisability criticisms raised by Abramson, Koçak and Magazinnik (2019).

probe whether it is possible to shift the behaviour of respondents via certain treatments – in this case, attributes and attribute-levels. Regardless of whether all relevant features of the choice are included, the internal validity of any individual marginal effect is broadly ensured by randomisation procedures. On the other hand, the "actual" effect of a feature – namely that effect we expect to observe in real-life contexts – is the generalised effect from a given experiment. All controlled experiments will, to varying extents, only approximate the "actual" effect. The mere fact of controlled randomisation means that, even in field experiments, there is some manipulation that makes the study unrealistic.[14] In the case of disclosure-related behaviour, both of these effects are relevant. We want to know both whether respondents are capable of reacting to financial disclosures, and separately whether this effect is robust to more realistic contexts in which respondents know more information about the candidates themselves.

To distinguish these two separate effects, I randomly vary the number of attributes, as well as the levels within each attribute (Sen, 2017). Randomisation at the attribute level ensures unconfoundedness between those presented the full set of conjoint attributes, and those presented just the subset of financial disclosure cues. Comparing the estimated marginal effects for the common set of attributes shared across both groups helps illustrate whether there is both any potential effect of disclosure, and how robust this is to the inclusion of other relevant signals.

This approach is particularly useful if we suspect that certain mechanisms may be swamped by stronger information signals. A conjoint containing all pertinent features may show that a given feature has no significant effect on voter behaviour, but it tells us very little about why that is the case. Researchers cannot discern from this design alone whether the potential effect is swamped by other features or whether the feature itself is simply inert on decisionmaking entirely. The reduced conjoint without additional attributes, moreover, gives a more sensitive analysis of which, if any, mechanisms disclosure operates through.

[14]The exception are natural experiments, where randomisation is a natural feature of the context of study and is not controlled by the researcher (Titiunik, 2020).

In short, randomly varying exposure to other politically relevant signals allows me to test whether disclosure has an effect on vote choice, and if so whether this effect is robust to the inclusion of other cues not related to campaign finance disclosure. This approach helps to tease out the difference between the "possible" causal effects of an attribute, and the effect of the same attribute in more realistic, information-rich contexts.

4.2.2 Design

Participants were asked to complete two separate conjoint experiments – one choosing candidates in a hypothetical state gubernatorial election, and one focussed on a set of four initiative policy proposals. Table 4.1 provides details of all the conjoint levels across the two experiments. The specific attribute-levels are designed to give clear comparisons between different sorts of campaign funding. Figure 4.3 shows a demonstration of the funding information as presented to respondents in the candidate conjoint, and a screenshot of the initiative conjoint can be found in the Appendix.

The dollar amounts in the conjoint levels are intended to clearly distinguish campaign finance profiles within each relevant attribute. It is worth noting, however, that legal contribution limits to gubernatorial candidates vary across states (no such limits exist for initiative campaigns). While individuals in Oregon can donate unlimited amounts to gubernatorial candidates, Colorado caps private individuals' donations at $575 per candidate (the lowest limit in the sample of states).[15] This variance limits the external validity of these findings - a large average donation will be implausible given some state's contribution limits. However, the scenarios were presented as hypothetical and these levels did not seem to hinder subjects' completion or comprehension of the survey. Moreover, given the salience and scale of the type of elections under consideration, these levels are broadly plausible even if there is some mismatch with actual state laws. For instance, while $70 million in funding is large, Beto O'Rourke's 2018 senate campaign had receipts in excess of this level, and Proposition 61

[15] See the appendix for a summary of contribution limits within each state in the sample.

All conjoints	
Total Donations	$100,000 to $200,000 $1 million to $10 million $70 million to $90 million
Average Donation	$75 $10,000 $1 million
Largest Donor	Private individual Political Advocacy Group Labor Union Corporation/Trade Association Identity not disclosed
Proportion of funding from largest donor	10% 50% 90%
Origin of donations	Majority from donors within the state Majority from donors out of state
Information Equivalence Treatment	
Party	Democrat Republican Independent
Ideology	Very liberal Moderate liberal Centrist Moderate conservative Very conservative
Elected to previous office?	No previous elected positions Elected to state office Elected to federal office

Table 4.1: Conjoint attributes and levels

(2016) in California saw opposition donations exceed $100 million suggesting they are not implausibly high values in abstract.

Candidate conjoint. Respondents were presented with the funding profiles of two candidates running for gubernatorial office. Half of all participants ($n = 204$) were randomly assigned to see three additional informational equivalence treatments: the candidates' party affiliations, their ideological position, and whether or not they have been elected to either state or federal office before. These three variables fix subjects' priors and thus enable us to test for the independent causal effect of disclosure on vote choice. The exact same funding attributes and levels were used across these two versions of the candidate conjoint. The value

In this section, you are going to be presented with the descriptions of two hypothetical candidates running for **state governor**. Again, you can imagine this sort of information as what you would see in the run up to voting in an election.

We would like you to first evaluate the two candidates, and then to indicate **which you would vote for if you had to choose**. You will also be asked to rate how strongly you approve or disapprove of each candidate, on a scale from 1-7.

The table summarises the candidate's campaign funding - including total donations, average donation, type of largest donor, the size of their contribution, and the origin of donations.

You will be asked to choose between 6 pairs of candidates.

Q56.

	Candidate A	**Candidate B**
Average Donation	$1 million	$1 million
Largest Donor	Political Advocacy Group	Political Advocacy Group
Proportion of Campaign Funds from Largest Donor	90%	90%
Origin of Donations	Majority from donors within the state	Majority from donors out of state
Total Donations	$70 million to $90 million	$70 million to $90 million

Figure 4.3: Screenshot of candidate conjoint (without ideological, partisan, and valence control attributes)

for each of attribute and candidate were randomly assigned. Figure 4.3 displays an example of what respondents were presented for each round of the candidate conjoint experiment.

Initiative conjoint. Participants were presented with four different initiative topics in separate conjoint rounds. The text for each initiative is shown in Table 4.2. These topics represent the sorts of issues likely to be considered on the ballot given proposals that have occurred in recent electoral cycles, and have been selected to capture a variety of different contemporary issues. The policy proposals are hypothetical but constructed to appear sufficiently realistic that a subject could imagine such proposals being placed on the ballot.

Alongside the initiative title and a brief description of the proposed policy, participants

Initiative title and description
Marijuana legalisation
If passed, this initiative would legalize the sale of marijuana within the state for recreational use for those aged 21 and over, subject to taxation and regulation by state authorities.
State minimum wage increase
If passed, this initiative would raise the state minimum wage for adult workers to $14 per hour within two months of enactment.
Bond issuance for sewage redevelopment
If passed, this initiative would authorise the state government to issue a bond worth $300 million in order to fund a sewage system redevelopment scheme, updating the sewage network within the state.
Carbon emissions tax
If passed, this initiative would impose a 5% emission-based CO2 tax on the sale of all non-electric and non-hybrid vehicles, as well as an additional point-of-sale surcharge of 2 cents per litre on all fuel purchases.

Table 4.2: Hypothetical initiative policies

were presented with the same funding table as in the candidate conjoint. Funding information was listed for both the support *and* opposition campaigns. Participants were asked to consider this information and choose whether they would vote 'for' or 'against' the proposed policy. Each participant makes a total of four choices within the initiative experiment – one per issue. The order in which the issues appear was randomised.

Note that, unlike most typical conjoint designs, in the case of the initiative conjoint there is only one campaign over which respondents make a choice – whether to accept or reject the campaign. That said, as in actual initiative elections, groups mobilise around support and opposition campaigns to persuade voters to vote for or against the initiative. I treat these opposing campaigns as if they are separate 'candidates' in a typical conjoint design. Once an initiative has been qualified to the ballot, both groups are able to separately raise and spend money. It is important to capture the fact that opposition campaigns in initiative races are their own separate entities that make strategic decisions and attempt to influence the race's outcome independently of the proposition's campaign (Gerber, 1999).

Randomisation procedure. The order in which the candidate and initiative conjoints were presented to each subject was randomised, as was the assignment to the two different candidate conjoints. Within the initiative conjoint component, subjects were shown each initiative issue once and the order of these issues was randomised to limit any order-effect. Finally, across all conjoints, attributes were randomised with minimal restrictions to prevent implausible attribute-level combinations.[16] The appendix presents a series of tests verifying that these randomisation procedures were successful.

Sample. To ensure greater external validity of the results of this study, and to ensure that the types of election and information presented are meaningful to respondents, the conjoint experiments were conducted using a U.S. specific subject pool maintained by the Centre for Experimental Social Science, at Nuffield College, University of Oxford.

Members of the subject pool were invited to participate if they were resident in a state which used the initiative process and therefore were likely to be familiar with the process. The first round of invitations was sent to those resident in California, Washington, Oregon, Arizona, Ohio, Florida, Colorado, and Massachusetts - all states with relatively high usage of the initiative process. Further invitations were then sent to those resident in the other 17 states where some form of initiative provision is in operation. In total, 390 eligible participants completed the experiment.

To ensure a similar baseline level of understanding across the sample, participants were asked to read a passage of text describing basic features of candidate and ballot initiative elections, as well as campaign finance. Participants were told they would have to answer three factual questions related to the text, and would be remunerated for each correct answer given. On average, subjects answered 2.4 questions correctly indicating a good level of understanding about these elections having read the information.

It is worth noting that there is a reasonably substantial bias in the sample with respect to partisan identification. 46 percent of respondents identified (post-experiment) as Democrats,

[16] For instance, a campaign raising $100,000-200,000 could not have an average donation of $1 million.

30 percent as independents, and 15 percent as Republicans. Initiative states may simply have more Democratic supporters (particularly those states on the West Coast oversampled in the first round of invitations). To assess the plausibility of the Democratic lean observed in the sample I take the average of the difference in party affiliations at the state level,[17] weighted by the proportion of respondents per state in the experimental sample. While the lean in the experimental sample is much larger, the expected lean towards the Democrats is nevertheless substantial (7.3 percentage points) suggesting that some Democratic bias is to be expected.[18] Further demographic information is available in the appendix.

Causal assumptions. For the models in conjoint analyses to have a causal interpretation, further assumptions about the design and implementation must be met (Hainmueller, Hopkins and Yamamoto, 2014). In short, these assumptions do hold. In the appendix I provide a detailed discussion of these criteria, and present a series of tests that verify each assumption – stability, no profile-order effects, randomisation, and balance – for this experiment.

4.3 Results

4.3.1 The effect of aggregate disclosure

The first candidate conjoint (fielded to half of all subjects) presented only financial information to voters (and without partisan, ideological, or valence cues). To recover the marginal effects of each attribute, I estimate the following model[19]:

$$\begin{aligned}\text{logit}(\text{Vote for candidate?}) = \beta_1 \times \text{Average Donation} + \beta_2 \times \text{Total Donations} + \\ \beta_3 \times \text{Largest Donor} + \beta_4 \times \text{Proportion from Largest Donor} + \beta_5 \times \text{Origin of Donations}. \quad (4.1)\end{aligned}$$

[17] Data taken from Gallup's 2017 summary of state party affiliation, available at https://news.gallup.com/poll/226643/2017-party-affiliation-state.aspx.

[18] Moreover, given the aim of the study is to test for the presence of disclosure mechanisms, rather than estimate their effect in the wider population, the representativeness of the sample is not of paramount importance.

[19] The results reported in this paper are not sensitive to the estimation strategy. Similar results can be obtained replacing the logistic estimator with OLS regressions (see replication code).

Equation 4.1 recovers an estimate of the marginal effect for each attribute-level, relative to a baseline. The model is wholly additive – I estimate the independent marginal effects for each attribute, and not the interaction between any two attributes. This strategy bests matches the theory in Section 4.1.3, which isolates non-interactive mechanisms through which different facets of vote choice might affect voter decisionmaking. The goal is to isolate what features of disclosure, if any, condition vote choice.[20] All standard errors are clustered at the individual-level to reflect the fact that individuals (from which we take multiple observations) are sampled from a much wider population of individuals (see Chapter 3).

Figure 4.4 plots the model coefficients. In this low-information environment, atypical of real electoral contexts, various disclosure mechanisms do have clear effects on vote choice. Voters are less likely to vote for candidates who receive a large proportion of their funds from a single donor, nor where the average donation amount is high. Donations mainly from within the state, labor union funds (relative to private individuals), and lower average donation amounts (relative to the baseline $10,000) all have a positive effect on the likelihood of a candidate being chosen.

These results show that aggregate disclosure *can* influence voter decisionmaking. In the absence of other cues, subjects are averse to instances where candidates appear to be 'captured' by a particular interest or group. In particular, voters appear to care about political capture rather than the scale of a campaign itself. This feature is reflected in the relative importance of the geographic, average donation, and proportion attributes. Notably, the total size of a campaign had no substantial effect on voters' choices.

[20] An interactive theory would suggest that voters combine multiple aspects of disclosure to form an assessment of candidates. For instance, they compare both the average donation size (say it is small) and the total raised (assume it is high) and infer that this candidate is widely supported by regular voters. These sorts of interactions are plausible, but in theory assume more complex heuristic processes. Given the limited research in this area exploring the fundamental effects of disclosure, I defer these theoretical issues to future research. More practically, to estimate the marginal interactive effects of attributes within this conjoint design would require a larger sample size. Without strong theoretical reasons to exclude certain interactions, a fully-interactive model would inflate the number of estimated parameters and therefore require more observations to maintain sufficient statistical power. Inferences on the basis of an underpowered interactive model would be misleading.

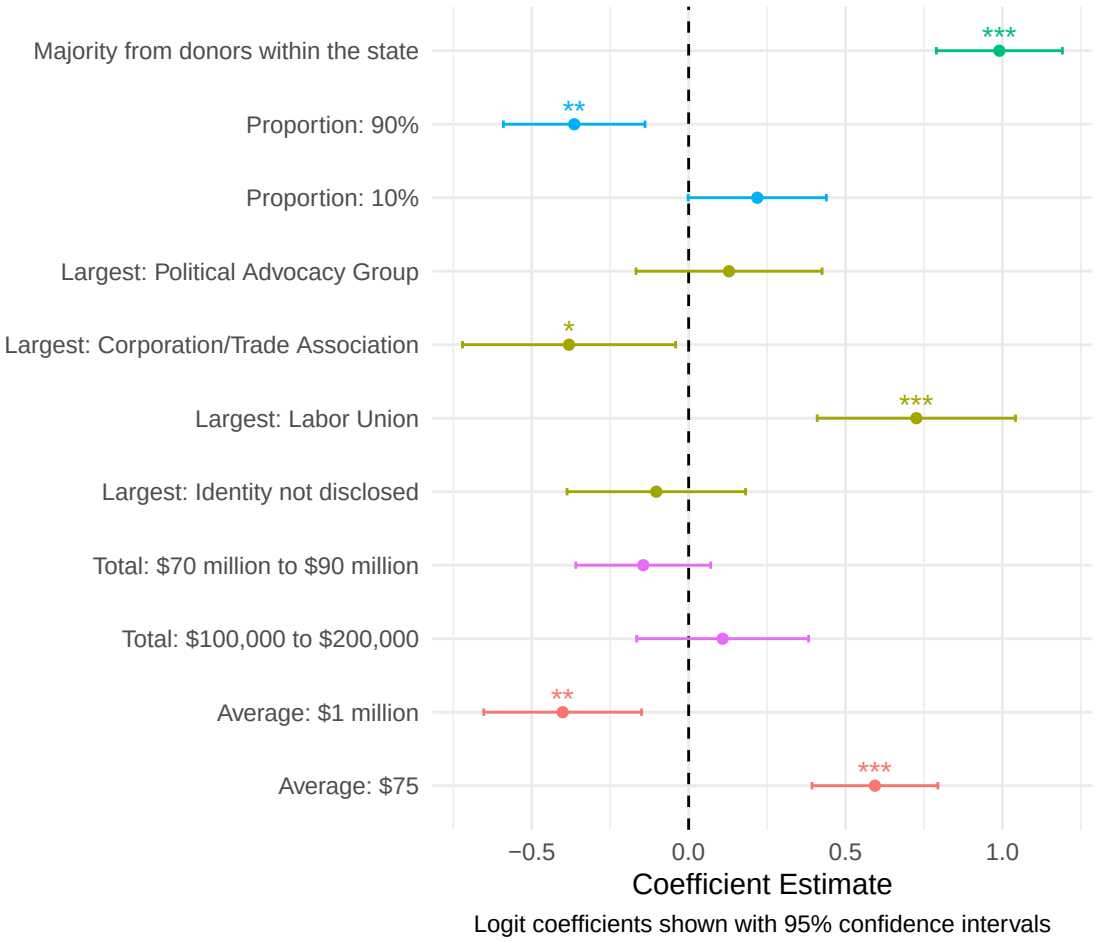


Figure 4.4: Candidate conjoint results (no additional information). Model coefficients are shown for the "control" experimental condition in which subjects only saw the funding information about the two candidates. All standard errors are clustered by participant, and the figure displays the 95 percent confidence interval around each estimate. Stars above coefficients indicate significance at $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$ respectively.

The substantively large and positive effect of the largest donor being a labor union (and likewise the negative effect of corporations) is perhaps telling of the Democratic bias in the sample. It is nevertheless noteworthy that voters adjust their vote choice when presented with this information (absent other cues). PACs did not affect vote choice which perhaps suggests that their ubiquity in the American electoral landscape has dulled their informative quality to voters.

Moreover, the "dark money" attribute level for the largest donor - "identity not disclosed"

- is insignificant (relative to individual donors). The absence of an effect for this opaque attribute level, which aims to mirror the obscurity granted to those who use non-profit corporations to hide their identity, suggests that donor anonymity does not deter individuals from voting for that campaign. In other words, there is little evidence in this experiment of electoral punishment for obscuring the source of one's donations.

Descriptively, the impact of these disclosure mechanisms are not dependent on subjects' own political identification. I calculate the difference in marginal means for the attributes given subjects' partisan affiliation (full results are reported in the appendix). There are no substantive divergences between the two groups in terms of effect direction. Only the difference between the marginal means of the PAC largest donor attribute-level is statistically significant at the 5 percent level, and then only between Democrat and Independent voters. Given the reduced sample size when considering these groups separately, however, caution should be exercised when extrapolating from these comparisons.

4.3.2 Priming subjects' perceptions of candidates' ideology and valence

Elections are rarely fought on the topic of disclosure. Indeed, the real empirical question is whether disclosure augments voter behaviour in *typical* electoral scenarios - ones in which candidate ideology, partisanship and valence characteristics are known. The first set of results demonstrate the ability of disclosure mechanisms to affect vote choice. But does disclosure have an independent effect once we control for the most relevant electoral variables?

For those subjects exposed to additional attributes, I run a similar logistic model to Equation 4.1 while also including additional parameters for the priming attributes of partisanship, ideology, and past experience. Given the inclusion of these variables, we should expect differential effects of the partisan and ideological primes dependent on subjects' own partisan affiliation. I therefore compare the revealed partisanship of the candidate to that of the subject, coding whether the partisanship is the "same" or "different". Similarly, for ideology, I project the ideological factor levels evenly between (0-10) and measure the absolute

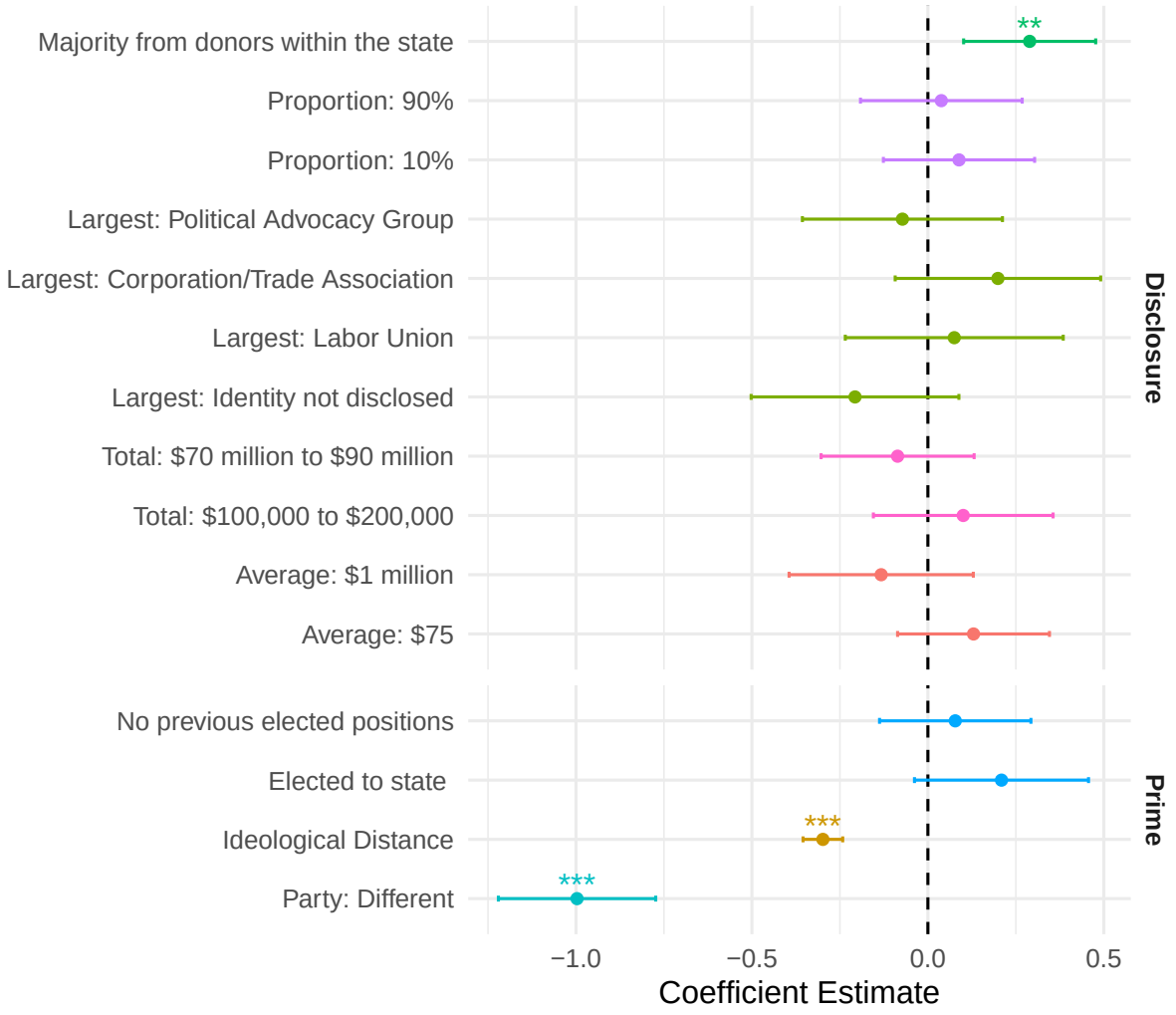


Figure 4.5: Candidate conjoint results - treatment condition in which partisan, ideological and previous experience attributes are included. Coefficient estimates for each attribute-level are shown with 95 percent confidence intervals. Stars above coefficients indicate significance at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$ respectively.

difference in ideology between candidate and subject:

$$\text{logit}(\text{Vote for candidate?}) = \sum_{i=1}^{5} \beta_i \times \text{Disclosure Mechanism}_i + \beta_6 \times \text{Party} + \beta_7 \times \text{Ideological Distance} + \beta_8 \times \text{Valence}. \quad (4.2)$$

Figure 4.5 displays the results of this model. The independent effects of all but one feature of disclosure are negligible once subjects' political priors are primed: only the positive effect of a majority of within-state donors has a significant effect on vote choice.

On the other hand, the political controls themselves have substantial effects. Voters unsurprisingly are averse to voting for candidates of a different party to themselves. Similarly, as the ideological distance between candidate and subject increases, this decreases the likelihood of voting for that candidate. The valence attribute, somewhat surprisingly, does not exhibit significant differences between the attribute levels.

In the Appendix I report two further analyses to test the conditional causal effects of disclosure based on subjects' partisanship. Broadly, I do not find that the effects of disclosure affect Democratic and Republican identifiers differently. Figure C.5 estimates separate models for Republican and Democrats in the sample.[21] These results are substantively similar to the analysis presented above. Partisan and ideological primes on the whole have large and opposing effects dependent on respondent's own partisan identification. Across all disclosure attributes, there are few substantive differences between Democrats and Republicans in terms of effect direction. Wide confidence intervals for the smaller Republican sample, however, suggest this model on its own is underpowered. Figure C.6 reports the difference in marginal means between Democrat and Republican identifying subjects. These descriptive results again show that while there are large and highly significant differences for partisan and ideological primes, all differences in marginal means for the disclosure attributes are indistinguishable from zero.

Taken together, the results of the two candidate conjoint conditions suggest that while disclosure *can* affect vote choice, it is relatively inert in more realistic contexts in which overt political cues are present. Disclosure mechanisms that were shown to affect vote choice when subjects consider funding information alone, fail to augment vote choice once respondents

[21]Running separate models for different subjects can be misleading (Leeper, Hobolt and Tilley, 2020). Comparing differences in the size of the causal effect tell us very little about preferences across subgroups. Suppose, for instance, we consider the effect of party label on Democrats and Republicans in separate models. If Republicans are already more supportive in general, the baseline level of support will also be high. Therefore, we would expect a smaller effect for Republicans compared to Democrats (given a bounded scale). The smaller effect, however, does not reflect less favourable attitudes by Republican respondents. Put simply, "similarity of conditional AMCEs therefore only convey similarity of the causal effect of the feature, but do not convey similarity of preferences unless preferences toward profiles with the reference category are equivalent across groups" (Leeper, Hobolt and Tilley, 2020, p.12).

are primed with candidates' ideological positions and partisan affiliation.

4.3.3 Disclosure's effect on initiative outcomes

Partisan and ideological cues appear to diminish the effect of aggregate disclosure considerably for candidate elections. But what about direct democratic races like initiative elections that typically lack these overt partisan and ideological cues?

Unlike the candidate conjoints where participants chose between two candidates, in the initiative conjoint participants chose whether to endorse or reject a given proposal. Table 4.3 shows the proportion of participants who voted for each proposal. The sewage bond, marijuana legalization and minimum wage increase initiatives all received greater than 70 percent support. Only the environmental protection initiative was a marginal race. Even without further analysis, therefore, it is clear that voters' opinions on these issues are relatively fixed (given that disclosure attribute levels are completely randomised).[22]

Table 4.3 also reports the mean approval rating of each side of the campaign (on a scale of 1-7), for each issue. Interestingly, there is some variation between subjects' approval of campaigns and their respective vote choice. While a majority voted against the environmental initiative, subjects were still marginally more favourable of the proponent side of the debate (albeit by a statistically insignificant amount, $p = 0.43$). And for the other three campaigns, despite the high proportions voting in favour of change, the difference in ratings are substantively closer (albeit still statistically significant) than the vote proportions would suggest.

[22]These levels of support accord with evidence at the national level that citizens overwhelmingly favour some policy changes. The Cooperative Congressional Elections Survey (CCES) fielded a set of hypothetical policy questions similar to those in this paper during the 2018 midterm election period (Ansolabehere, Schaffner and Luks, 2019). 69.9 percent of those interviewed nationwide favoured an increase in the state minimum wage to $12 an hour (question CC18_414A); a millionaire's tax to fund school and road spending was supported by 70.6 percent of budgets (question CC18_414B); policies granting the Environmental Protection Agency the power to regulate Carbon Dioxide (question CC18_415a) or requiring states to use a minimum amount of renewable fuels (question CC18_415c) both received 61.2 percent support. These questions are not perfectly comparable with those fielded in this conjoint experiment, but they do at least suggest that the levels of support are feasible and not incongruent with other surveys on citizens' policy positions.

Issue	Vote	Oppose	Support
Bond issuance	0.74	3.51	4.80
Environment tax	0.49	3.96	4.07
Marijuana legalisation	0.78	3.24	5.17
Wage increase	0.72	3.27	5.11

Table 4.3: Subject support and rating of initiative campaigns

In the remainder of this analysis, I report the results using the binary vote choice variable, although the substantive results of the model do not differ if I instead use subjects' ratings of the two campaigns. Figure 4.6 displays the results of a logistic regression very similar to Equation 4.1, except I also include issue fixed effects.

The only statistically significant disclosure attribute is the origin of donations. This attribute's effect is robust across all three conjoint experiments in this paper. Voters do seem to favour those campaigns that are funded mostly by donors within their own state. I return to the implications of this finding in the discussion.

For all other disclosure mechanisms tested in this model, there are no significant effects on vote choice (or the ratings of the two campaigns). Even in the absence of partisan cues, voters do not appear to change their opinions on both the campaigns or the issues when given information about the funding profiles of direct democratic campaigns (geographic disclosure not withstanding). It appears that voters' attitudes towards policies are reasonably fixed and that these issue preferences swamp any impact of disclosure on vote choice or individuals' assessments of campaigns.

The model reported in Figure 4.6 controls for the latent support of each issue area but assumes that disclosure affects voters the same way for each type of issue. It is plausible, however, that voters may use disclosure information to different extents when considering different issues. To check this argument I estimate separate logit models for each initiative issue separately.

Figure 4.7 plots the estimated coefficients. Large campaign finance totals and low average donations are statistically significant, positive predictors of support for the sewage bond

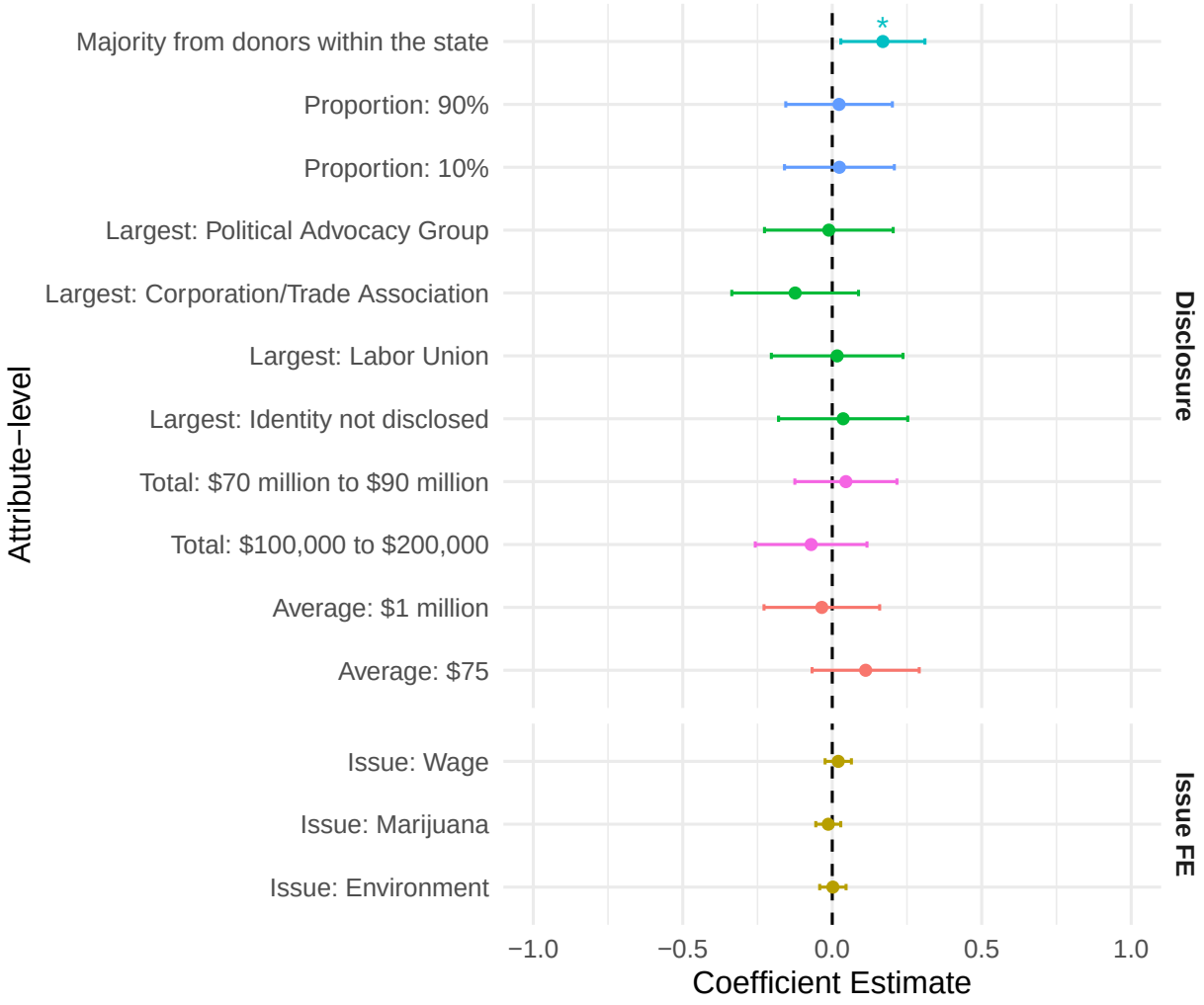


Figure 4.6: Initiative conjoint results. All responses across the four issues are pooled, and issue fixed effects are included to control for varying levels of support not contingent on disclosure mechanisms. Coefficients are shown with 95 percent confidence intervals. Stars above coefficients indicate significance at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$ respectively.

issue. No attribute level is significant for either the marijuana legalization or environmental taxation issues, and only the majority of within-state donations attribute is impactful for the minimum wage issue. These results are suggestive that for issues that are less salient (where we might expect voters' issue preferences to be less strongly held, or which covary less with voters' ideological position) that disclosure can alter vote choice. These results are purely indicative, however, and further experimental evidence is needed to test this hypothesis further.

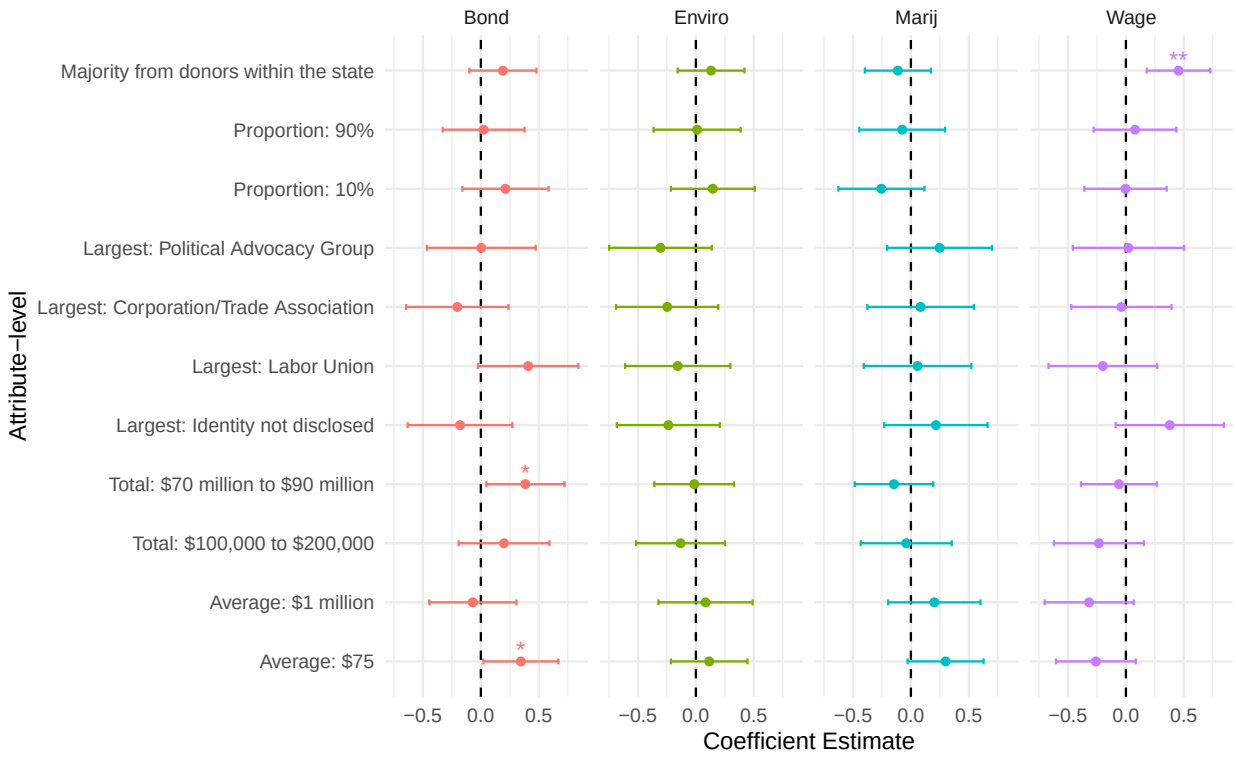


Figure 4.7: Initiative conjoint results with separate logit models run for each of the four issue areas posed to voters. All coefficients are shown with 95 percent confidence intervals. Stars above coefficients indicate significance at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$ respectively.

To ensure that the lack of differences is not driven by different effects of the baseline category across issues (Leeper, Hobolt and Tilley, 2020) I also compare the differences in marginal means for the attributes by issue. Figure C.7 in the Appendix plots these results. On the whole, differences in the marginal means of attribute levels across the four issues are insignificant. The marginal mean of an anonymous largest donor is significantly larger for the comparison between minimum wage and bond initiatives, and labor unions significantly lower for the comparisons between minimum wage and bond initiatives as well as between marijuana and bond initiatives. All other attributes are statistically indistinguishable from zero at conventional levels of significance. These comparisons provide further evidence that, at least with the power available in this study, voters do not exhibit clear differences in how they act on disclosure information across issues.

4.4 Discussion

In the absence of other relevant information, disclosure can influence vote choice. Voters seem particularly concerned about features of campaign funding profiles that indicate the extent to which campaigns are captured by narrow interests, or by out-of-state actors. Voters are less responsive to the overall size of campaigns.

Crucially, however, once voters are primed with overt partisan and ideological cues, features that are ubiquitous to candidate campaigns, the effects of disclosure all but disappear. In initiative campaigns, where voters must decide whether to endorse particular policies, disclosure is similarly ineffective at influencing vote choice. It appears that the issue prime that is necessarily a part of each initiative swamps any effect of disclosure. Moreover, while recognising the lower power of the models subsetting this data, the results do not suggest that the (null) effects of disclosure vary meaningfully across issue areas.

One exception is the positive effect of having a majority of in-state donations, which is robust across all the main models in this paper. Voters appear to be very averse to the influence of actors outside their state, even as they overlook other forms of capture by narrow interests once a candidates' partisanship and ideology are revealed. The strength of this effect, and its resilience to both partisan information and issue primes, is notable. Future work should consider precisely why this effect is so robust when other disclosure mechanisms are not.

Future work should further probe how partisan and ideological features mediate the effects of disclosure. Does partisan affiliation and ideology swamp disclosure, to the extent disclosure mechanisms are separable but merely substantively very small? Or do the effects of disclosure cue voters' beliefs about partisanship and ideology, such that priming voters renders the effects of disclosure redundant?

It is also worth noting that the conjoint attribute levels were deliberately abstract. One limitation of the conjoint method, and particularly those presenting initiative races, is that it becomes less believable if the exact same entity is both the major donor for and against

an initiative, for example. Increasing the number of unique names to decrease the likelihood of same-name pairs decreases the power of one's survey instrument (holding the number of participants fixed). Researchers may wish to develop strategies that compare the effects of disclosure when (un)recognisable names are used. Some recent evidence suggests that there are few differences between attitudes when voters are exposed to charged statements about 'secret donors' compared to a more neutral framing (Rhodes et al., 2019). That said, it would be particularly interesting to consider names with varying local salience, to see if the results presented in this paper replicate when "stronger" labels are used. This would also potentially allow researchers to understand why voters react so negatively to the effect of out-of-state donors.

From a policy perspective, the implication of these findings suggest the informational benefit of disclosure appears to be highly dependent on the wider informational context in which such information is given. In the presence of overtly political cues, across both representative and direct democratic races, disclosure does not appear to substantively impact vote choice. That is not to say that the disclosure in uninformative in a broader sense, but it does not appear to drive vote choice (as measured in this study). Voters' lack of sensitivity to the funding profile of candidates at the point of voting,[23] challenges the informational benefit logic of disclosure-based regimes. This finding is particularly concerning at a time when rates of contributions and political spending are rising.

A consequence of these findings is that campaigns (and donors) should be unperturbed by the visibility of their donations. The trade-off that Ashworth (2006) poses, between greater exposure and voters' aversion to interest group involvement, does not seem to exist in this study. Spending has consequences not only for the success of campaigns (Stratmann, 2005, 2010; Figueiredo et al., 2011; Rogers and Middleton, 2015) but also the representation of sectional interests (Gilens, 2015; Przeworski, 2015). The relatively limited effects of disclosure on voter behaviour observed in this study suggest we should think carefully about how best

[23]The effects of less immediate disclosure methods like information pamphlets may well be further attenuated (which can be empirically tested).

to curb the influence of unrepresentative interests. Disclosure alone appears insufficient.

References

Abramson, Scott F, Korhan Koçak and Asya Magazinnik. 2019. "What Do We Learn About Voter Preferences From Conjoint Experiments?".

Ansolabehere, Stephen, Brian F. Schaffner and Sam Luks. 2019. "Cooperative Congressional Election Study, 2018: Common Content.".

Arceneaux, Kevin. 2010. "The Benefits of Experimental Methods for the Study of Campaign Effects." *Political Communication* 27(2):199–215.

Ashworth, Scott. 2006. "Campaign Finance and Voter Welfare with Entrenched Incumbents." *American Political Science Review* 100(1):55–68.

Briffault, Richard. 2010. "Campaign Finance Disclosure 2.0." *Election Law Journal: Rules, Politics, and Policy* 9(4):273–303.

Brown, Adam R. 2013. "Does Money Buy Votes? The Case of Self-Financed Gubernatorial Candidates, 1998–2008." *Political Behavior* 35(1):21–41.

Buckley v. Valeo 424 U.S. 1. 1976.

Bump, Philip. 2016*a*. "Bernie Sanders Keeps Saying His Average Donation Is $27, but His Own Numbers Contradict That." https://www.washingtonpost.com/news/the-fix/wp/2016/04/18/bernie-sanders-keeps-saying-his-average-donation-is-27-but-it-really-isnt/.

Bump, Philip. 2016*b*. "Donald Trump's July Fundraising Numbers Will Reveal Just How Unusual His Campaign Actually Is - The Washington Post." https://www.washingtonpost.com/news/the-fix/wp/2016/08/02/donald-trumps-july-fundraising-numbers-will-reveal-just-how-unusual-his-campaign-actually-is/?noredirect=on&utm_term=.1d7185777d6b.

Carpini, Michael X. Delli and Scott Keeter. 1996. *What Americans Know about Politics and Why It Matters*. Yale University Press.

Citizens United v. Federal Election Commission, 558 U.S. 310. 2010.

Coate, Stephen. 2004. "Pareto-Improving Campaign Finance Policy." *The American Economic Review* 94(3):628–655.

Dafoe, Allan, Baobao Zhang and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26(04):399–416.

Dowling, Conor M. and Amber Wichowsky. 2013. "Does It Matter Who's Behind the Curtain? Anonymity in Political Advertising and the Effects of Campaign Finance Disclosure." *American Politics Research* 41(6):965–996.

Dowling, Conor M. and Amber Wichowsky. 2015. "Attacks without Consequence? Candidates, Parties, Groups, and the Changing Face of Negative Advertising: The Changing Face of Negative Advertising." *American Journal of Political Science* 59(1):19–36.

Dowling, Conor M. and Michael G. Miller. 2016. "Experimental Evidence on the Relationship Between Candidate Funding Sources and Voter Evaluations." *Journal of Experimental Political Science* 3(02):152–163.

Figueiredo, De, John M, Chang Ho Ji and Thad Kousser. 2011. "Financing Direct Democracy: Revisiting the Research on Campaign Spending and Citizen Initiatives." *The Journal of Law, Economics, and Organization* 27(3):485–514.

Finkelstein, Josh. 2018. "Grassroots Candidates Tout Small Average Contributions." https://www.opensecrets.org/news/2018/02/grassroots-candidates-tout-small-average-contributions/.

First National Bank of Boston v. Bellotti, 435 U.S. 765. 1978.

Garrett, Elizabeth and Daniel A. Smith. 2005. "Veiled Political Actors and Campaign Disclosure Laws in Direct Democracy." *Election Law Journal: Rules, Politics, and Policy* 4(4):295–328.

Gerber, Elisabeth R. 1999. *The Populist Paradox: Interest Group Influence and the Promise of Direct Legislation*. Princeton University Press.

Gigerenzer, Gerd and Wolfgang Gaissmaier. 2011. "Heuristic Decision Making." *Annual Review of Psychology* 62(1):451–482.

Gilens, Martin. 2015. "Descriptive Representation, Money, and Political Inequality in the United States." *Swiss Political Science Review* 21(2):222–228.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(01):1–30.

Jiang, Lear. 2018. Disclosure's Last Stand? The Need to Clarify the 'Informational Interest' Advanced by Campaign Finance Disclosure. SSRN Scholarly Paper ID 3262349.

La Raja, Raymond J. 2014. "Political Participation and Civic Courage: The Negative Effect of Transparency on Making Small Campaign Contributions." *Political Behavior* 36(4):753–776.

Leeper, Thomas J, Sara B Hobolt and James Tilley. 2020. "Measuring Subgroup Preferences in Conjoint Experiments." *Political Analysis* 28(2):207–221.

Lupia, Arthur. 1994. "Shortcuts Versus Encyclopedias: Information and Voting Behavior in California Insurance Reform Elections." *The American Political Science Review* 88(1):63–76.

Mayer, Jane. 2016. *Dark Money: The Hidden History of the Billionaires behind the Rise of the Radical Right*. First edition. ed. New York: Doubleday.

McIntyre v. Ohio Elections Commission, 514 U.S. 334. 1995.

Prat, Andrea, Riccardo Puglisi and James M. Snyder Jr. 2010. "Is Private Campaign Finance a Good Thing? Estimates of the Potential Informational Benefits." *Quarterly Journal of Political Science* 5(3):291–318.

Primo, David M. 2013. "Information at the Margin: Campaign Finance Disclosure Laws, Ballot Issues, and Voter Knowledge." *Election Law Journal: Rules, Politics, and Policy* 12(2):114–129.

Przeworski, Adam. 2015. Economic Inequality, Political Inequality, and Redistribution. Working Paper Department of Politics, New York University.

Rhodes, Samuel C., Michael M. Franz, Erika Franklin Fowler and Travis N. Ridout. 2019. "The Role of Dark Money Disclosure on Candidate Evaluations and Viability." *Election Law Journal: Rules, Politics, and Policy* 18(2):175–190.

Ridout, Travis N., Michael M. Franz and Erika Franklin Fowler. 2015. "Sponsorship, Disclosure, and Donors: Limiting the Impact of Outside Group Ads." *Political Research Quarterly* 68(1):154–166.

Rogers, Todd and Joel Middleton. 2015. "Are Ballot Initiative Outcomes Influenced by the Campaigns of Independent Groups? A Precinct-Randomized Field Experiment Showing That They Are." *Political Behavior* 37(3):567–593.

Sampson v. Buescher, 625 F.3d 1247. 2010.

Sances, Michael W. 2013. "Is Money in Politics Harming Trust in Government? Evidence from Two Survey Experiments." *Election Law Journal: Rules, Politics, and Policy* 12(1):53–73.

Sen, Maya. 2017. "How Political Signals Affect Public Support for Judicial Nominations: Evidence from a Conjoint Experiment." *Political Research Quarterly* 70(2):374–393.

Spencer, Douglas M and Alexander G Theodoridis. 2020. "The "Appearance of Corruption" Linking Public Opinion and Campaign Finance Reform." *Election Law Journal* 19(Forthcoming):32.

Stratmann, Thomas. 2005. "Some Talk: Money in Politics. A (Partial) Review of the Literature." *Public Choice* 124(1-2):135–156.

Stratmann, Thomas. 2010. Campaign Spending and Ballot Measures. In *Financing Referendum Campaigns*, ed. Karin Gilland Lutz and Simon Hug. London: Palgrave Macmillan UK.

Titiunik, Rocio. 2020. "Natural Experiments." *arXiv:2002.00202 [econ, stat]* .

Wood, Abby K. 2017. "Show Me the Money: 'Dark Money' and the Informational Benefit of Campaign Finance Disclosure." *SSRN Electronic Journal* .

Wood, Abby K. 2018. "Campaign Finance Disclosure." *Annual Review of Law and Social Science* 14(1):11–27.

Wood, Abby K. and Douglas M. Spencer. 2016. "In the Shadows of Sunlight: The Effects of Transparency on State Political Campaigns." *Election Law Journal: Rules, Politics, and Policy* 15(4):302–329.

5 Conclusions

In the study of representation, the motivations and behaviours of actors are only ever partially observed. Legislators rarely reveal why they did or did not act on a given policy proposal directly (and such revelations would be prone to biases). In some contexts, observations are entirely precluded because the outcome itself entails some form of non-action. For instance, when legislators are responsive to citizens preferences, there is no need to draft or vote on the corresponding ballot initiative.

How we as researchers accommodate these limitations is paramount to the validity of our claims about political phenomena. Across the three papers in this thesis, I explore different ways in which missingness influences the validity of our inferences. Chapter 2 provides a framework in which researchers can make informative claims despite unavoidable selection problems. In Chapter 3 I present applied guidance for how to handle group-level uncertainty as a function of treatment and sampling. Within political economy, but also more broadly, it is surprising how often this component of statistical estimation is neglected. Particularly in experimental contexts, where randomisation allows researchers to estimate valid causal effects, it is crucial that these results are not undermined by incorrect estimates of the corresponding uncertainty. Finally, Chapter 4 adapts conventional conjoint analyses to test whether observed causal effects are undermined by the presence of other information signals, distinguishing between the possible effects of disclosure and those within more realistic contexts.

Substantively, the results in Chapters 2 and 4 advance our understanding of the representa-

tive linkages between voters and elected officials. In Chapter 2 I demonstrate that legislators are not always willing to act on popular policy demands, and this appears to occur when support for a given policy lies outside the mainstream policy community. If a popular policy is novel, legislators deem it unwise, or it goes against the preferences of legislators' financial supporters, representatives will not necessarily act on that policy demand.

This finding raises interesting questions about how we should make sense of our study of representation. Legislators who exercise policy judgement and reach alternative conclusions to their constituents are not necessarily violating democratic norms. Policy expertise and legislative experience may temper the extent to which legislators should support policy congruent with their constituents' preferences. On the other hand, when legislators are more responsive to donors than other citizens, they undermine the fundamental norm of political equality crucial to democratic policymaking. Making sense of the various facets of representation is far beyond the scope of any single thesis, but this research suggests that we should be sensitive to how we conceptualise and study representation. Moreover, at the electoral level, Chapter 3 suggests voters themselves are unresponsive to information about the sorts of donors that condition the behaviour of legislators (once elected). In fact, I demonstrate that partisanship's influence on voters acts to crowd out any possible effects of disclosure. These results suggest federal court rulings on the efficacy of disclosure are wrong, and that voters do not discriminate between candidates on the basis of their funding profiles.

Understanding the influence of money in politics, and its impact on representation in both electoral and policymaking arenas, is paramount to our ability to safeguard democratic norms of political equality. The mechanisms that dictate how donors, parties, and other actors within policy communities interact, to determine policy equilibria, will affect the substantive representation of ordinary citizens. This thesis contributes to our understanding of these dynamics, and opens new avenues of research to explore representation and responsiveness in modern democratic systems. There is, however, still much more to be done.

Appendix A Supporting information for Chapter 2

A.1 Estimating legislator's ideal responses: robustness tests

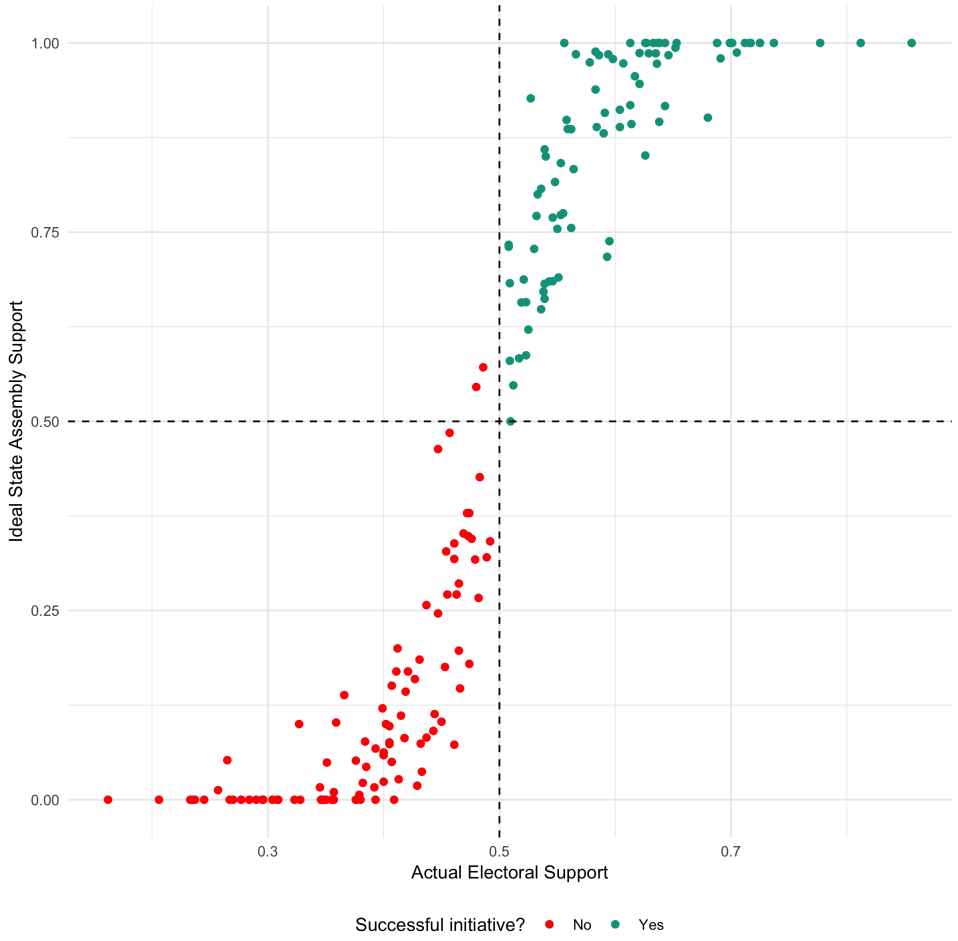


Figure A.1: Estimating the ideal legislative support where legislators *abstain* in a hypothetical policy vote if the proportion of constituents supporting an issue is between 0.475 and 0.525.

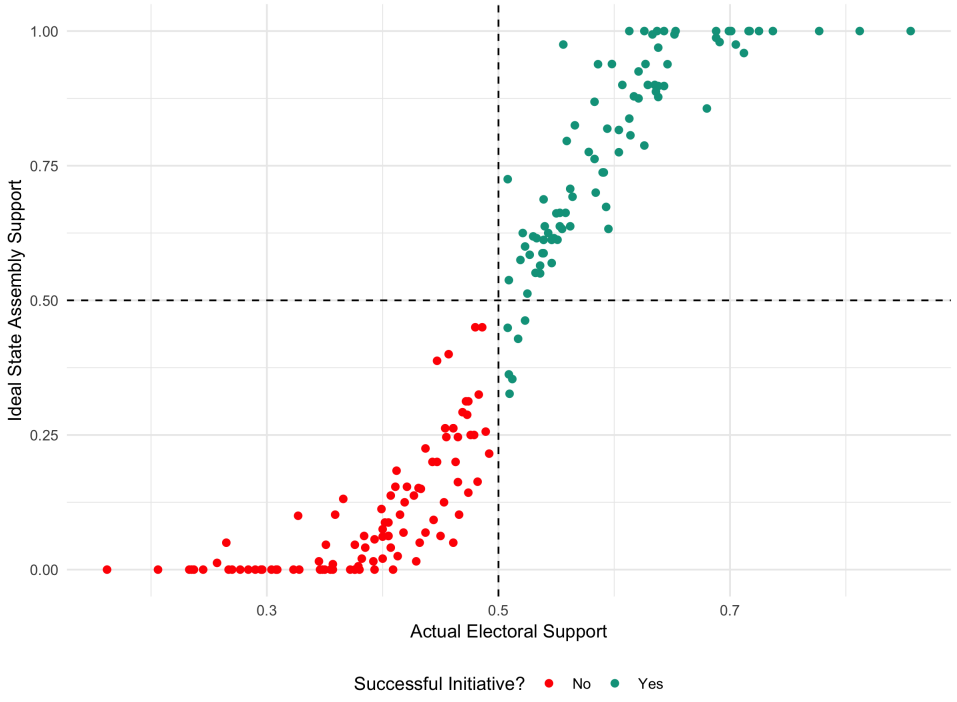


Figure A.2: Estimating the ideal legislative support where legislators hypothetically *vote against* a policy proposal if the proportion of their constituency in favour of the proposal (p) is less than 0.525.

A.2 Ideological distribution of support: further analyses

A.2.1 Proportion of extreme initiatives by year

Table A.2 displays the count of centrist and extreme initiatives by year. Except for 2000 (where there are only 2 initiatives within the sample), there are always a greater proportion of "extreme" initiatives. In all years except 2010, this proportion is above 70 percent.

A.2.2 Proportion of extreme initiatives by issue area

Another way to disaggregate initiatives is by the issue area. This is particularly useful if we believe that legislators may respond to different types of issue in different ways (holding constant levels of public support). Unfortunately, the size of the sample of initiatives makes a conclusive analysis of differences across a multiplicity of issue areas difficult. Descriptively, however, it is worth noting the policy issues touched upon by ballot initiatives. To do so, I use the National Conference of State Legislature's (NCSL; 2018) coding of initiative policy across 40 different topics. Note that the same initiative can be coded as including multiple topics. For example, an initiative raising tax levels on tobacco will be coded under both 'tax and revenue' and 'health' topics.

		Initiative	p-value	
Cycle	**Initiative**	**Donor N**	**Democratic Party**	**Republican Party**
AZ2004	200	550	0.000***	0.000***
AZ2006	201	302	0.000***	0.000***
AZ2006	203	304	0.000***	0.000***
AZ2006	204	1737	0.000***	0.000***
AZ2006	207	101	0.000***	0.000***
AZ2010	203	130	1.000	0.000***
CA2000	35	4657	0.000***	0.000***
CA2000	39	658	0.000***	0.000***
CA2002	49	636	0.000***	0.000***
CA2002	50	266	0.000***	0.000***
CA2004	64	551	0.000***	0.001***
CA2004	63	259	0.000***	0.000***
CA2004	71	252	0.234	0.000***
CA2006	83	158	0.000***	0.000***
CA2006	84	157	0.001***	0.000***
CA2008	8	6944	0.000***	0.000***
CA2008	11	661	0.000***	0.007**
CA2008	2	6183	0.000***	0.000***
CA2010	25	75	0.000***	0.000***
CA2010	22	1050	0.000***	0.000***
CA2010	20	41	0.000***	0.482
CA2010	26	107	0.000***	0.038*
CA2012	30	2157	0.000***	0.000***
CA2012	35	364	0.000***	0.000***
CA2012	36	47	0.000***	0.000***
CA2012	39	350	0.000***	0.000***
MA2008	3	2729	0.000***	0.000***
MA2008	2	372	0.000***	0.000***
MA2010	1	266	0.000***	0.000***
MA2012	3	100	0.000***	0.000***
OR2002	25	52	0.000***	0.000***
OR2004	36	2614	0.000***	0.000***
OR2004	37	184	0.000***	0.000***
OR2006	44	846	0.000***	0.000***
OR2012	79	81	0.000***	0.000***
WA2004	297	118	0.000***	0.000***
WA2004	872	213	0.000***	0.165
WA2006	937	1404	0.000***	0.000***
WA2008	1000	8865	0.000***	0.000***
WA2010	1053	816	0.000***	0.000***
WA2012	502	620	0.000***	0.000***
WA2012	1240	194	0.000***	0.000***
WA2012	1185	667	0.000***	0.000***

***=$p < 0.001$,**=$p < 0.01$,*=$p < 0.05$

Table A.1: Mann-Whitney U-tests for initiative campaigns and state legislative parties.

Year	Centrist	Extreme	N
2000	0.50	0.50	2
2002	0.25	0.75	4
2004	0.20	0.80	20
2006	0.27	0.73	15
2008	0.17	0.83	12
2010	0.36	0.64	11
2012	0.15	0.85	13

Table A.2: Cross-tabulation of relative initiative positions by year.

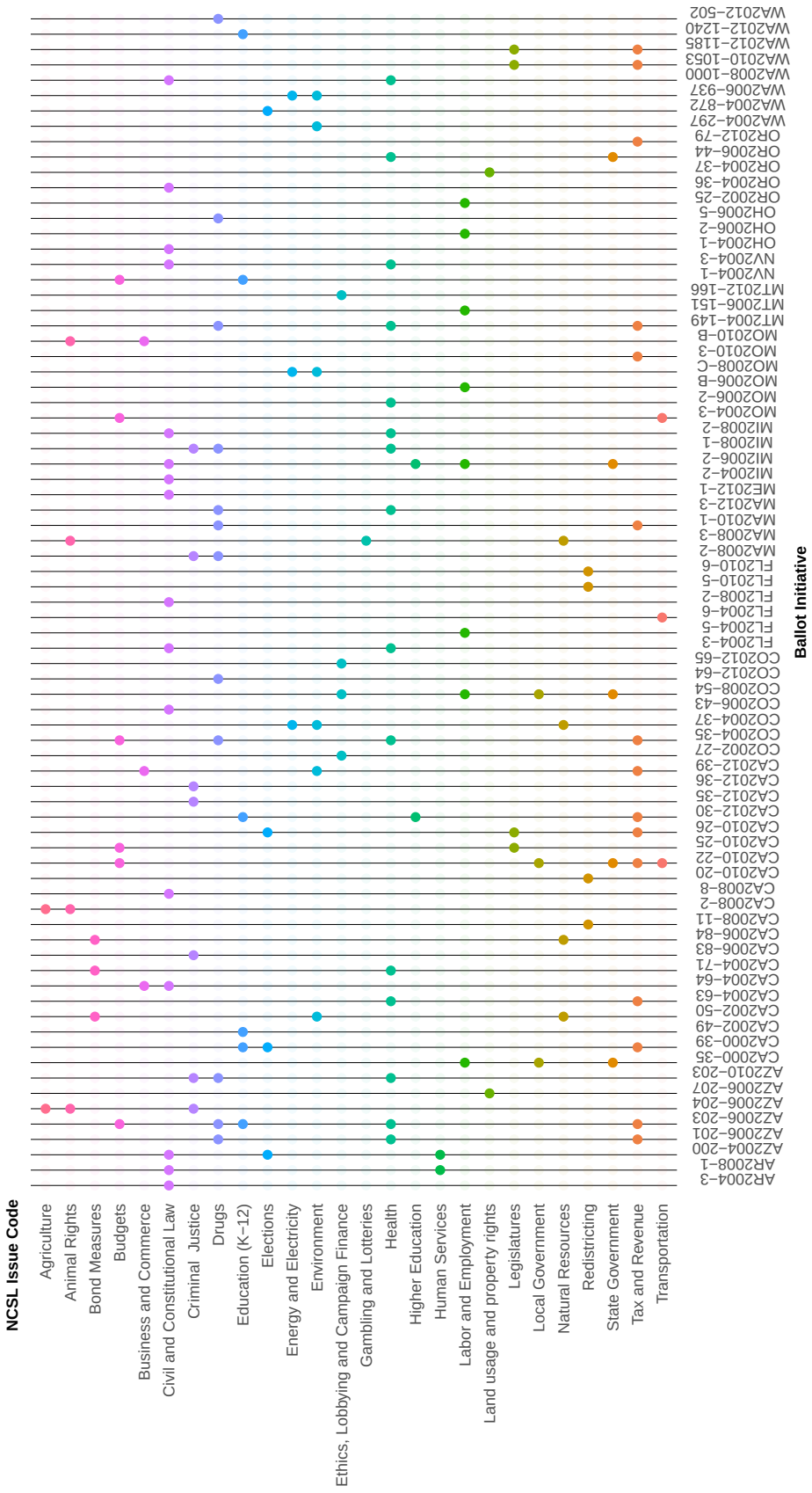


Figure A.3: The (multiple) policy topics covered by ballot initiatives in the donor data analysis

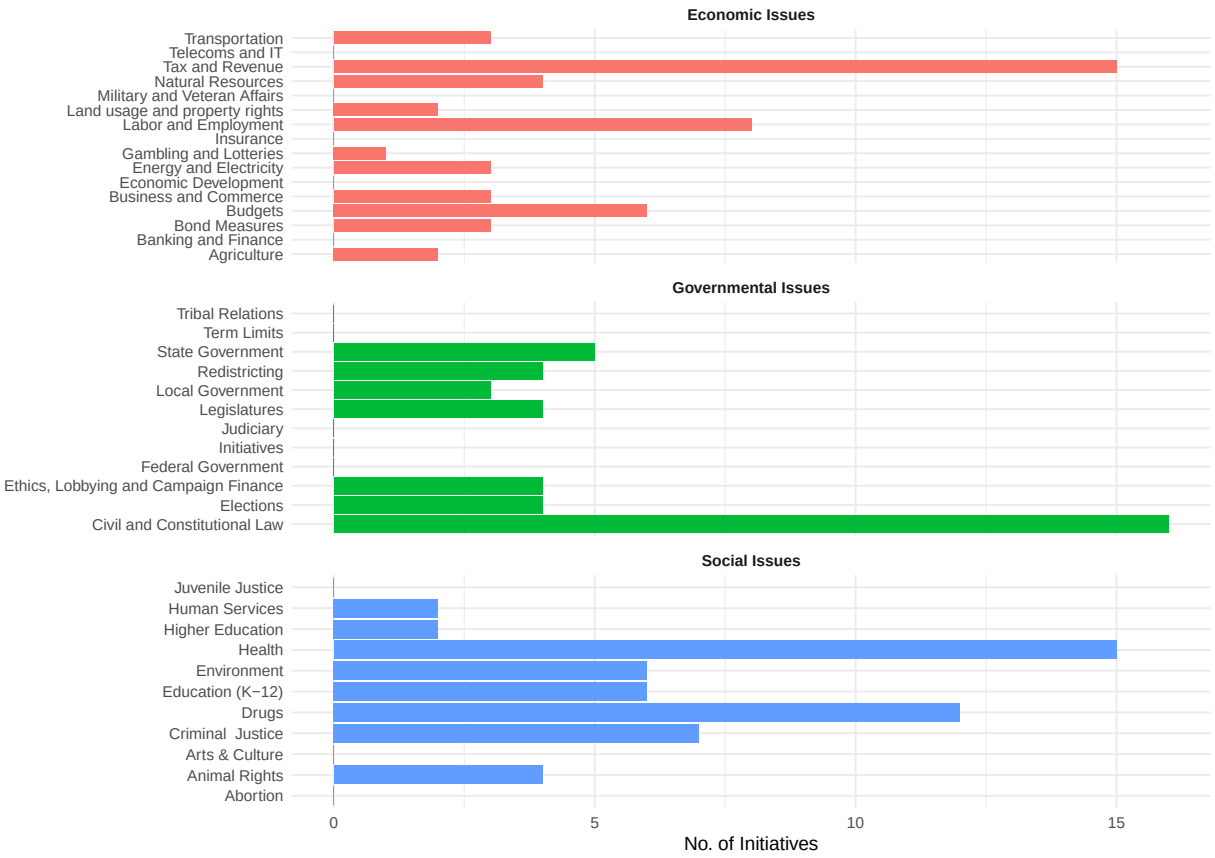


Figure A.4: Total count of each policy topic across all initiatives considered in the donor data analysis

Figure A.3 plots which topics each initiative is coded under in the NCSL database. Note that most topics are coded as including both economic and social topics. Clearly, initiative policymaking is multi-dimensional despite the singular focus of initiative legislation.

I further aggregate the individual NCSL topics into social, economic, and governmental[1] categories. Figure A.4 plots the total counts of each category over the 77 initiatives in the donor data sample. I also include those topics not covered by any initiative to show which issues were not legislated using direct democratic policymaking tools.

The modal topic across the entire list is 'civil and constitutional law' – a very broad category. Other notably frequent topics of initiatives include taxation, as well as health, drugs

[1]Governmental pertains to policies that regulate how state politics is conducted.

policy, and labor and employment law. More generally, while there are fewer 'governmental' topics covered, it is not the case that successful initiatives predominantly focus on either social or economic issues. Of the 77 initiatives considered, 53 percent cover social topics, 52 percent cover economic topics, and 43 percent cover governmental topics.

	Relative position		
	Left	Centre	Right
Social initiatives	0.512	0.293	0.195
Economic initiatives	0.45	0.275	0.275
Governmental initiatives	0.273	0.182	0.545

Table A.3: Cross-tabulation of relative initiative positions by issue category.

Do the distribution of ideological positions differ between these three aggregate categories? Table A.3 reports the proportion of left, centre, and right positioned initiatives (relative to parties) by category, using the estimated positions in the main text of this paper.[2] Across all three categories there is a clear tendency for initiatives to be extreme. A majority of initiatives covering social issues are left-extreme, and only about 20 percent have support that is on average more conservative than the Republican party. Similarly, for economic initiatives, a clear plurality of initiatives are left-extreme. Interestingly, a majority of governmental issues are right-extreme. The distribution of positions for these initiatives is almost the complete opposite of that for social and economic initiatives. This difference merits further study.

[2]Note that individual initiatives may be included in multiple counts where they cover more than one topic.

A.3 Robustness tests: stringency of exclusion criteria

Table A.4 demonstrates the robustness of the main paper's findings on ideological distributions, by excluding infrequent donors from the analysis. Overall, the results are substantively unchanged across the increasingly strict exclusion criteria. Only four initiative's support base change positions as the stringency of exclusion criteria increases (highlighted in bold). Proposition 35 (2000) moves from a Centre to Right position and then back again; Propositions 11 (2008) and 26 (2010) in California move from Centre to Right positions; and Ohio Issue 5 (2006) switches from Centre to Left position. These are all cases where the support distribution was already very close to the party mean, and do not alter the substantive findings of this analysis.

Table A.4: Relative position of initiatives when controlling for number of unique donations per contributor

		$n \geq 2$			$n \geq 4$			$n \geq 8$		
Cycle	Initiative	N_I	Position	Distance (σ)	N_I	Position	Distance (σ)	N_I	Position	Distance (σ)
AR2004	3	147	Right	1.82	80	Right	1.81	56	Right	2.01
AR2008	1	147	Right	1.40	69	Right	1.43			
AZ2004	200	243	Right	1.84	119	Right	1.48	69	Right	1.43
AZ2006	201	159	Left	0.34	95	Left	0.26	60	Left	0.57
AZ2006	203	175	Centre	0.38	145	Centre	0.27	106	Centre	0.14
AZ2006	204	425	Left	0.72	181	Left	0.64	85	Left	0.88
AZ2006	207	75	Right	1.24	60	Right	1.22	45	Right	1.19
AZ2010	203	68	Left	0.09	48	Left	0.17			
CA2000	35	1405	Right	0.01	652	**Centre**	0.01	338	Right	0.04
CA2000	39	380	Centre	0.48	307	Centre	0.43	245	Centre	0.43
CA2002	49	423	Right	0.09	333	Right	0.08	278	Right	0.17
CA2002	50	195	Centre	0.35	144	Centre	0.36	99	Centre	0.32
CA2004	63	190	Left	0.99	95	Left	0.80	59	Left	0.69
CA2004	64	260	Centre	0.09	187	Centre	0.07	146	Centre	0.07
CA2004	71	198	Centre	0.11	160	Centre	0.10	135	Centre	0.08
CA2006	83	109	Right	0.76	87	Right	0.76	63	Right	0.83
CA2006	84	122	Centre	0.18	96	Centre	0.19	78	Centre	0.25
CA2008	11	572	Centre	0.19	500	Centre	0.15	421	**Right**	0.00
CA2008	2	5229	Left	1.37	1682	Left	1.02	831	Left	1.01
CA2008	8	3811	Right	2.61	1301	Right	2.13	603	Right	1.97
CA2010	22	548	Centre	0.44	345	Centre	0.36	221	Centre	0.26
CA2010	25	70	Left	0.58	59	Left	0.68	47	Left	0.81
CA2010	26	94	Centre	0.06	80	Centre	0.03	69	**Right**	0.01
CA2012	30	1503	Left	0.87	889	Left	0.62	641	Left	0.60
CA2012	35	293	Left	0.99	100	Left	0.54	73	Left	0.58
CA2012	36	43	Left	0.96						
CA2012	39	300	Left	2.33	142	Left	1.90	74	Left	1.75
CO2002	27	189	Left	0.74	148	Left	0.76	115	Left	0.80
CO2004	35	307	Centre	0.23	231	Centre	0.16	162	Centre	0.10
CO2004	37	469	Left	0.51	333	Left	0.51	228	Left	0.56
CO2006	43	237	Right	1.19	127	Right	1.23	82	Right	1.39
CO2008	54	49	Right	0.78						
CO2012	64	692	Centre	0.45	429	Centre	0.33	276	Centre	0.17

Continued on next page

Table A.4 – *Continued from previous page*

		$n \geq 2$			$n \geq 4$			$n \geq 8$		
Cycle	Initiative	N_I	Position	Distance (σ)	N_I	Position	Distance (σ)	N_I	Position	Distance (σ)
CO2012	65	264	Left	0.54	229	Left	0.57	191	Left	0.61
FL2004	3	1734	Right	0.67	1057	Right	0.57	600	Right	0.63
FL2004	5	58	Left	2.19	51	Left	2.32	47	Left	2.52
FL2004	6	2272	Right	1.62	1057	Right	1.70	461	Right	1.80
FL2008	2	764	Right	2.25	341	Right	2.16	168	Right	2.43
FL2010	5	3512	Left	2.17	2987	Left	2.27	2288	Left	2.48
FL2010	6	3512	Left	2.17	2987	Left	2.27	2288	Left	2.48
MA2008	2	265	Left	0.37	206	Left	0.31	146	Left	0.38
MA2008	3	649	Left	0.52	270	Left	0.67	111	Left	0.66
MA2010	1	154	Centre	0.68						
MA2012	3	44	Left	0.49						
ME2012	1	5866	Left	0.59	3969	Left	0.52	2330	Left	0.49
MI2004	2	454	Right	1.20	210	Right	1.18	104	Right	1.25
MI2006	2	921	Right	0.90	670	Right	0.92	430	Right	1.04
MI2008	1	364	Left	0.36	238	Left	0.50	150	Left	0.67
MI2008	2	255	Left	0.18	216	Left	0.25	176	Left	0.32
MO2004	3	180	Centre	0.70	142	Centre	0.64	109	Centre	0.56
MO2006	2	359	Left	0.08	221	Left	0.09	142	Left	0.00
MO2006	B	80	Left	0.53	76	Left	0.56	74	Left	0.60
MO2008	C	94	Left	0.94	76	Left	0.98	56	Left	1.11
MO2010	B	318	Left	0.43	218	Left	0.41	145	Left	0.36
MT2004	149	64	Left	1.76	48	Left	0.76	40	Left	0.21
MT2006	151	72	Left	0.28	69	Left	0.29	66	Left	0.32
MT2012	166	63	Left	0.49	50	Left	0.55	40	Left	0.58
NV2004	1	67	Right	0.42	58	Right	0.47	46	Right	0.58
NV2004	3	721	Right	0.77	468	Right	0.88	269	Right	0.96
OH2004	1	121	Right	1.48	65	Right	1.43			
OH2006	2	65	Left	0.70	61	Left	0.71	58	Left	0.76
OH2006	5	334	Centre	0.03	183	**Left**	0.14	97	**Left**	0.42
OR2002	25	47	Left	0.70						
OR2004	36	579	Right	2.11	228	Right	2.15	96	Right	2.15
OR2004	37	102	Right	1.34	72	Right	1.42	52	Right	1.64
OR2006	44	273	Left	4.53	52	Left	1.73			
OR2012	79	40	Right	1.72						
WA2004	297	94	Left	1.07	86	Left	1.10	65	Left	1.14

Continued on next page

Table A.4 – *Continued from previous page*

		$n \geq 2$			$n \geq 4$			$n \geq 8$		
Cycle	Initiative	N_I	Position	Distance (σ)	N_I	Position	Distance (σ)	N_I	Position	Distance (σ)
WA2004	872	113	Centre	0.09	81	Centre	0.12	58	Centre	0.02
WA2006	937	1071	Left	0.94	839	Left	0.96	606	Left	1.01
WA2008	1000	5905	Left	0.70	4087	Left	0.73	2520	Left	0.80
WA2010	1053	725	Right	0.99	581	Right	1.03	411	Right	1.07
WA2012	1185	593	Right	1.38	445	Right	1.39	298	Right	1.51
WA2012	1240	147	Centre	0.21	112	Centre	0.14	81	Centre	0.08
WA2012	502	455	Left	0.57	341	Left	0.57	232	Left	0.60

A.4 Robustness tests: Bootstrapped difference-in-means estimates

To compute the bootstrapped difference-in-means for an initiative's distribution of support (of size n) and a party's distribution of support (of size m), I merge the two groups of support into a single cumulative distribution. From this combined distribution, I take two random samples *with replacement* of sizes n and m and calculate the means of both sampled groups. I then subtract one from the other to produce a hypothetical difference-in-means estimate ($\hat{x}$). By repeating this process 10,000 times per pairing, I generate a distribution of hypothetical difference-in-means. I then rank these calculated differences and assess the position of the *actual* difference-in-means ($\hat{x}*$) with respect to this distribution. Doing so enables me to test how outlying the actual, observed difference is from the distribution of *potential* differences. Hence, this bootstrap procedure allows me to estimate a p-value for the difference-in-means, calculated as:

$$p = \frac{\sum_i^{10000}(|\hat{x}^*| > \hat{x}_i)}{10000} \tag{A.1}$$

Table A.5 displays the results of this procedure. The results are substantively unchanged compared to the results in the main text. The four initiatives that fail to reach statistical significance in the Mann-Whitney tests also have statistically insignificant differences between their mean support and that of the closest party. All other initiatives (when projected donors are included) exhibit statistically significant difference-in-means to both parties, even when the absolute distance between the two groups is relatively small.

As the stringency of donor exclusion criteria increases, fewer initiatives reach statistical significance. For instance, when the number of distinct donations required is $n \geq 2$, we cannot reject the null hypothesis that mean support for Proposition 64 (California 2004) is different from the mean of successful Republican legislators. In general, those initiatives which become insignificant are those closest to the party means. Even when the number of

distinct donations required is ≥ 8, however, the vast majority of initiatives are statistically distinguishable from their legislative counterparts.

Min. donations		$n \geq 1$		$n \geq 2$		$n \geq 4$		$n \geq 8$	
Cycle	**No.**	**Dem. Diff.**	**Rep. Diff.**	**Dem. Diff.**	**Rep. Diff.**	**Dem. Diff.**	**Rep. Diff.**	**Dem. Diff.**	**Rep. Diff.**
AZ2004	200	2.047***	1.289***	1.585***	0.919***	1.354***	0.74***	1.262***	0.705***
AZ2006	201	-0.335***	-1.185***	-0.247***	-1.009***	-0.192*	-0.902***	-0.411***	-1.057***
AZ2006	203	0.557***	-0.293***	0.277***	-0.486***	0.2**	-0.51***	0.1	-0.547***
AZ2006	204	-0.728***	-1.578***	-0.528***	-1.29***	-0.465***	-1.175***	-0.636***	-1.283***
AZ2006	207	1.587***	0.737***	1.394***	0.632***	1.325***	0.615***	1.231***	0.585***
AZ2010	203	0.01	-1.168***	-0.068	-1.127***	-0.127	-1.082***		
CA2000	35	0.838***	0.04***	0.726***	0.004	0.656***	-0.007	0.623***	0.023
CA2000	39	0.213***	-0.585***	0.24***	-0.482***	0.218***	-0.446***	0.215***	-0.385***
CA2002	49	0.952***	0.054**	0.894***	0.05	0.835***	0.041	0.806***	0.09**
CA2002	50	0.159***	-0.738***	0.185***	-0.659***	0.189	-0.605***	0.168**	-0.548***
CA2004	63	-0.538***	-1.435***	-0.519***	-1.355***	-0.419***	-1.208***	-0.361***	-1.076***
CA2004	64	0.836***	-0.061*	0.79***	-0.047	0.75***	-0.038	0.678***	-0.037
CA2004	71	0.05	-0.847***	0.056	-0.78***	0.055	-0.734***	0.039	-0.676***
CA2006	83	1.245***	0.424***	1.167***	0.393***	1.114***	0.389***	1.073***	0.416***
CA2006	84	0.097*	-0.724***	0.096*	-0.678***	0.103	-0.621***	0.133*	-0.523***
CA2008	11	0.742***	-0.13***	0.702***	-0.115***	0.675***	-0.089**	0.691***	0.001
CA2008	2	-0.734***	-1.606***	-0.737***	-1.554***	-0.553***	-1.317***	-0.536***	-1.226***
CA2008	8	2.524***	1.652***	2.377***	1.56***	2.025***	1.261***	1.813***	1.123***
CA2010	20	0.884***	0.012						
CA2010	22	0.259***	-0.613***	0.236***	-0.578***	0.193***	-0.557***	0.135	-0.525***
CA2010	25	-0.279***	-1.151***	-0.309***	-1.123***	-0.361***	-1.111***	-0.423***	-1.082***
CA2010	26	0.814***	-0.058	0.779***	-0.035	0.734***	-0.015	0.665***	0.005
CA2012	30	-0.53***	-1.322***	-0.456***	-1.196***	-0.324***	-1.005***	-0.309***	-0.918***
CA2012	35	-0.512***	-1.304***	-0.521***	-1.26***	-0.284***	-0.965***	-0.298***	-0.907***
CA2012	36	-0.478***	-1.27***	-0.504***	-1.244***				
CA2012	39	-1.208***	-1.999***	-1.22***	-1.959***	-0.999***	-1.681***	-0.903***	-1.512***
MA2008	2	-0.129***	-1.101***	-0.139***	-1.014***	-0.118***	-0.931***	-0.139***	-0.872***
MA2008	3	-0.379***	-1.351***	-0.195***	-1.07***	-0.25***	-1.064***	-0.238***	-0.971***
MA2010	1	0.633***	-0.469***	0.578***	-0.417***				
MA2012	3	-0.101**	-1.206***	-0.182**	-1.172***				
OR2002	25	-0.321***	-1.152***	-0.353***	-1.117***				
OR2004	36	1.939***	1.15***	1.653***	0.925***	1.588***	0.917***	1.482***	0.884***
OR2004	37	1.473***	0.685***	1.317***	0.589***	1.278***	0.608***	1.269***	0.672***
OR2006	44	-2.612***	-3.241***	-2.25***	-2.83***	-0.857***	-1.402***		
OR2012	79	1.57***	0.85***	1.362***	0.688***				
WA2004	297	-0.651***	-1.797***	-0.593***	-1.688***	-0.621***	-1.651***	-0.643***	-1.563***
WA2004	872	1.157***	0.011	1.041***	-0.053	0.959***	-0.07	0.91***	-0.01
WA2006	937	-0.551***	-1.632***	-0.548***	-1.554***	-0.561***	-1.497***	-0.585***	-1.414***
WA2008	1000	-0.429***	-1.568***	-0.407***	-1.505***	-0.425***	-1.455***	-0.46***	-1.368***
WA2010	1053	1.92***	0.544***	1.879***	0.597***	1.808***	0.612***	1.685***	0.621***
WA2012	1185	2.285***	0.859***	2.204***	0.871***	2.102***	0.864***	1.985***	0.9***
WA2012	1240	1.056***	-0.371***	1.204***	-0.13*	1.15***	-0.088	1.035***	-0.05
WA2012	502	-0.303***	-1.73***	-0.308***	-1.642***	-0.317***	-1.556***	-0.341***	-1.426***

$*** = p < 0.001$, $** = p < 0.01$, $* = p < 0.05$. All initiatives have $n \geq 40$; blank entries indicate initiatives that drop below this threshold.

Table A.5: Non-parametric bootstrapped difference in means test, by number of distinct donations.

A.5 Kernel density plots

This section includes the complete set of kernel density plots for each ballot initiative per state-legislative cycle. Blue-shaded distributions depict the ideological distribution of donors to the Democratic Party. Red-shaded distributions depict the same but for the Republican party. The black line draws the distribution of ideological support for the specific initiative in question.

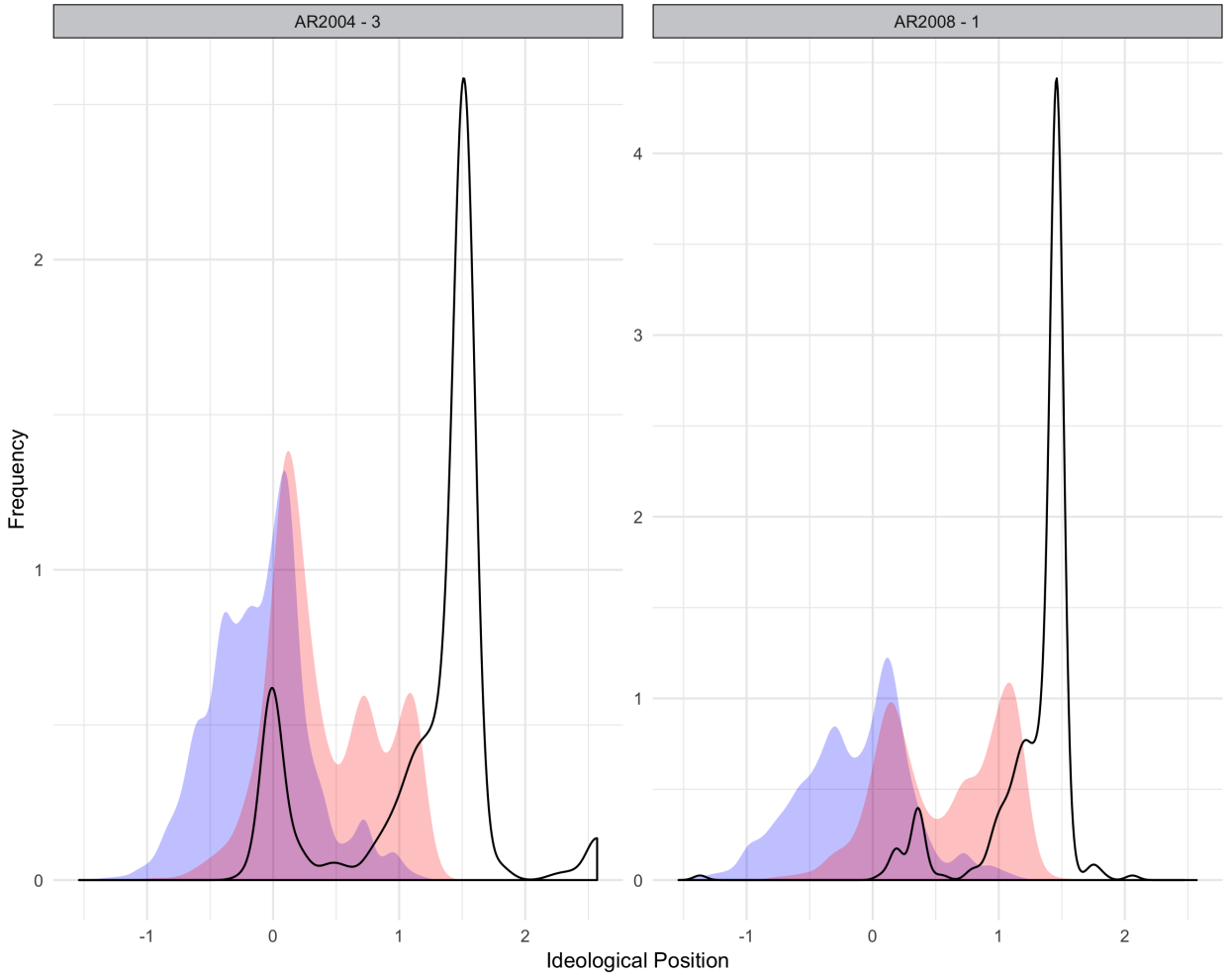


Figure A.5: Arkansas kernel density plots of initiative and legislative party donors' ideology

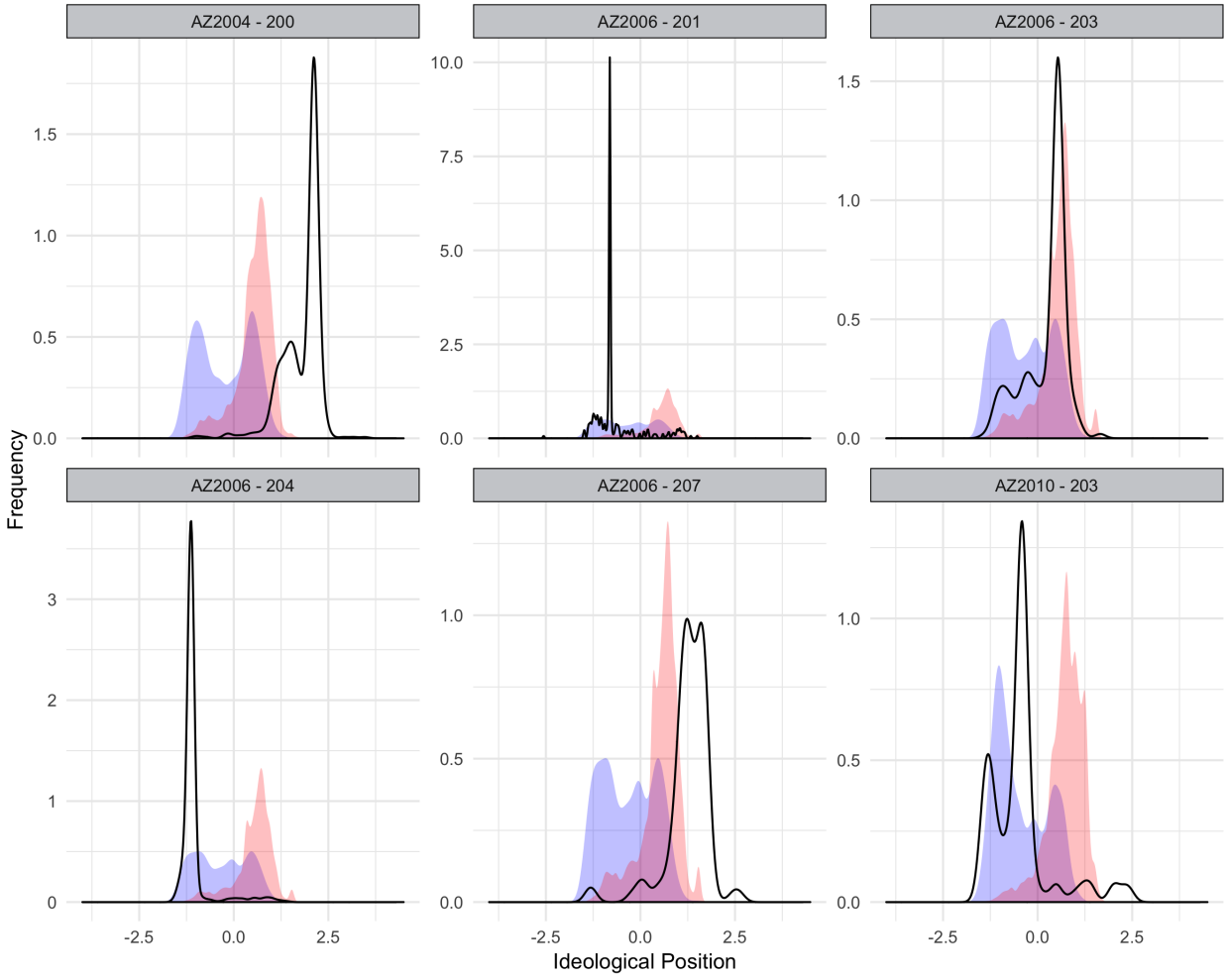


Figure A.6: Arizona kernel density plots of initiative and legislative party donors' ideology

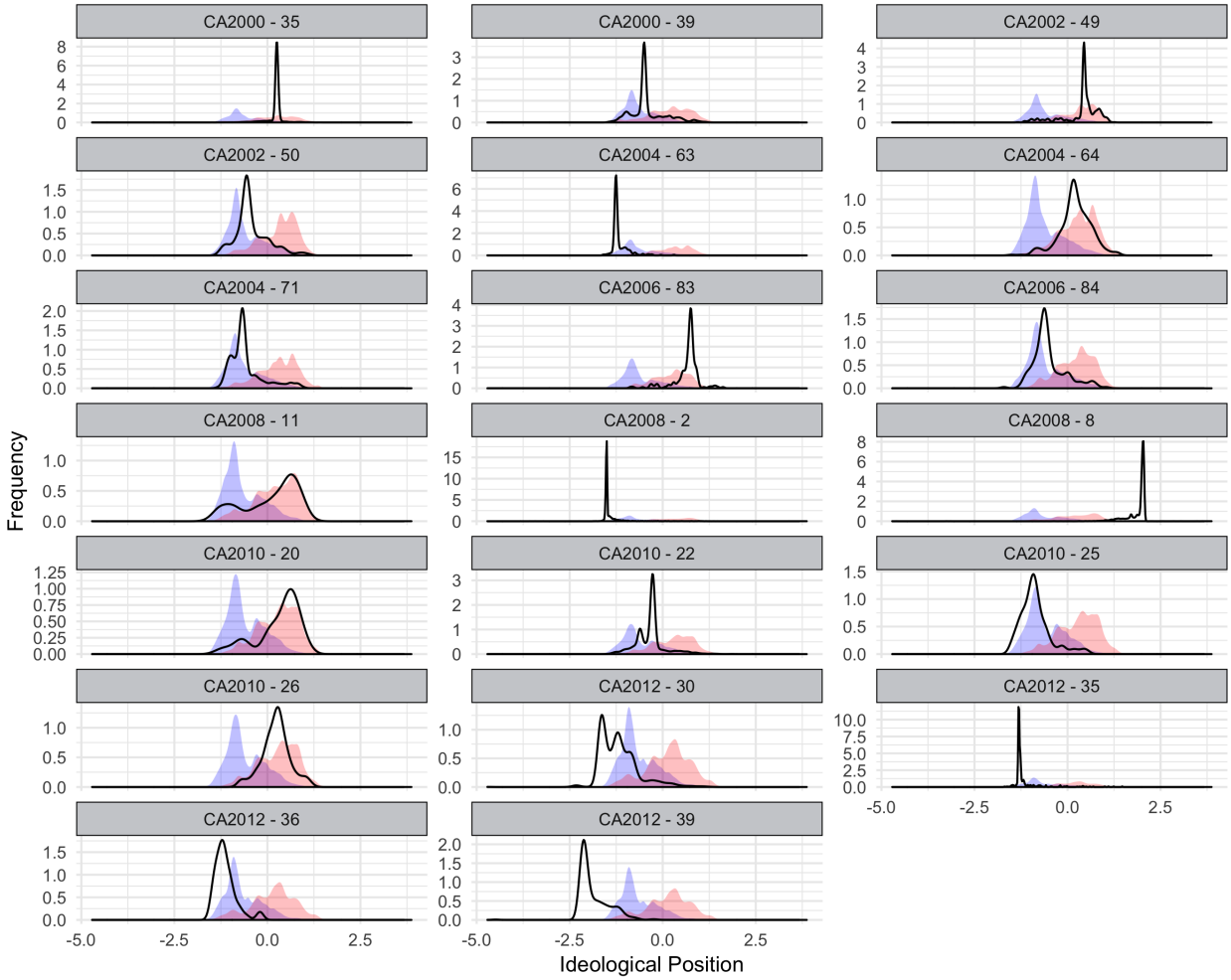


Figure A.7: California kernel density plots of initiative and legislative party donors' ideology

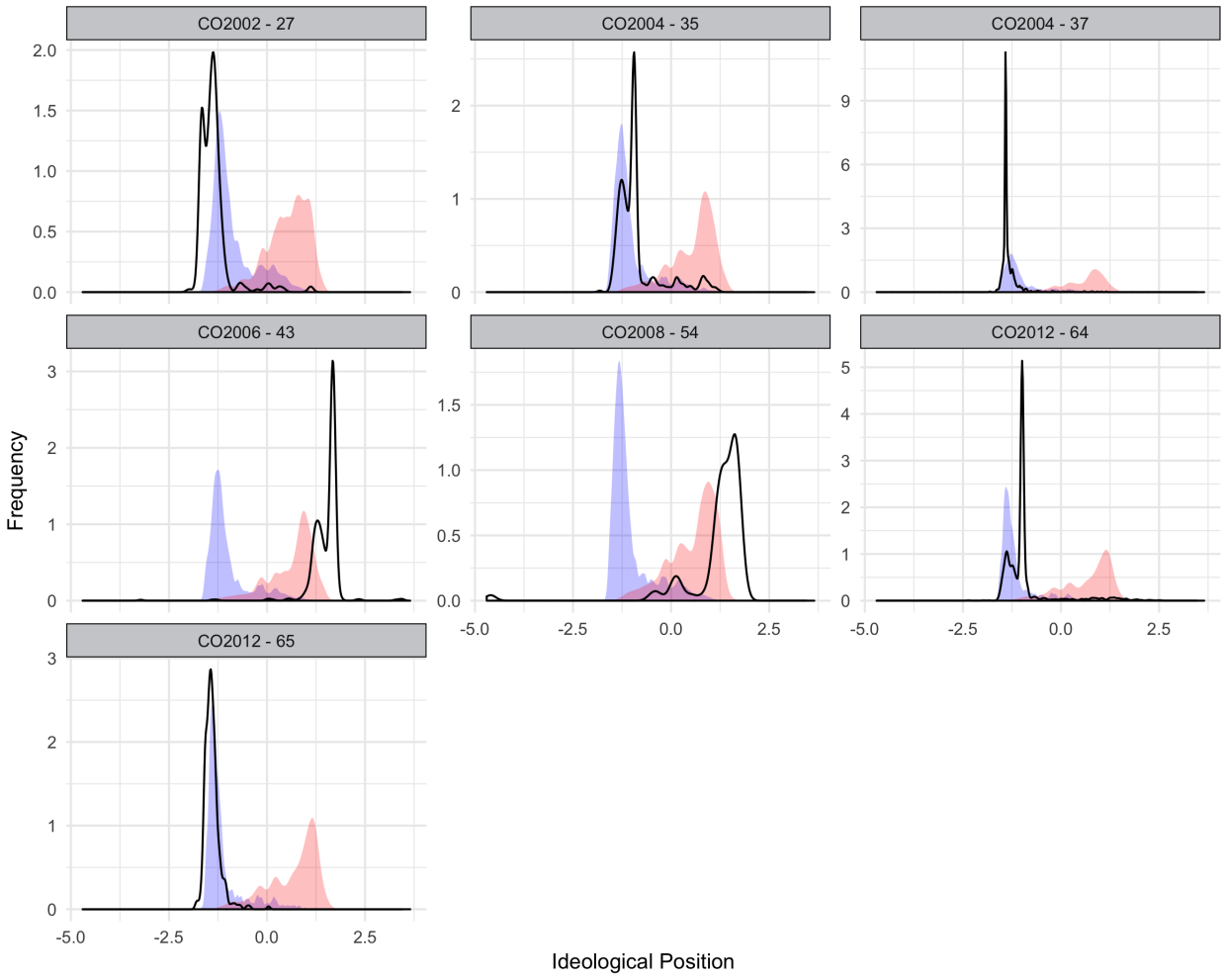


Figure A.8: Colorado kernel density plots of initiative and legislative party donors' ideology

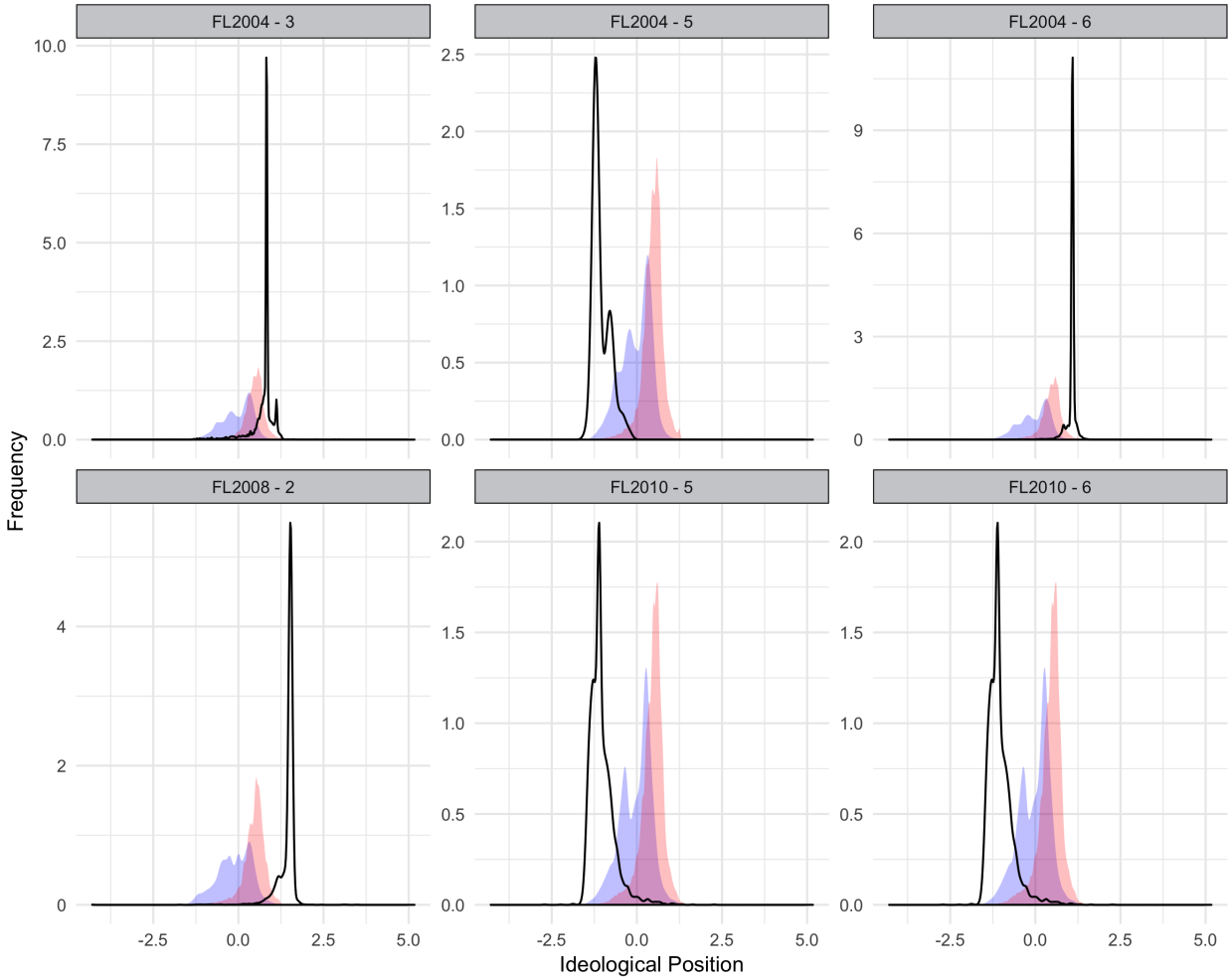


Figure A.9: Florida kernel density plots of initiative and legislative party donors' ideology

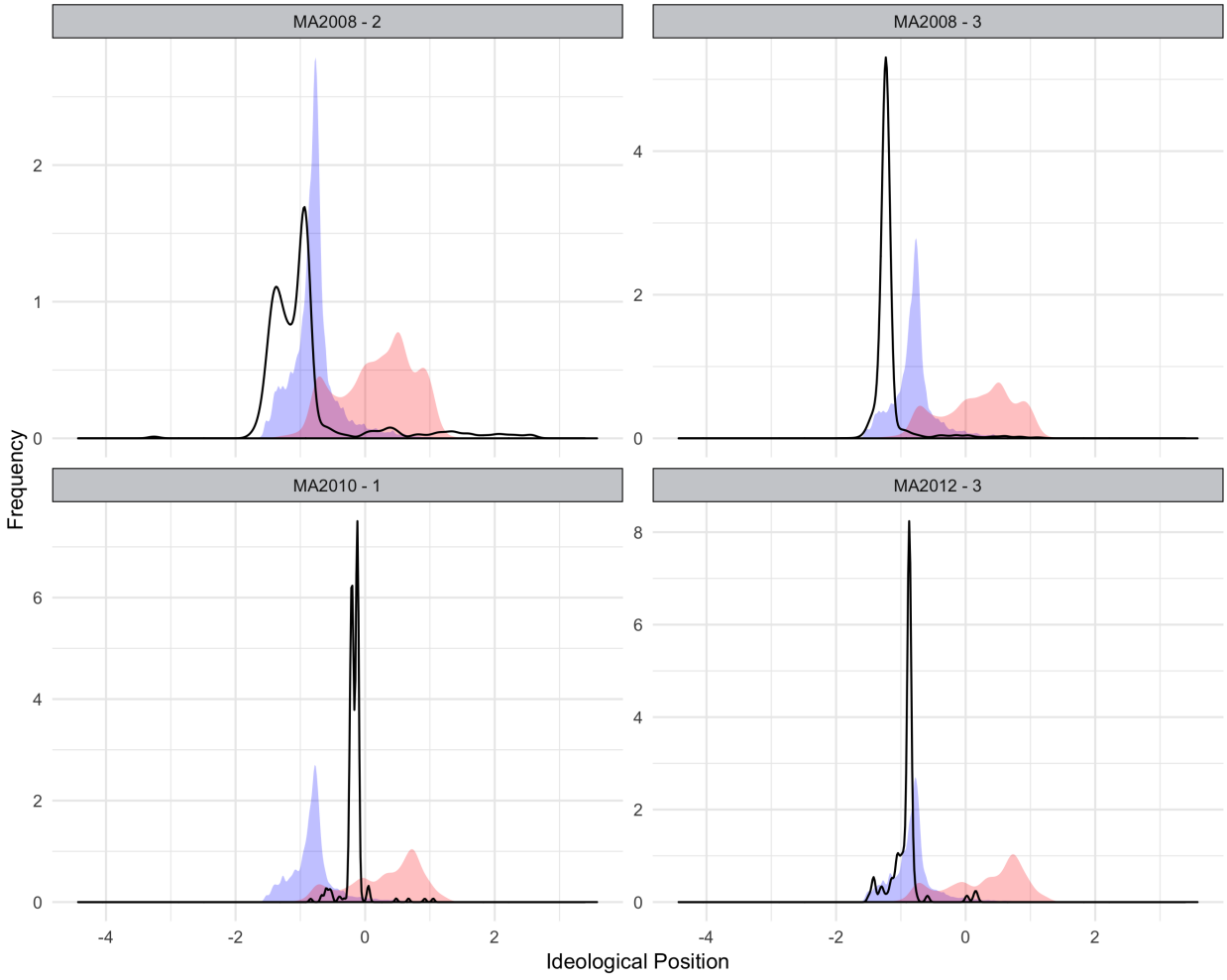


Figure A.10: Massachusetts kernel density plots of initiative and legislative party donors' ideology

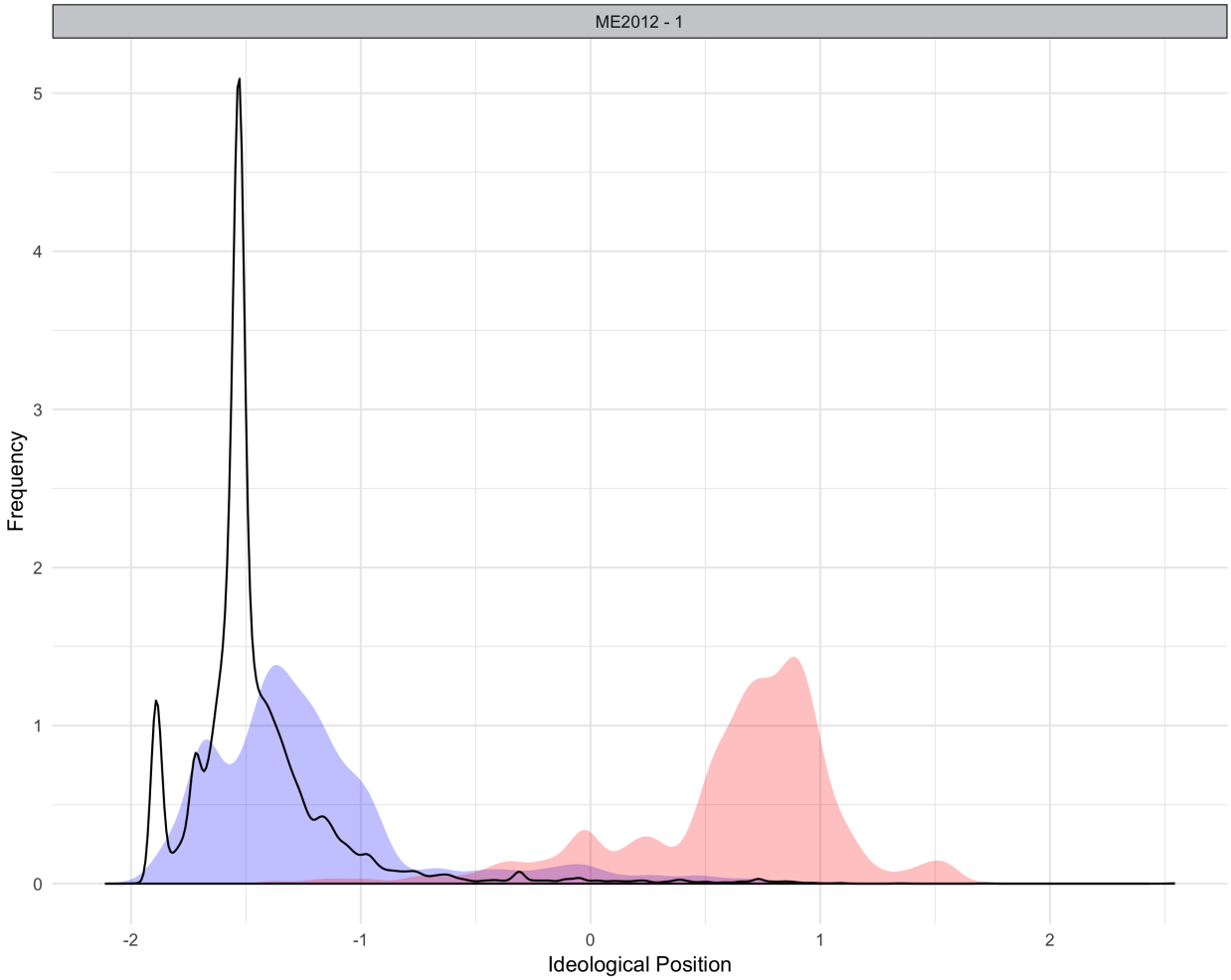


Figure A.11: Maine kernel density plots of initiative and legislative party donors' ideology

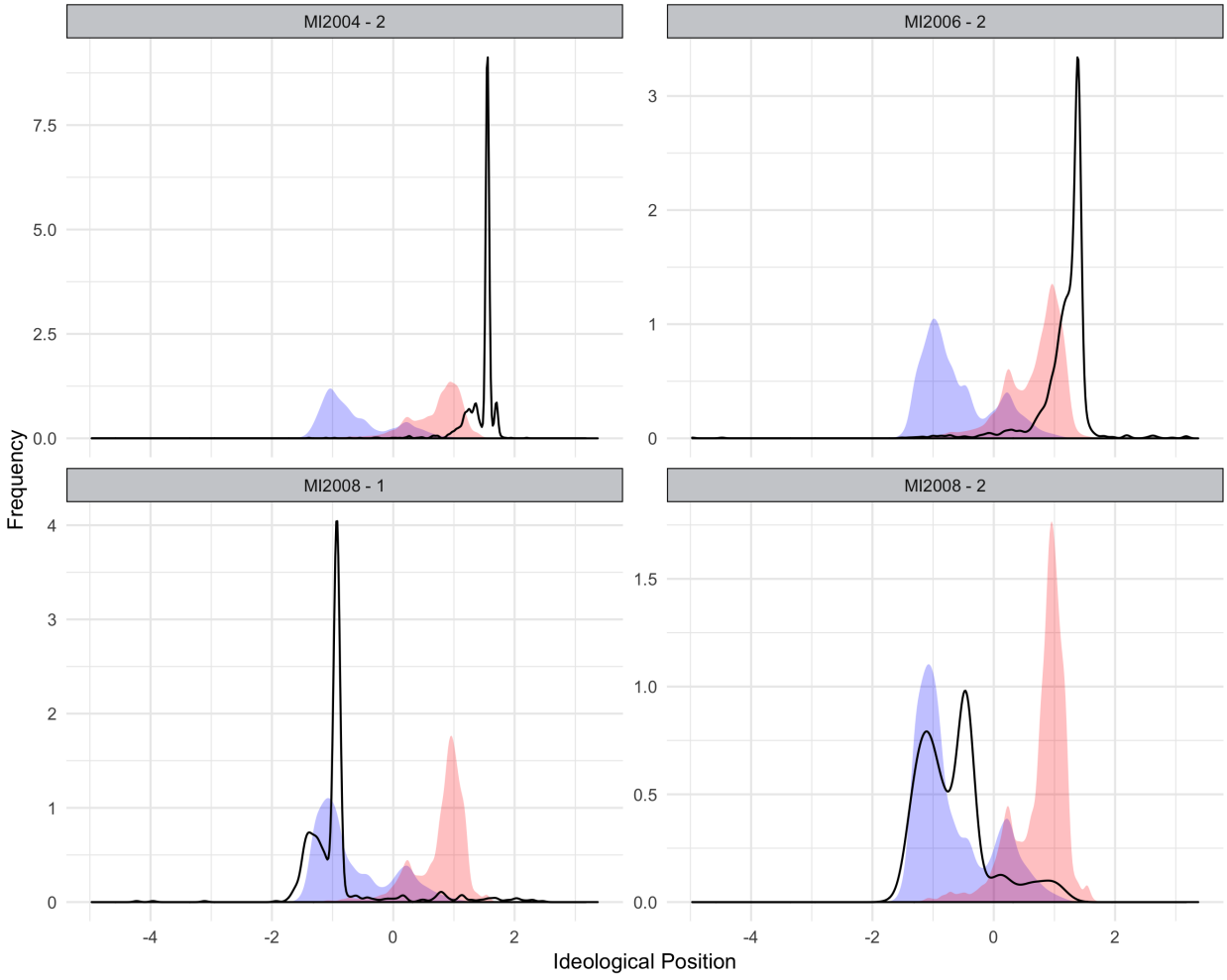


Figure A.12: Michigan kernel density plots of initiative and legislative party donors' ideology

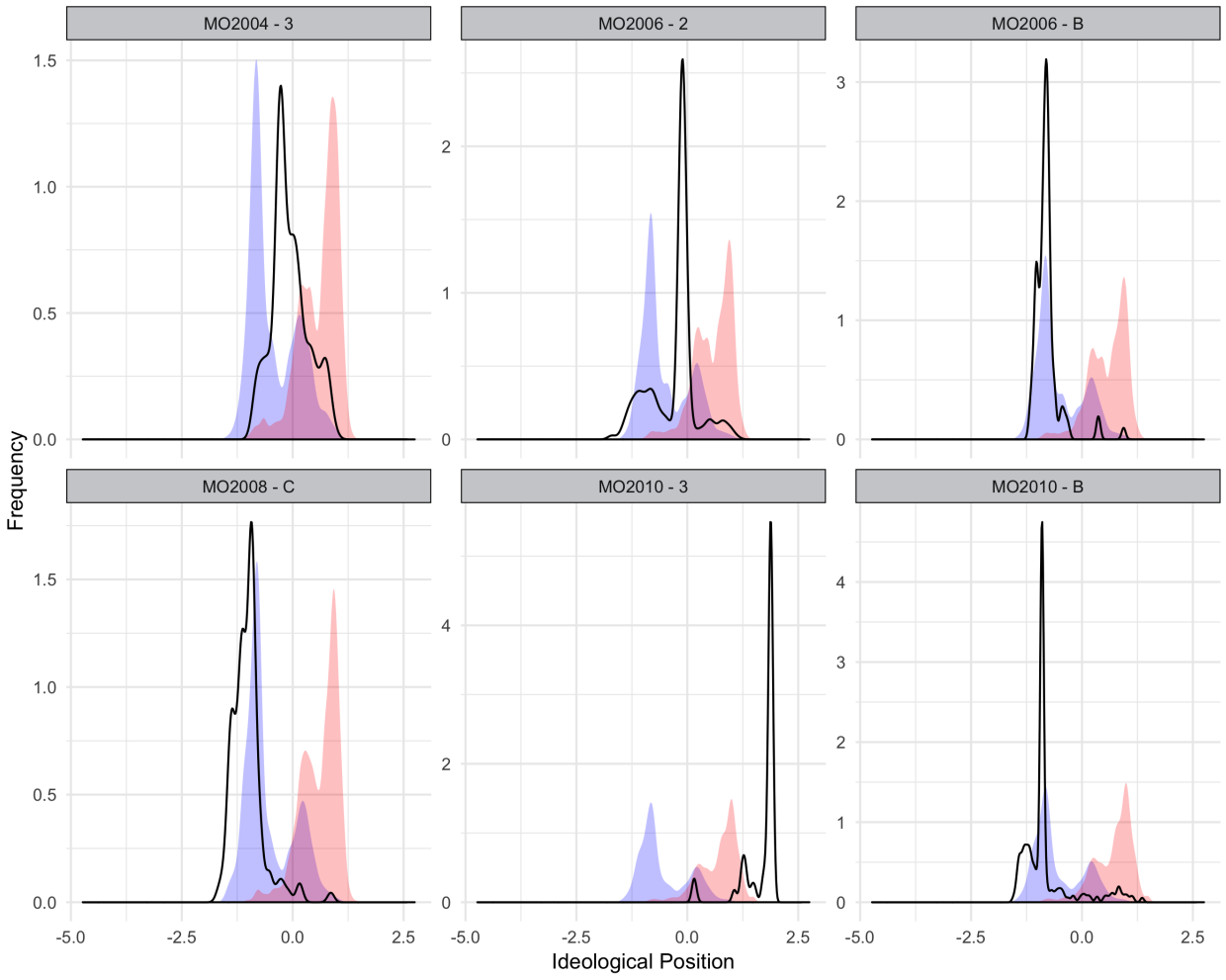


Figure A.13: Montana kernel density plots of initiative and legislative party donors' ideology

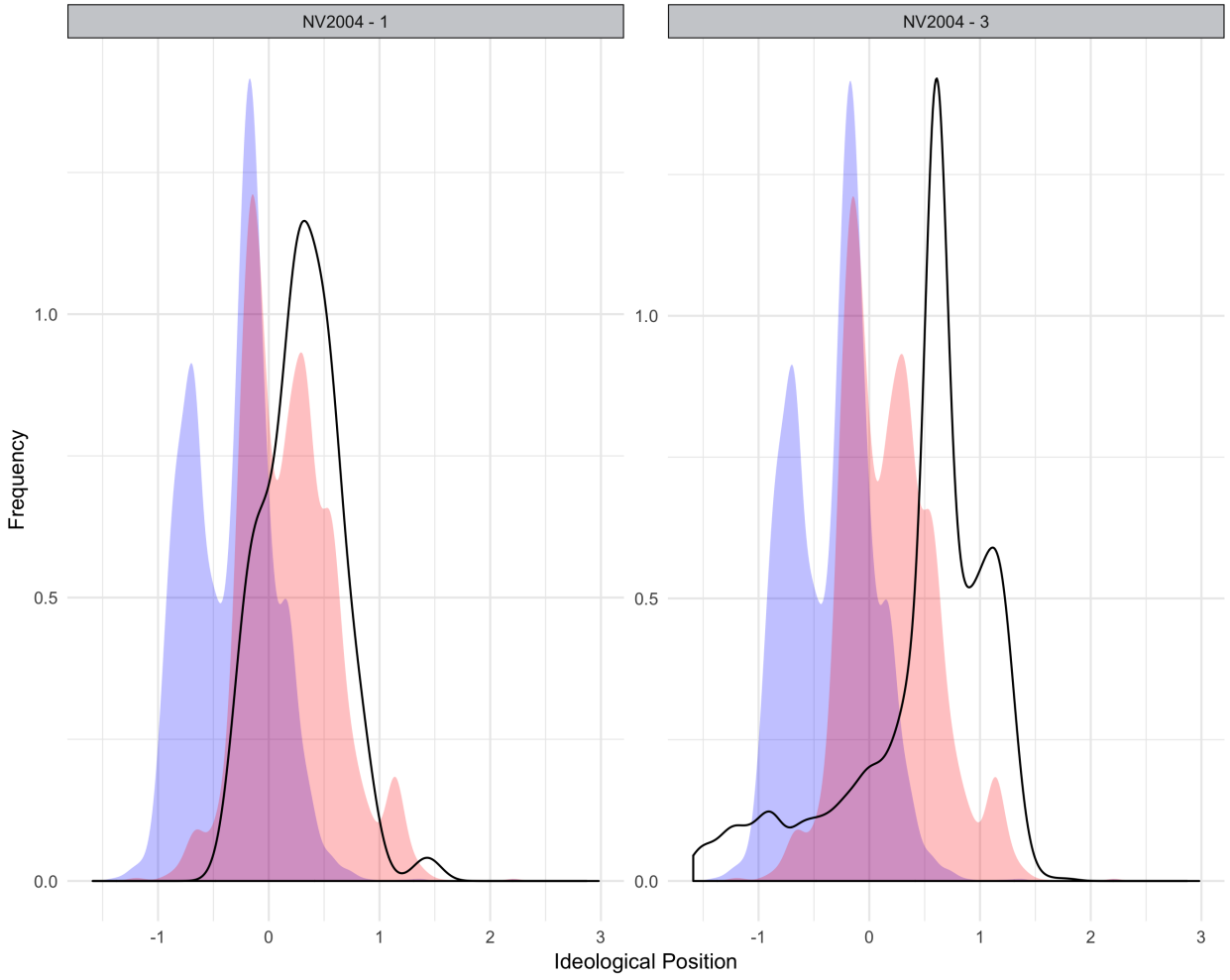


Figure A.14: Nevada kernel density plots of initiative and legislative party donors' ideology

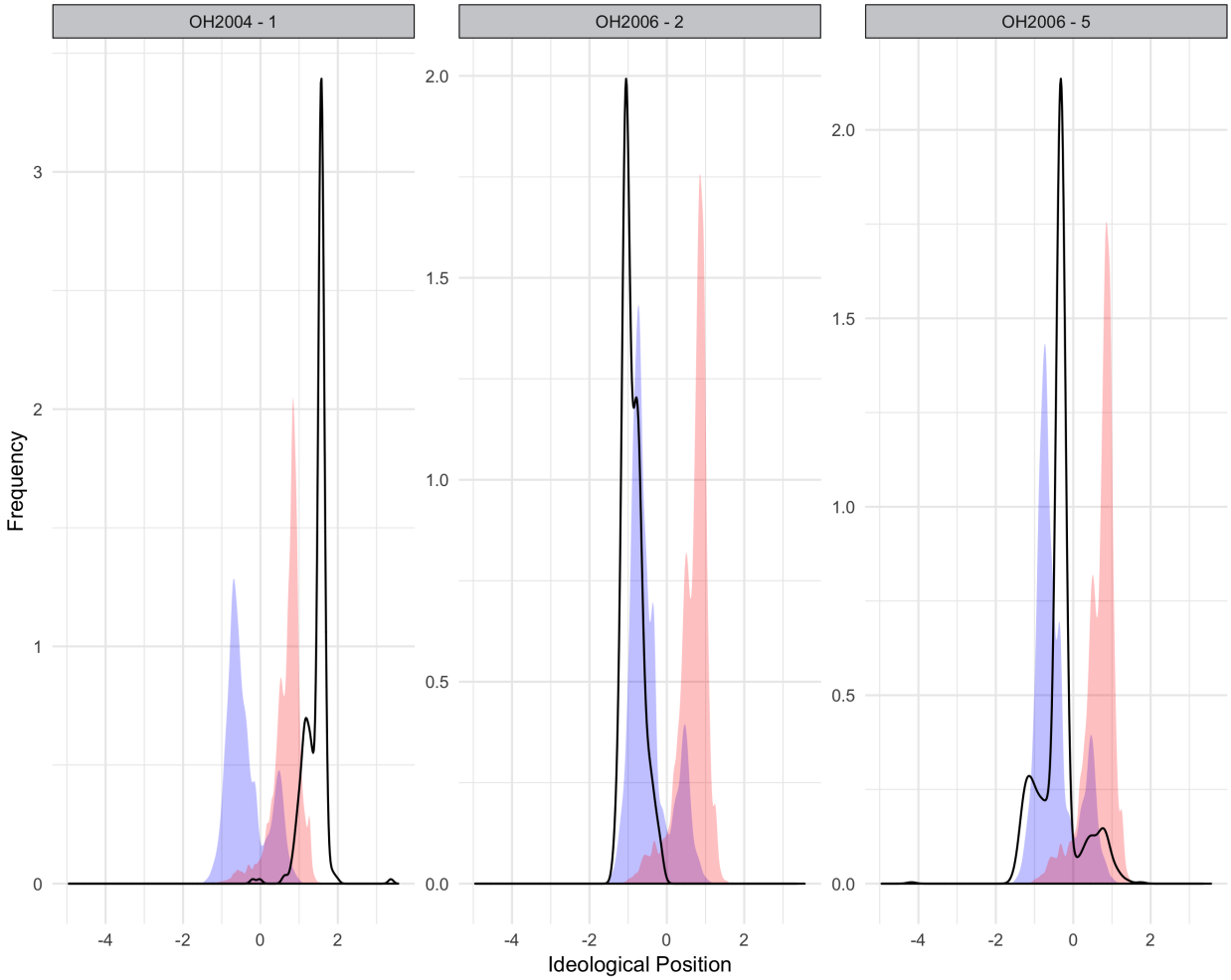


Figure A.15: Ohio kernel density plots of initiative and legislative party donors' ideology

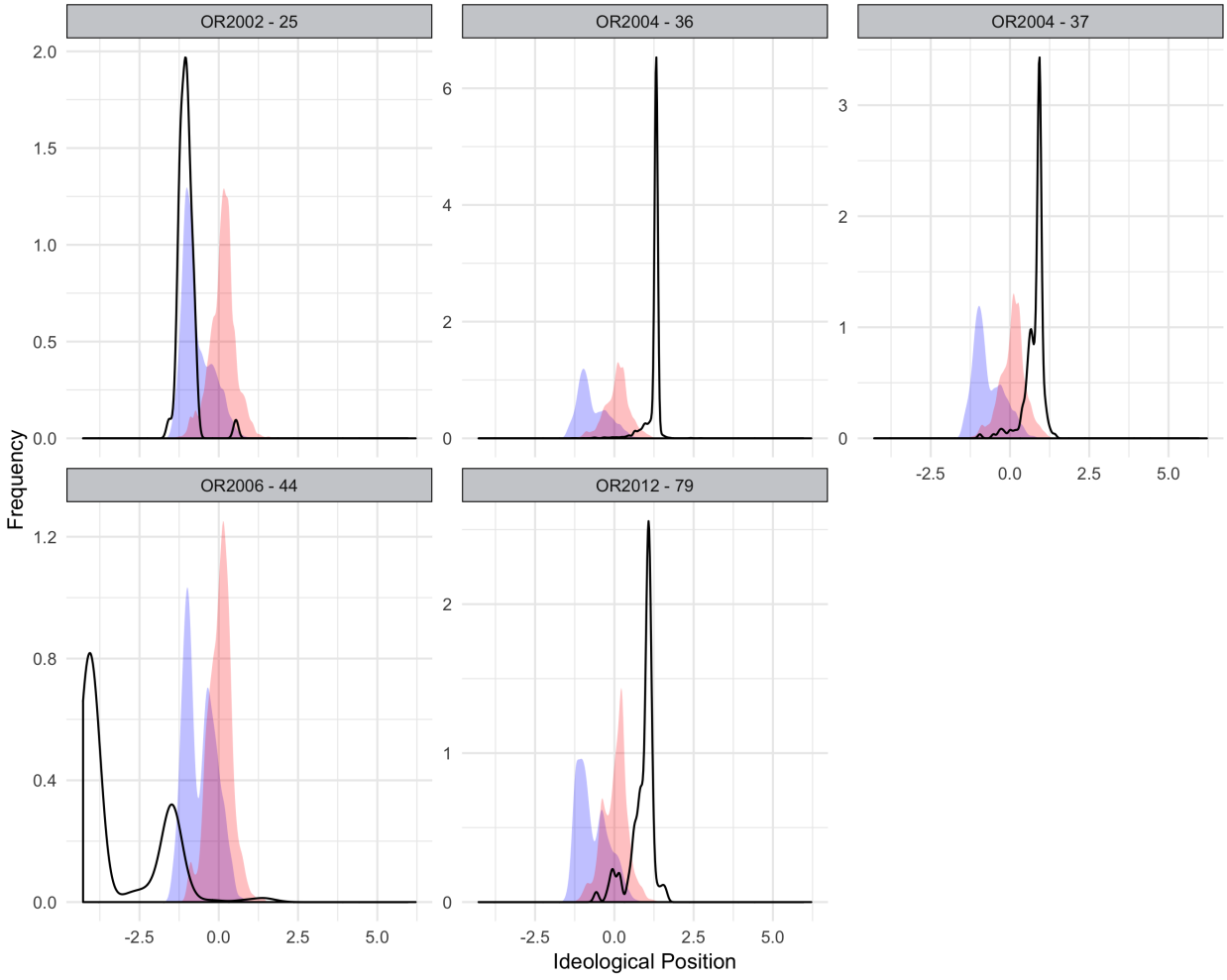


Figure A.16: Oregon kernel density plots of initiative and legislative party donors' ideology

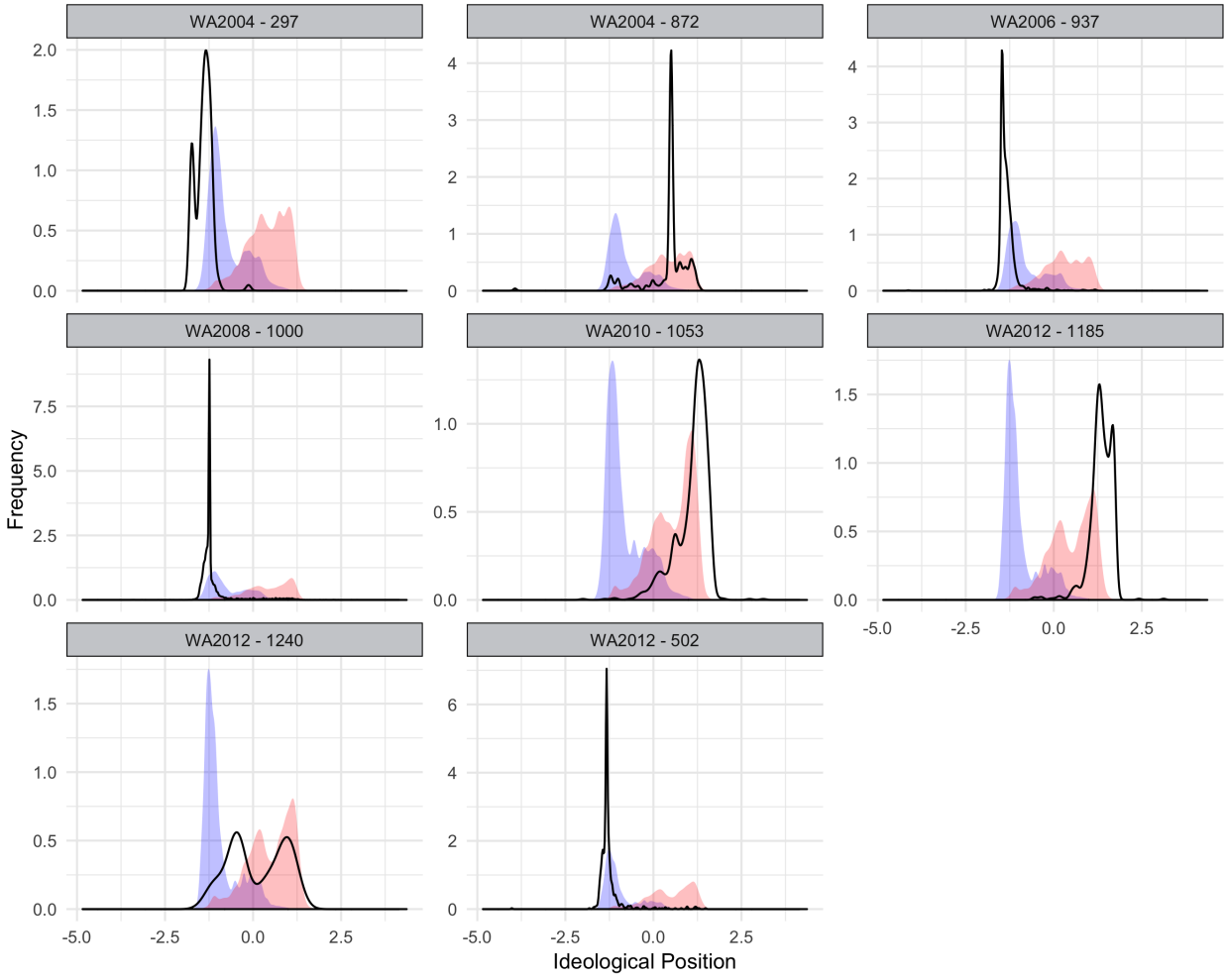


Figure A.17: Washington kernel density plots of initiative and legislative party donors' ideology

A.6 Factors deterring legislative action, by issue area

	Bad Policy	Econ. Costs	Interest Groups	Policy Sensitivity	Stalemate	Time	Uncertainty
Drugs	2.67	3.67	5.22	6.00	4.90	1.56	3.30
Elections	2.75	2.29	5.00	3.88	4.11	2.29	4.00
Environment	3.00	3.60	5.60	3.20	4.00	2.60	2.25
Guns	3.25	0.25	5.00	6.33	6.00	2.40	2.75
Tax	4.56	4.88	5.33	5.00	3.25	0.88	2.38
Wage	2.13	2.80	5.07	3.47	4.13	2.33	1.93

Table A.6: Legislator survey: mean importance of factors for deterring legislative action, by issue area

References

National Conference of State Legislatures. 2018. "Ballot Measure Database." http://www.ncsl.org/Default.aspx?TabId=16580&language=en-US.

Appendix B Supporting information for Chapter 3

B.1 Uncertainty as a function of randomisation

In the main text, Table 3.1 presents a particular case in which there is uncertainty over the treatment effect due to randomisation – what Abadie et al. (2014) call "causal uncertainty". In short, these issues arise because of the many potential random treatment assignments only one is ever realised. Given we can only ever estimate the treatment effect (we do not observe the full outcomes schedule) our estimate may differ dependent on the specific treatment assignment vector drawn. There are other plausible configurations of potential outcomes that generate similar concerns about causal uncertainty.

This section briefly motivates some alternative cases that collectively highlight why experimentalists should care about this causal uncertainty. In what follows, assume that our data is the full relevant population, such that sampling concerns drop out.

The first configuration demonstrates a Type I error whereby an estimated effect ($\hat{\tau}$) is non-zero when the true causal effect (τ) is precisely zero. In which case, without an appropriate measure of uncertainty, a single randomisation will yield a false positive result. Table B.1 presents a potential outcomes schedule and treatment assignment for which this state of affairs holds.

Given these potential outcomes τ is precisely zero – there is no effect on any subject

i	$Y_i(1)$	$Y_i(0)$	D_i
1	10	10	1
2	5	5	0
3	10	10	1
4	5	5	0

Table B.1: $\tau = 0, \hat{\tau} > 0$

within the population. However, given the random assignment D_i, the best estimate of the treatment effect is the difference in means between those in treatment and control i.e. $10 - 5 = 5$. This estimate is, technically, unbiased. As we increase the N the estimate will tend towards zero. But with this particular random assignment vector, it would appear there is a positive and substantial causal affect. To prevent a false positive, our measure of uncertainty should account for this type of chance occurrence even as the true population parameter is zero and the potential outcomes remain fixed.

In the configuration above, while the causal effect (of 0) is constant across individuals, the control/treatment outcomes vary. This sort of scenario depicts a highly simplified case where individuals hold different preferences/outcomes of which none are affected by an experimental stimulus. The treatment itself is homogeneous (it is precisely zero) but heterogeneity in the potential outcomes across individuals creates some leeway for our unbiased estimator to return a non-zero effect estimate.

It is also possible to construct cases where causal uncertainty yields a Type II error. Table B.2 presents an alternative potential outcomes schedule that yields the converse to Table B.1.

In this case, there are clearly two separate effects of treatment. For individuals $i = 1$ and $i = 2$ the effect of treatment is negative, whereas for the other two individuals it is positive. This might be the case in some sort of experiment where an information prime corrects individuals prior perceptions of some phenomena or quantity. Since the absolute size of the effect is the same across these two subgroups, the true population *average* treatment effect

i	$Y_i(1)$	$Y_i(0)$	D_i
1	5	10	1
2	5	10	1
3	5	0	0
4	5	0	0

Table B.2: $\tau > 0, \hat{\tau} = 0$

is zero. Under D_i in Table B.2, the estimated average treatment effect is positive.

As in the first potential outcomes schedule in Table B.1, this configuration exhibits heterogeneity in the potential outcomes. Unlike in the first case, the heterogeneity is present both in the control condition and the direction of the treatment effect. The treatment itself has a heterogeneous effect. Individuals' outcomes under treatment are identical precisely because the treatment acts in opposing directions for the two subgroups with different control outcomes.

The example potential outcomes in the main text demonstrate another issue where the true population average treatment effect is smaller than the estimated average treatment effect but still non-zero given some non-compliance. Common to this example and the further two supplementary configurations in this appendix section, uncertainty in the estimated effect arises as a result of stochastic assignment to either treatment or control.

Of course, the extent to which all members of particular subgroups are assigned to the same treatment status diminishes with the number of observations (assuming a truly random assignment). Part of the attraction of experimental methods is that we expect balance across covariates precisely for this reason – with equiprobable treatment probabilities and a sufficiently large sample any feature that may confound the effect estimate should be equally distributed among treatment conditions.

That being said, a truly random procedure will not necessarily ensure perfect balance. If there is any heterogeneity to the treatment effect, or to outcomes among individuals, different equiprobable assignments will lead to different estimated average treatment effects.

As researchers, we aim to making inferences about the true treatment effect. The confidence we place in our estimates should therefore capture this uncertainty as a function of randomisation, separate from any sampling-related concerns.

B.2 Derivation of variance

Suppose some treatment D is linearly related to an outcome y:[1]

$$Y = D\beta + \epsilon, \tag{B.1}$$

To mimimise the squared residuals, rearrange Equation B.1 to define the residual as $\epsilon = Y - X\beta$, multiply this expression by its transpose, and expand to get the quantity of interest[2]:

$$\epsilon'\epsilon = y'y - 2\beta' D'y + \beta' D'D\beta. \tag{B.2}$$

To find $\hat{\beta}$, we differentiate with respect to β, set the first-order condition to zero and solve:

$$\begin{aligned}
\frac{d\epsilon'\epsilon}{d\beta} &= -2D'y + 2D'D\hat{\beta} \Rightarrow 0 \\
D'D\hat{\beta} &= D'y \\
(D'D)^{-1}D'D\hat{\beta} &= (D'D)^{-1}D'y \\
\hat{\beta} &= (D'D)^{-1}D'y
\end{aligned} \tag{B.3}$$

For convenience later on, we can substitute Equation B.1 back into Equation B.3, to express the estimated coefficients in terms of the true coefficients and the residual:[3]

[1] This derivation is heavily based on course notes from by Michael J. Rosenfeld entitled "OLS in Matrix Form", which can be accessed at `https://web.stanford.edu/~mrosenfe/soc_meth_proj3/matrix_OLS_NYU_notes.pdf`.

[2] This step relies on the transposition of matrices. First as it applies to the expansion of the squared residuals:

$$\begin{aligned}
\epsilon'\epsilon &= (y - D\beta)'(y - D\beta) \\
&= (y' - \beta' D')(y - D\beta) \\
&= y'y - y'D\beta - \beta' D'y + \beta' D'D\beta,
\end{aligned}$$

and second as it applies to simplifying the expanded regression:

$$\begin{gathered}
y'D\beta \equiv \beta' D'y \\
\therefore y'y - y'D\beta - \beta' D'y + \beta' D'D\beta \equiv y'y - 2\beta' D'y + \beta' D'D\beta
\end{gathered}$$

[3] Here, note that $(AB)^{-1}(AB) = I$ and therefore $(X'X)^{-1}X'X$ drops out of the equation.

$$\begin{aligned}\hat{\beta} &= (D'D)^{-1}D'(D\beta + \epsilon) \\ &= (D'D)^{-1}D'D\beta + (D'D)^{-1}D'\epsilon \\ &= \beta + (D'D)^{-1}D'\epsilon \end{aligned} \tag{B.4}$$

Our estimate of variance $Var(\hat{\beta})$ is the squared difference between the "true" coefficient (β) and our best estimate ($\hat{\beta}$), such that:

$$Var(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \tag{B.5}$$

Substituting $\hat{\beta}$ with Equation B.4,

$$\begin{aligned} Var(\hat{\beta}) &= [((D'D)^{-1}D'\epsilon)((D'D)^{-1}D'\epsilon)'] \\ &= E[(D'D)^{-1}D'\epsilon)(\epsilon'D(D'D)^{-1})] \\ &= E[(D'D)^{-1}D'\epsilon\epsilon'D(D'D)^{-1})] \\ &= (D'D)^{-1}D'E[\epsilon\epsilon']D(D'D)^{-1} \end{aligned} \tag{B.6}$$

B.3 JEPS review articles

Table B.3: All articles published in *JEPS* between 2017 and 2019

Authors	Type of experiment	Sample	Main result	Observation-level	Obs	Cluster?
Pérez and Tavits (2019)	Survey experiment (phone)	Bilingual adults	Table 1	Individual	262	No
Morton and Ou (2019)	Lab experiment	Students	Figure 1	Individual-period	140	No
Loewen and MacKenzie (2019)	Survey (elite)	Political elites	Table 2	Politician-contact	404	Yes
Levine and Kline (2019)	Field experiment	Women members of community group	Table 1	Individual	100708	No
Israel-Trummel and Schachter (2019)	Survey experiment	African American adults	Figure 1	Individual	1200	No
Huddleston (2019)	Online survey experiment	Mturk	Figure 2	Individual	1512	No
Hou and Quek (2019)	Survey experiment	Nationally representative	Table 1	Individual	824	No
Green, Davenport and Hanson (2019)	Natural/survey experiment	Men	Table 5	Individual	900	Yes
Flynn and Krupnikov (2019)	Survey experiment	Nationally representative	Table 2	Individual	1031	No

Dietrich, Hyde and Winters (2019)	Lab experiment	Students	Table 5	Individual	777	No
DeScioli and Kimbrough (2019)	Lab experiment	*Not stated*	Table 1	Individual-round	4000	Yes
Denny and Driscoll (2019)	Survey experiment	Nationally representative	Table 2	Individual	120	No
Demel et al. (2019)	Lab-in-the-field experiment	Employed, unemployed, and students	Table 1	Individual-allocation	1092	Yes
Bokemper, Descioli and Kline (2019)	Online experiment	Mturk	Table 2	Individual	542	No
Bassi (2019)	Lab experiment	Students/university employees	Table 5	Individual-round	1660	No
Zhang (2018)	Lab experiment	Students	Table 2	Individual	371	Yes
White et al. (2018)	Online experiment	Mturk	Figure 1	Individual	2006	No
Trump and White (2018)	Survey experiment	Nationally representative	Figure 1	Individual	1014	No
Severson (2018)	Survey experiment	Nationally representative	Figure 2	Individual	918	No
Merolla and Zechmeister (2018)	Survey experiment	Adult	Figure 1	Individual	Numerous	No
Linos and Twist (2018)	Survey experiment	Nationally representative	Table 1	Individual	1000	No
Kingsley and Muise (2018)	Lab experiment	*Not stated*	Table 2	Individual-round	1260	Yes
Ihme and Tausendpfund (2018)	Survey experiment	Students	Table 1	Individual	377	No

Enos and Celaya (2018)	Online Experiment	Mturk/Student	Table 1	Individual-round	1081	Yes
Clayton (2018)	Natural experiment	Adults	Table 1	Individual	996	Yes
Butler and Pereira (2018)	Survey experiment	Political elites	Table 1	Individual	463	No
Busby and Druckman (2018)	Survey experiment	Students	Table 1	Individual	384	No
Buntaine and Prather (2018)	Online experiment	Mturk	Figure 1	Individual	1140	No
Bonilla and Mo (2018)	Survey experiment	Nationally representative	Figure 1	Individual	514	No
Andersen and Lau (2018)	Survey experiment	Mturk	Table 1	Individual	364	No
Amira (2018)	Survey experiment	Mturk	Figure 2	Individual	593	No
Soroka et al. (2017)	Survey experiment	Adults	Table 1	Individual	3783	No
Scotto et al. (2017)	Online survey experiment	Adults	Table 1	Individual	1010	No
Rogers (2017)	Lab experiment	Students	Figures 1/2	Individual-round	109	No
Peterson and Simonovits (2017)	Survey experiment	Nationally representative	Table 2	Individual	1924	No
Pedersen (2017)	Online survey experiment	Adults	Figure 1	Individual	1210	No
Lorenz, Paetzel and Tepe (2017)	Lab experiment	*Not stated*	Table 1	Individual-round	480	No
Li and Zeng (2017)	Online survey experiment	Adults	Figure 1	Individual-round	8240	Yes
Leeper (2017)	Online survey experiment	Nationally representative	Table 1	Individual	879	No

Kobayashi, Miura and Inamasu (2017)	Lab experiment	Students	Table 1	Individual	104	No
Klofstad (2017)	Online experiment	Adults	Table 1	Candidate	400	Yes
Jeon, Johnson and Robinson (2017)	Lab-in-the-field experiment	Two ethnic groups	Table 2	Game	89	No
Green and Zelizer (2017)	Field experiment	Republican women	Table 2	Individual	58751	No
Franco et al. (2017)	NA	NA	NA	NA	NA	NA
Costa (2017)	NA	NA	NA	NA	NA	NA
Butler and Crabtree (2017)	Field experiment	Political elites	Table 1	Individual	11801	No
Bechtel and Scheve (2017)	Online survey experiment	Nationally representative	Figure 3	Individual	8497	No
Barber et al. (2017)	Online survey experiment	Mturk	Table 2	Individual-choice	5720	No
Arceneaux (2017)	Online survey experiment	Mturk (white adults)	Table A6	Individual choice	1654	No

References

Abadie, Alberto, Susan Athey, Guido W Imbens and Jeffrey M Wooldridge. 2014. Finite Population Causal Standard Errors. Working Paper 20325 National Bureau of Economic Research.

Amira, Karyn. 2018. "Do People Contrast and Assimilate Candidate Ideology? An Experimental Test of the Projection Hypothesis." *Journal of Experimental Political Science* 5(3):195–205.

Andersen, David J. and Richard R. Lau. 2018. "Pay Rates and Subject Performance in Social Science Experiments Using Crowdsourced Online Samples." *Journal of Experimental Political Science* 5(3):217–229.

Arceneaux, Kevin. 2017. "Anxiety Reduces Empathy Toward Outgroup Members But Not Ingroup Members." *Journal of Experimental Political Science* 4(1):68–80.

Barber, Michael, David Gordon, Ryan Hill and Joseph Price. 2017. "Status Quo Bias in Ballot Wording." *Journal of Experimental Political Science* 4(2):151–160.

Bassi, Anna. 2019. "Weather, Risk, and Voting: An Experimental Analysis of the Effect of Weather on Vote Choice." *Journal of Experimental Political Science* 6(1):17–32.

Bechtel, Michael M. and Kenneth F. Scheve. 2017. "Who Cooperates? Reciprocity and the Causal Effect of Expected Cooperation in Representative Samples." *Journal of Experimental Political Science* 4(3):206–228.

Bokemper, Scott E., Peter Descioli and Reuben Kline. 2019. "Unfair Rules for Unequal Pay: Wage Discrimination and Procedural Justice." *Journal of Experimental Political Science* 6(3):180–191.

Bonilla, Tabitha and Cecilia Hyunjung Mo. 2018. "Bridging the Partisan Divide on Immigration Policy Attitudes through a Bipartisan Issue Area: The Case of Human Trafficking." *Journal of Experimental Political Science* 5(2):107–120.

Buntaine, Mark T. and Lauren Prather. 2018. "Preferences for Domestic Action Over International Transfers in Global Climate Policy." *Journal of Experimental Political Science* 5(2):73–87.

Busby, Ethan C. and James N. Druckman. 2018. "Football and Public Opinion: A Partial Replication and Extension." *Journal of Experimental Political Science* 5(1):4–10.

Butler, Daniel M. and Charles Crabtree. 2017. "Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions." *Journal of Experimental Political Science* 4(1):57–67.

Butler, Daniel M. and Miguel M. Pereira. 2018. "Are Donations to Charity an Effective Incentive for Public Officials?" *Journal of Experimental Political Science* 5(1):68–70.

Clayton, Amanda. 2018. "Do Gender Quotas Really Reduce Bias? Evidence from a Policy Experiment in Southern Africa." *Journal of Experimental Political Science* 5(3):182–194.

Costa, Mia. 2017. "How Responsive Are Political Elites? A Meta-Analysis of Experiments on Public Officials." *Journal of Experimental Political Science* 4(3):241–254.

Demel, Simona, Abigail Barr, Luis Miller and Paloma Ubeda. 2019. "Commitment to Political Ideology Is a Luxury Only Students Can Afford: A Distributive Justice Experiment." *Journal of Experimental Political Science* 6(1):33–42.

Denny, Elaine K. and Jesse Driscoll. 2019. "Calling Mogadishu: How Reminders of Anarchy Bias Survey Participation." *Journal of Experimental Political Science* 6(2):81–92.

DeScioli, Peter and Erik O. Kimbrough. 2019. "Alliance Formation in a Side-Taking Experiment." *Journal of Experimental Political Science* 6(1):53–70.

Dietrich, Simone, Susan D. Hyde and Matthew S. Winters. 2019. "Overseas Credit Claiming and Domestic Support for Foreign Aid." *Journal of Experimental Political Science* 6(3):159–170.

Enos, Ryan D. and Christopher Celaya. 2018. "The Effect of Segregation on Intergroup Relations." *Journal of Experimental Political Science* 5(1):26–38.

Flynn, D. J. and Yanna Krupnikov. 2019. "Misinformation and the Justification of Socially Undesirable Preferences." *Journal of Experimental Political Science* 6(1):5–16.

Franco, Annie, Neil Malhotra, Gabor Simonovits and L. J. Zigerell. 2017. "Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments." *Journal of Experimental Political Science* 4(2):161–172.

Green, Donald P. and Adam Zelizer. 2017. "How Much GOTV Mail Is Too Much? Results from a Large-Scale Field Experiment." *Journal of Experimental Political Science* 4(2):107–118.

Green, Donald P., Tiffany C. Davenport and Kolby Hanson. 2019. "Are There Long-Term Effects of the Vietnam Draft on Political Attitudes or Behavior? Apparently Not." *Journal of Experimental Political Science* 6(2):71–80.

Hou, Yue and Kai Quek. 2019. "Violence Exposure and Support for State Use of Force in a Non-Democracy." *Journal of Experimental Political Science* 6(2):120–130.

Huddleston, R. Joseph. 2019. "Think Ahead: Cost Discounting and External Validity in Foreign Policy Survey Experiments." *Journal of Experimental Political Science* 6(2):108–119.

Ihme, Toni Alexander and Markus Tausendpfund. 2018. "Gender Differences in Political Knowledge: Bringing Situation Back In." *Journal of Experimental Political Science* 5(1):39–55.

Israel-Trummel, Mackenzie L. and Ariela Schachter. 2019. "Does Shared Social Disadvantage Cause Black–Latino Political Commonality?" *Journal of Experimental Political Science* 6(1):43–52.

Jeon, Sangick, Tim Johnson and Amanda Lea Robinson. 2017. "Nationalism and Social Sanctioning Across Ethnic Lines: Experimental Evidence from the Kenya–Tanzania Border." *Journal of Experimental Political Science* 4(1):1–20.

Kingsley, David C. and Daniel Muise. 2018. "More Talk, Less Need for Monitoring: Communication and Deterrence in a Public Good Game." *Journal of Experimental Political Science* 5(2):88–106.

Klofstad, Casey A. 2017. "Looks and Sounds Like a Winner: Perceptions of Competence in Candidates' Faces and Voices Influences Vote Choice." *Journal of Experimental Political Science* 4(3):229–240.

Kobayashi, Tetsuro, Asako Miura and Kazunori Inamasu. 2017. "Media Priming Effect: A Preregistered Replication Experiment." *Journal of Experimental Political Science* 4(1):81–94.

Leeper, Thomas J. 2017. "How Does Treatment Self-Selection Affect Inferences About Political Communication?" *Journal of Experimental Political Science* 4(1):21–33.

Levine, Adam Seth and Reuben Kline. 2019. "Loss-Framed Arguments Can Stifle Political Activism." *Journal of Experimental Political Science* 6(3):171–179.

Li, Xiaojun and Ka Zeng. 2017. "Individual Preferences for FDI in Developing Countries: Experimental Evidence from China." *Journal of Experimental Political Science* 4(3):195–205.

Linos, Katerina and Kimberly Twist. 2018. "Diverse Pre-Treatment Effects in Survey Experiments." *Journal of Experimental Political Science* 5(2):148–158.

Loewen, Peter John and Michael Kenneth MacKenzie. 2019. "Service Representation in a Federal System: A Field Experiment." *Journal of Experimental Political Science* 6(2):93–107.

Lorenz, Jan, Fabian Paetzel and Markus Tepe. 2017. "Just Don't Call It a Tax! Framing in an Experiment on Voting and Redistribution." *Journal of Experimental Political Science* 4(3):183–194.

Merolla, Jennifer L. and Elizabeth J. Zechmeister. 2018. "Threat and Information Acquisition: Evidence from an Eight Country Study." *Journal of Experimental Political Science* 5(3):167–181.

Morton, Rebecca B. and Kai Ou. 2019. "Public Voting and Prosocial Behavior." *Journal of Experimental Political Science* 6(3):141–158.

Pedersen, Rasmus T. 2017. "Politicians Appear More Competent When Using Numerical Rhetoric." *Journal of Experimental Political Science* 4(2):129–150.

Pérez, Efrén O. and Margit Tavits. 2019. "Language Heightens the Political Salience of Ethnic Divisions." *Journal of Experimental Political Science* 6(2):131–140.

Peterson, Erik and Gabor Simonovits. 2017. "Costly Values: The Limited Benefits and Potential Costs of Targeted Policy Justifications." *Journal of Experimental Political Science* 4(2):95–106.

Rogers, Jonathan. 2017. "Nothing to Lose: Charitable Donations as Incentives in Risk Preference Measurement." *Journal of Experimental Political Science* 4(1):34–56.

Scotto, Thomas J., Jason Reifler, David Hudson and Jennifer vanHeerde-Hudson. 2017. "We Spend How Much? Misperceptions, Innumeracy, and Support for the Foreign Aid in the United States and Great Britain." *Journal of Experimental Political Science* 4(2):119–128.

Severson, Alexander W. 2018. "Partisan Affiliation and the Evaluation of Non-Prototypical Candidates." *Journal of Experimental Political Science* 5(2):121–147.

Soroka, Stuart, Matthew Wright, Richard Johnston, Jack Citrin, Keith Banting and Will Kymlicka. 2017. "Ethnoreligious Identity, Immigration, and Redistribution." *Journal of Experimental Political Science* 4(3):173–182.

Trump, Kris-Stella and Ariel White. 2018. "Does Inequality Beget Inequality? Experimental Tests of the Prediction That Inequality Increases System Justification Motivation." *Journal of Experimental Political Science* 5(3):206–216.

White, Ariel, Anton Strezhnev, Christopher Lucas, Dominika Kruszewska and Connor Huff. 2018. "Investigator Characteristics and Respondent Behavior in Online Surveys." *Journal of Experimental Political Science* 5(1):56–67.

Zhang, Nan. 2018. "Institutions, Norms, and Accountability: A Corruption Experiment with Northern and Southern Italians." *Journal of Experimental Political Science* 5(1):11–25.

Appendix C Supporting information for Chapter 4

C.1 Subject description

All subjects were recruited via the CESS Online USA subject pool, according to the exclusion restrictions noted in the main body of the text. Responses were collected between 18th February and 8th March 2019. Table C.1 outlines key demographic information about the subjects, and Figure C.1 plots the frequency of participants by state included within the sample. Post data-collection, 13 respondents who answered that they lived in Alabama were excluded from all analyses since Alabama does not have the initiative process.

Variable	Value	Freq (%)
Age	*Mean*	38.00
	Standard Deviation	14.14
Gender	Female	53.80
	Male	44.90
	Other:	0.30
	Prefer not to say	0.30
	Transgender	0.80
Ethnic	American Indian or Alaska Native	1.50
	Asian	6.70
	Black or African American	8.70
	Hispanic or Latino	3.30
	Native Hawaiian or Pacific Islander	0.30
	Other	5.40
	Prefer not to say	2.10
	White	71.80
	(Missing)	0.30
Party ID	A Democrat	46.20
	A Republican	14.60
	An independent	29.70
	Other	4.10
	Prefer not to say	1.30
	(Missing)	4.10
Ideology	*Mean*	4.26
	Standard Deviation	2.39

Table C.1: Descriptive summary of key demographics for conjoint experiment subjects.

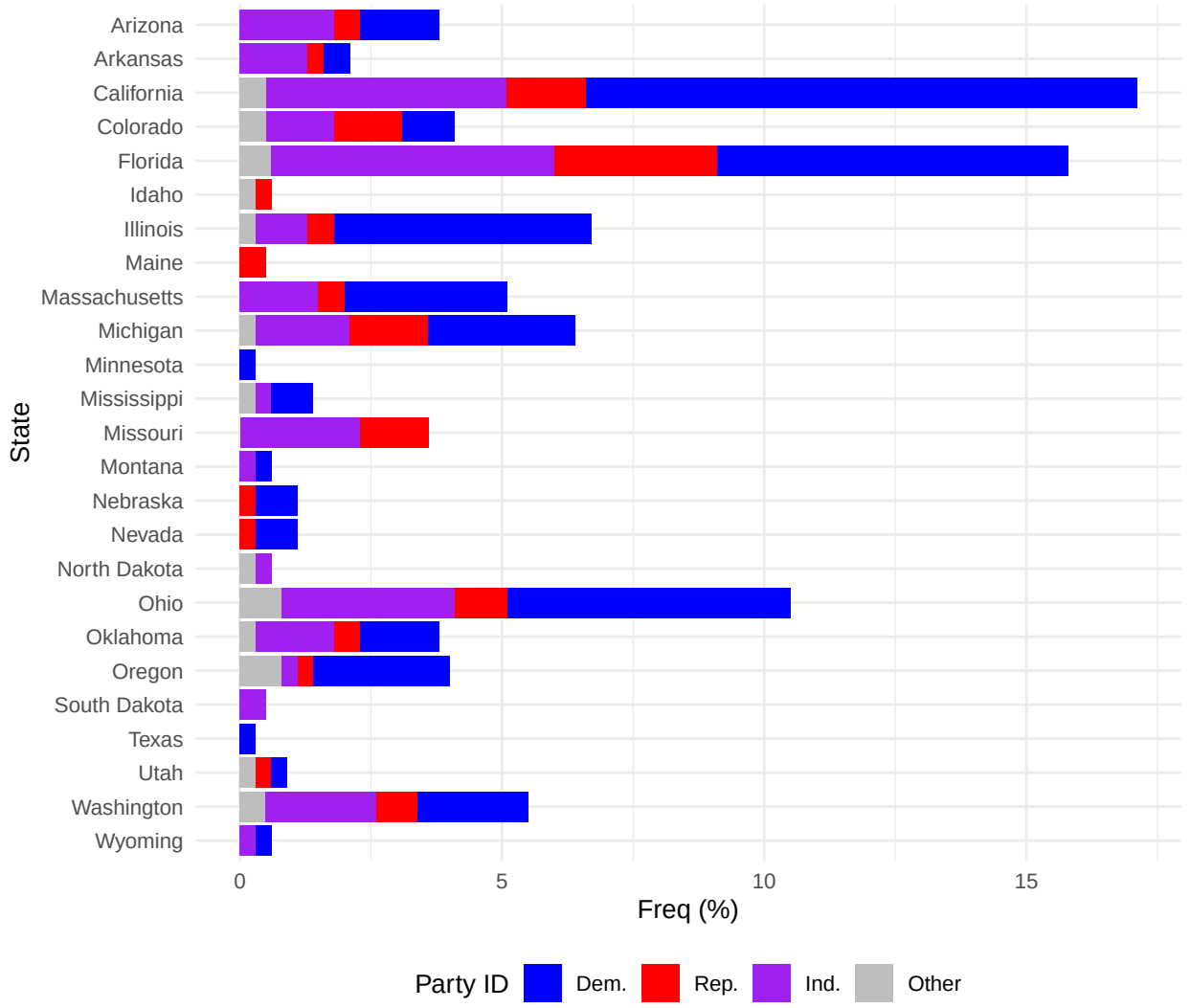


Figure C.1: Frequency of subject participation by states

C.2 Citizens United and its implications for disclosure

The *Citizens United v. FEC* (2010) decision struck down independent expenditure limits on corporations and unions. It did so on the basis of the First Amendment, ruling that freedom of speech extends to incorporated interests. Preventing these organisations from spending on political activity was, the Court ruled, an effective but unconstitutional impediment to their political expression.

As a result of this ruling businesses, unions, and other organisations like non-profit groups can now spend unlimited amounts of money in political races, so long as it is independently coordinated from candidates. This latter clause is crucial – corporations cannot donate unlimited amounts to candidates, there are strict limits at both the federal and state level (which I outline in further detail in the following sectioin). Instead, corporations can *spend* unlimited amounts in favour of a given candidate – for example producing advertisements and signage, holding events, and distributing pamphlets.

Citizens United represents a general weakening of regulation aimed at curbing expenditure in political campaigns. Those institutions typically most able to spend comparatively large amounts of money, compared to regular voters, face fewer obstacles when trying to advocate for candidates or parties. However, businesses themselves have been wary of direct spending in campaigns, and therefore the most visible impact immediately following the ruling has been with respect to non-profit spending (Briffault, 2011).

In and of itself the ruling is not about disclosure. But the changing regulatory landscape after *Citizens United* has at least two notable implications for the disclosure regulation in the United States. First, as limits on political spending are generally weakened, the regulatory 'workload' of disclosure has increased (Jiang, 2018). Simply put, without spending limits there are fewer pre-donation regulations that limit the acitivities of various classes of actors. Thus, the relative input of disclosure on regulating activity increases.

Second, the increase in political spending by non-profit organisations specifically has implications for the transparency of campaign finance activity. As mentioned in the main

text, non-profit entities registered under 501(c)(4) terms do not face the same disclosure requirements as other organisations. While these non-profits must disclose what *they* spend, they are not obliged to reveal their donation sources in the same way as candidates or candidate-affiliated PACs. As Mayer (2016) demonstrates, this can lead to complicated and convoluted networks of donations routed via 501(c)(4) organisations such that the original source of campaign spending is not revealed.

This latter feature is extremely consequential for disclosure construed in normative terms. Corporations can use hidden funding routes to spend in campaigns without being discovered, potentially misleading voters about the sources of electoral resources and who, ultimately, the benefactors are. If a donor has a negative reputation, or a corporation has a clear vested interest, strategically altering how electoral spending is presented to voters has the potential to change voter behaviour and potentially change electoral results (Wood and Spencer, 2016; Wood, 2017; Oklobdzija, 2019).

C.3 Summary of state campaign finance laws

As noted in the main text, campaign finance laws for gubernatorial candidates vary in terms of the contribution limits by state. Table C.2 details the total amount individuals, PACs, corporations and unions can donate to a single candidate per year (or electoral cycle where relevant). To reiterate, contribution limits in ballot initiative races are proscribed by federal court rulings.

[1]"Mega" PAC's can contribute $10,100

[2]$5675 for small donor committees.

[3]$5675 for small donor committees.

[4]Gubernatorial candidates with more than $250,000 independent expenditures are exempt, or if opposition candidate is self-funded (spending over $250,000.)

[5]Independent PACs can contribute up to $68,000.

State	Individual	PAC	Corporate	Union
Arizona	$5100	$5100[1]	Prohibited	Prohibited
Arkansas	$2700	$2700	Prohibited	Prohibited
California	$29200	$29,200	$29,200	$29,200
Colorado	$575	$575[2]	Prohibited	$575[3]
Florida	$3000	$3000	$3000	$3000
Idaho	$5000	$5000	$5000	$5000
Illinois	$5600[4]	$55,400	$11,100	$11,100
Maine	$1600	$1600	$1600	$1600
Massachusetts	$1000	$500	Prohibited	$500
Michigan	$6800	$6800[5]	Prohibited	Prohibited
Minnesota	$4000	$4000	Prohibited	$4000
Mississippi	Unlimited	Unlimited	$1000	Unlimited
Missouri	$2600	$2600	Prohibited	Prohibited
Montana	$1990	$10610	Prohibited	Prohibited
Nebraska	Unlimited	Unlimited	Unlimited	Unlimited
Nevada	$5000	$5000	$5000	$5000
North Dakota	Unlimited	Unlimited	Prohibited	Prohibited
Ohio	$12707.79	$12707.79	Prohibited	Prohibited
Oklahoma	$2700	$5000	Prohibited	Prohibited
Oregon	Unlimited	Unlimited	Unlimited	Unlimited
South Dakota	$4000	Unlimited	$4000	$4000
Texas	Unlimited	Unlimited	Prohibited	Prohibited
Utah	Unlimited	Unlimited	Unlimited	Unlimited
Washington	$2000	$2000	$2000	$2000
Wyoming	$2500	Unlimited	Prohibited	Prohibited

Table C.2: Contribution limits per year for individuals, PACs, corporations, and unions. Amounts quoted are for gubernatorial candidates. Data from National Conference of State Legislatures.

C.3.1 Example of initiative conjoint round

Initiative Title: Marijuana legalisation

If passed, this initiative would legalize the sale of marijuana within the state for recreational use for those aged 21 and over, subject to taxation and regulation by state authorities.

	Support	Opposition
Origin of Donations	Majority from donors out of state	Majority from donors within the state
Average Donation	$75	$75
Total Donations	$100,000 to $200,000	$70 million to $90 million
Largest Donor	Private individual	Corporation/Trade Association
Proportion of Campaign Funds from Largest Donor	90%	50%

init_marij_choice. If you had to choose, would you vote for or against this initiative?

- (•) For
- ( ) Against

init_marij_rate.
On a scale from 1 to 7, where 1 indicates that you strongly disapprove of the campaign and 7 indicates that you strongly approve of the campaign, how would you rate the two sides of the campaign?

1 = you strongly **disapprove** of the campaign
7 = strongly **approve** of the campaign

	Strongly Disapprove 1	2	3	4	5	6	Strongly Approve 7
Support	( )	( )	( )	( )	(•)	( )	( )
Opposition	( )	( )	( )	(•)	( )	( )	( )

Figure C.2: Screenshot of initiative conjoint

C.4 Causal assumptions

Stability and no carryover. In line with similar conjoint experiments about political candidates, I do not expect there to be carryover effects between rounds of the same conjoint experiment. The marginal effect of disclosing a majority of out-of-state donations, for instance, should remain stable whether it is presented in the first or last round of the experiment. To ensure this assumption holds, I reran the logistic regressions including a numeric control variable for the round the choice-profile was presented in (1-6). When this variable is included, the coefficients of the candidate conjoint attribute-levels are substantively unchanged, nor is the round variable statistically significant – suggesting that the stability assumption holds. Figure C.3 demonstrates these results compared to the original models.[6]

No profile-order effects. The profile-order assumption states that there is no distinct effect of the order of profiles within each task i.e. that any effect of a large total donations is constant whether it appears under Candidate A or B. This concern is mitigated, in part, by randomizing the order of attributes across profiles. To the extent I recover the average marginal effect by pooling across subjects and conjoint rounds, any profile-order effect (if present) should be netted-out. As a further robustness check, I regress a new model on the control-condition candidate data, interacting the disclosure variables with the profile indicator ("A" and "B"). None of the interactive terms approach conventional levels of statistical significance ($0.19 < p < 0.95$; see replication code for the full models), suggesting there is no difference whether an attribute was displayed in the first or second profile.

Randomized and atypical profiles. For the estimated marginal effects to be causally robust, the conjoint design should in theory assign non-zero probabilities to every possible

[6]This check assumes that the direction of any carryover-effect is uniform across attributes. Of course, the cumulative carryover effect could be statistically indistinguishable from zero whereas the marginal carryover effect for each attribute is non-zero. As a further check, one could rerun models on the subset of data for respondents' 'uncontaminated' first choice task alone. The diminished number of observations in this case, however, limits the extent to which this is a useful check.

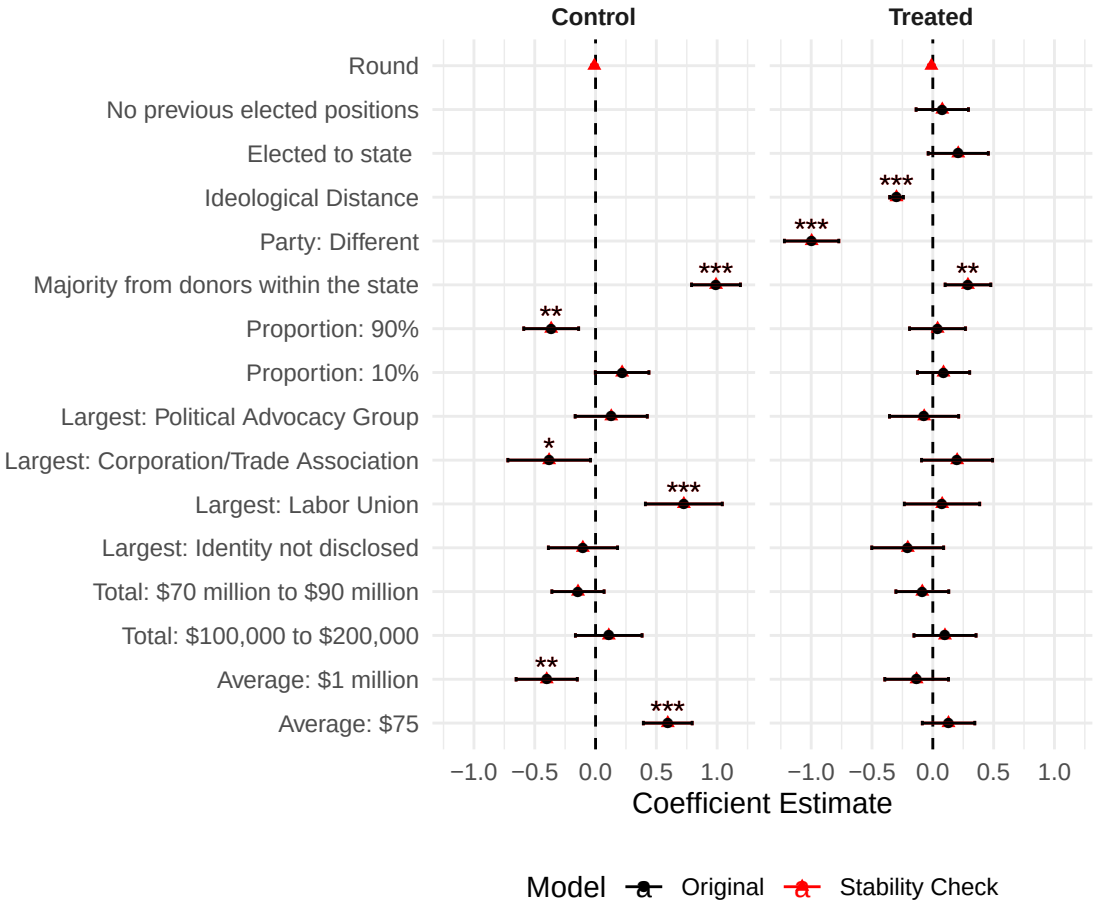


Figure C.3: Comparison of coefficients between models reported in the main text and models including a continuous variable of the conjoint round, to check for stability and any carryover effects. Coefficients are shown with 95 percent confidence intervals. Stars above coefficients indicate statistical significance at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

vector of treatments. Across the three conjoint experiments, however, I impose a very limited set of restrictions to ensure that the conjoint profiles are plausible. Across all three experiments, I prevented profiles where the average donation exceeded the total value of donations. Furthermore, for the informational equivalence candidate conjoint, I prevented profiles where the candidate was both an "extreme" liberal (conservative) and a Republican (Democrat). Given the limited set of restrictions imposed, the advantages of external validity

and subject engagement outweighed the smaller benefits of including atypical profiles.[7]

Balanced profiles. Finally, as with any randomisation procedure, it is crucial to show that the mechanics of said randomisation in fact lead to both balanced attribute profiles and subject characteristics. As a first check, Tables C.3 and C.4 show the proportion of times each attribute-level was displayed within the three conjoint experiments. No attribute-level was displayed a significantly higher or lower amount of times, relative to the other levels for the same attribute except for those attributes subject to restrictions. Here, it is the case that the unrestricted levels are nonetheless relatively equal in the proportion of times they were presented.

Moreover, I regress subjects' characteristics (age, gender, ideology) on profile attribute-levels using multinomial logit models. These tests check for imbalances in the assignment of levels across demographic categories. Overall, the attribute levels are well-balanced based on subjects' individual characteristics with *very* few attribute levels having a statistically higher or lower likelihood of being shown.[8] Overall, the evidence suggests the randomisation procedures across the experiment were successful.

[7]Indeed, very early on in the implementation, a coding error led to a limited number of profiles displaying implausible attribute combinations (these observations were subsequently excluded from the analysis). This prompted a respondent to email the experimental administrator to point out the incomprehensibility of the profiles, suggesting omitting these profiles is indeed the correct design decision.

[8]Given the number of models, by chance we would expect to observe some statistically significant coefficient irrespective of balance.

Attribute	Level	Bond	Enviro.	Marij.	Wage
Average	$1 million	0.27	0.23	0.26	0.27
Average	$10,000	0.38	0.37	0.37	0.35
Average	$75	0.35	0.40	0.37	0.38
Largest	Corporation/Trade Association	0.22	0.20	0.18	0.23
Largest	Identity not disclosed	0.19	0.22	0.21	0.18
Largest	Labor Union	0.21	0.20	0.18	0.21
Largest	Political Advocacy Group	0.17	0.19	0.22	0.19
Largest	Private individual	0.21	0.19	0.20	0.19
Origin	Majority from donors out of state	0.47	0.49	0.50	0.50
Origin	Majority from donors within the state	0.53	0.51	0.50	0.50
Prop	10%	0.31	0.36	0.35	0.32
Prop	50%	0.31	0.33	0.28	0.33
Prop	90%	0.38	0.31	0.37	0.35
Total	$1 million to $10 million	0.38	0.35	0.35	0.39
Total	$100,000 to $200,000	0.25	0.27	0.27	0.23
Total	$70 million to $90 million	0.37	0.37	0.38	0.38

Table C.3: Balance test: proportion of times each attribute-level was displayed to participants in the initiative conjoint

Attribute	Level	Control	Treat
Average	$1 million	0.24	0.24
Average	$10,000	0.38	0.40
Average	$75	0.38	0.36
Cand. Ideology	Centrist		0.22
Cand. Ideology	Moderate conservative		0.23
Cand. Ideology	Moderate liberal		0.24
Cand. Ideology	Very conservative		0.16
Cand. Ideology	Very liberal		0.16
Cand. Ideology	–	1.00	
Largest	Corporation/Trade Association	0.19	0.21
Largest	Identity not disclosed	0.20	0.21
Largest	Labor Union	0.21	0.19
Largest	Political Advocacy Group	0.21	0.20
Largest	Private individual	0.20	0.19
Office	Elected to federal office		0.34
Office	Elected to state office		0.34
Office	No previous elected positions		0.33
Office	–	1.00	
Origin	Majority from donors out of state	0.49	0.49
Origin	Majority from donors within the state	0.51	0.51
Party	Democrat		0.30
Party	Independent		0.37
Party	Republican		0.33
Party	–	1.00	
Prop	10%	0.35	0.33
Prop	50%	0.33	0.35
Prop	90%	0.33	0.31
Total	$1 million to $10 million	0.38	0.39
Total	$100,000 to $200,000	0.25	0.26
Total	$70 million to $90 million	0.37	0.35

Table C.4: Balance test: proportion of times each attribute-level was displayed to participants in the candidate conjoints

	Level:	
	\$10,000	\$75
Age	−0.004	−0.001
	(0.004)	(0.004)
Gender: Male	−0.160	−0.078
	(0.112)	(0.129)
Ideology	−0.019	−0.053*
	(0.023)	(0.027)
Constant	0.324	−0.169
	(0.182)	(0.208)
Akaike Inf. Crit.	3,802.436	3,802.436
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table C.5: Multinomial balance test for candidate (control) attribute: Average

	Level:
	Majority from donors out of state
Age	0.001
	(0.003)
Gender: Male	−0.042
	(0.098)
Ideology	−0.044*
	(0.020)
Constant	0.272
	(0.159)
Akaike Inf. Crit.	2,444.169
Note:	*p<0.05; **p<0.01; ***p<0.001

Table C.6: Multinomial balance test for candidate (control) attribute: Origin

	Level:	
	50%	10%
Age	0.010*	0.001
	(0.004)	(0.004)
Gender: Male	0.068	0.053
	(0.120)	(0.122)
Ideology	−0.007	0.005
	(0.025)	(0.025)
Constant	−0.317	−0.088
	(0.193)	(0.197)
Akaike Inf. Crit.	3,881.700	3,881.700
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table C.7: Multinomial balance test for candidate (control) attribute: Prop

	Level:	
	$1 million to $10 million	$100,000 to $200,000
Age	0.002	0.003
	(0.004)	(0.004)
Gender: Male	−0.043	0.013
	(0.126)	(0.113)
Ideology	−0.002	−0.021
	(0.026)	(0.023)
Constant	−0.480*	−0.102
	(0.204)	(0.182)
Akaike Inf. Crit.	3,826.984	3,826.984
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table C.8: Multinomial balance test for candidate (control) attribute: Total

	Level:			
	Private individual	Labor Union	Political Advocacy Group	Identity not disclosed
Age	−0.003	0.003	−0.004	−0.005
	(0.005)	(0.005)	(0.005)	(0.005)
Gender: Male	0.106	0.300	0.121	0.257
	(0.154)	(0.155)	(0.154)	(0.157)
Ideology	0.053	0.007	0.039	0.019
	(0.032)	(0.032)	(0.032)	(0.033)
Constant	−0.174	−0.280	−0.055	−0.099
	(0.248)	(0.248)	(0.248)	(0.252)
Akaike Inf. Crit.	5,697.256	5,697.256	5,697.256	5,697.256

Note: *p<0.05; **p<0.01; ***p<0.001

Table C.9: Multinomial balance test for candidate (control) attribute: Largest

	Level:	
	\$10,000	\$75
Age	0.004	0.004
	(0.004)	(0.005)
Gender: Male	−0.111	0.024
	(0.114)	(0.128)
Ideology	−0.021	−0.008
	(0.024)	(0.027)
Constant	−0.109	−0.658**
	(0.191)	(0.216)
Akaike Inf. Crit.	3,672.010	3,672.010
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table C.10: Multinomial balance test for candidate (control) attribute: Average

	Level:			
	Private individual	Labor Union	Political Advocacy Group	Identity not disclosed
Age	0.0004	0.004	0.009	−0.002
	(0.006)	(0.006)	(0.006)	(0.006)
Gender: Male	−0.125	−0.117	−0.023	0.029
	(0.157)	(0.160)	(0.156)	(0.158)
Ideology	0.019	0.048	0.036	0.028
	(0.033)	(0.034)	(0.033)	(0.034)
Constant	0.031	−0.312	−0.397	−0.019
	(0.264)	(0.269)	(0.264)	(0.266)
Akaike Inf. Crit.	5,489.995	5,489.995	5,489.995	5,489.995
Note:				*p<0.05; **p<0.01; ***p<0.001

Table C.11: Multinomial balance test for candidate (control) attribute: Largest

	Level:
	Majority from donors out of state
Age	0.004
	(0.004)
Gender: Male	0.106
	(0.099)
Ideology	0.035
	(0.021)
Constant	−0.335*
	(0.167)
Akaike Inf. Crit.	2,356.619
Note:	*p<0.05; **p<0.01; ***p<0.001

Table C.12: Multinomial balance test for candidate (control) attribute: Origin

	Level:	
	50%	10%
Age	0.005	0.002
	(0.004)	(0.004)
Gender: Male	0.009	−0.047
	(0.120)	(0.122)
Ideology	0.058*	0.056*
	(0.026)	(0.026)
Constant	−0.435*	−0.361
	(0.203)	(0.205)
Akaike Inf. Crit.	3,739.065	3,739.065
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table C.13: Multinomial balance test for candidate (control) attribute: Prop

	Level:	
	$1 million to $10 million	$100,000 to $200,000
Age	−0.002	−0.004
	(0.005)	(0.004)
Gender: Male	0.014	0.188
	(0.126)	(0.115)
Ideology	−0.012	0.053*
	(0.027)	(0.024)
Constant	−0.292	−0.268
	(0.210)	(0.195)
Akaike Inf. Crit.	3,687.805	3,687.805
Note:		*p<0.05; **p<0.01; ***p<0.001

Table C.14: Multinomial balance test for candidate (control) attribute: Total

	Level:	
	$10,000	$75
Age	−0.003	−0.0001
	(0.003)	(0.004)
Gender: Male	−0.011	0.128
	(0.094)	(0.103)
Ideology	0.010	−0.014
	(0.020)	(0.022)
Constant	0.106	−0.315
	(0.157)	(0.172)
Akaike Inf. Crit.	5,451.445	5,451.445
Note:		*p<0.05; **p<0.01; ***p<0.001

Table C.15: Multinomial balance test for initiative conjoint attribute: Average

	Level:			
	Private individual	Political Advocacy Group	Labor Union	Corporation/Trade Association
Age	0.014**	0.006	0.011*	0.011*
	(0.005)	(0.005)	(0.005)	(0.005)
Gender: Male	−0.010	−0.015	−0.086	−0.054
	(0.130)	(0.127)	(0.128)	(0.128)
Ideology	−0.027	−0.048	−0.015	−0.024
	(0.027)	(0.027)	(0.027)	(0.027)
Constant	−0.482*	0.041	−0.290	−0.275
	(0.218)	(0.212)	(0.213)	(0.214)
Akaike Inf. Crit.	8,069.141	8,069.141	8,069.141	8,069.141

Note: *p<0.05; **p<0.01; ***p<0.001

Table C.16: Multinomial balance test for initiative conjoint attribute: Largest

	Level:
	Majority from donors out of state
Age	0.002
	(0.003)
Gender: Male	0.058
	(0.081)
Ideology	−0.034*
	(0.017)
Constant	0.092
	(0.135)
Akaike Inf. Crit.	3,472.717
Note:	*p<0.05; **p<0.01; ***p<0.001

Table C.17: Multinomial balance test for initiative conjoint attribute: Origin

	Level:	
	50%	10%
Age	0.002	0.004
	(0.004)	(0.003)
Gender: Male	−0.016	0.165
	(0.101)	(0.099)
Ideology	0.015	−0.009
	(0.021)	(0.021)
Constant	−0.074	−0.075
	(0.168)	(0.165)
Akaike Inf. Crit.	5,501.713	5,501.713
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table C.18: Multinomial balance test for initiative conjoint attribute: Prop

	Level:	
	$1 million to $10 million	$100,000 to $200,000
Age	−0.003	−0.001
	(0.004)	(0.003)
Gender: Male	−0.080	0.090
	(0.105)	(0.093)
Ideology	0.005	0.015
	(0.022)	(0.020)
Constant	−0.230	−0.044
	(0.174)	(0.156)
Akaike Inf. Crit.	5,437.388	5,437.388
Note:		*p<0.05; **p<0.01; ***p<0.001

Table C.19: Multinomial balance test for initiative conjoint attribute: Total

C.5 Additional models

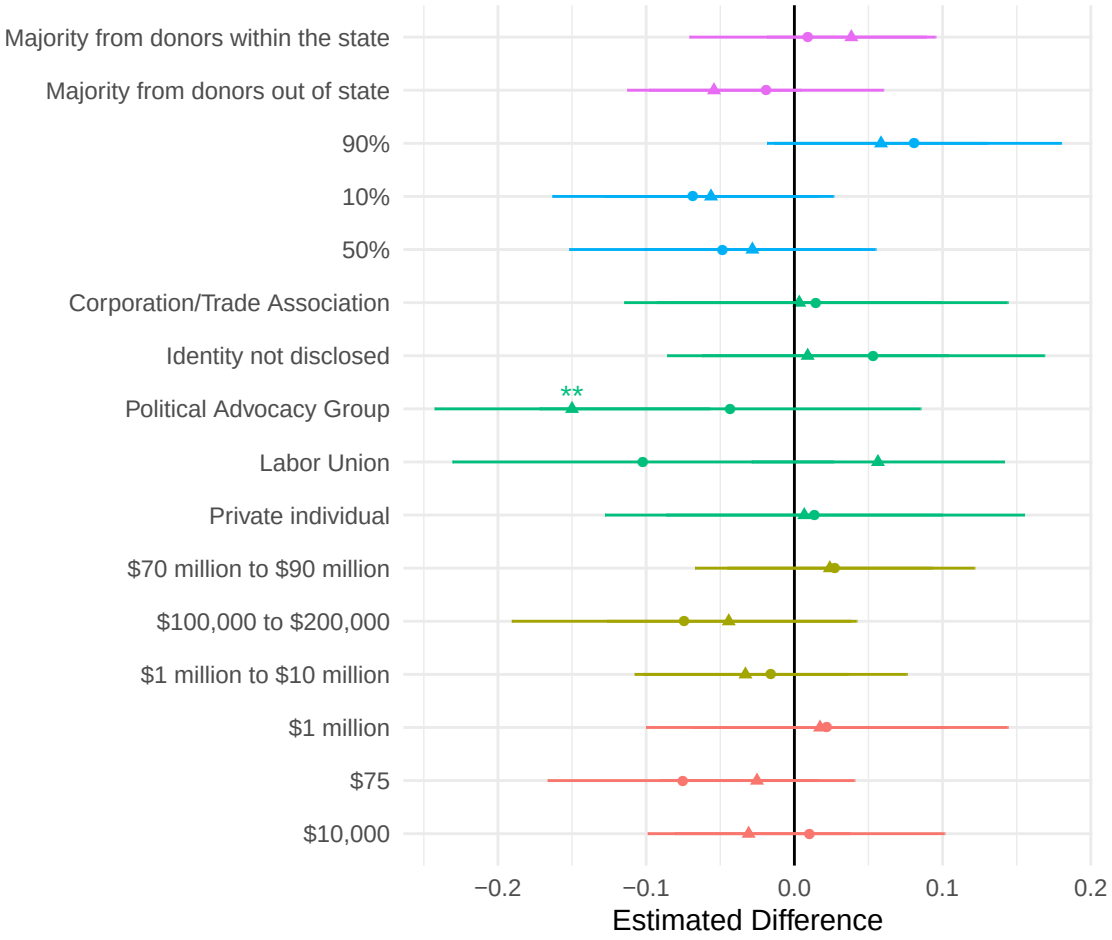


Figure C.4: Difference in marginal means for each attribute level by respondents' party identification, for subjects **not** exposed to additional party, ideology, and valence attributes. Estimated differences are shown with 95 percent confidence intervals. Stars indicate statistically significant differences at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

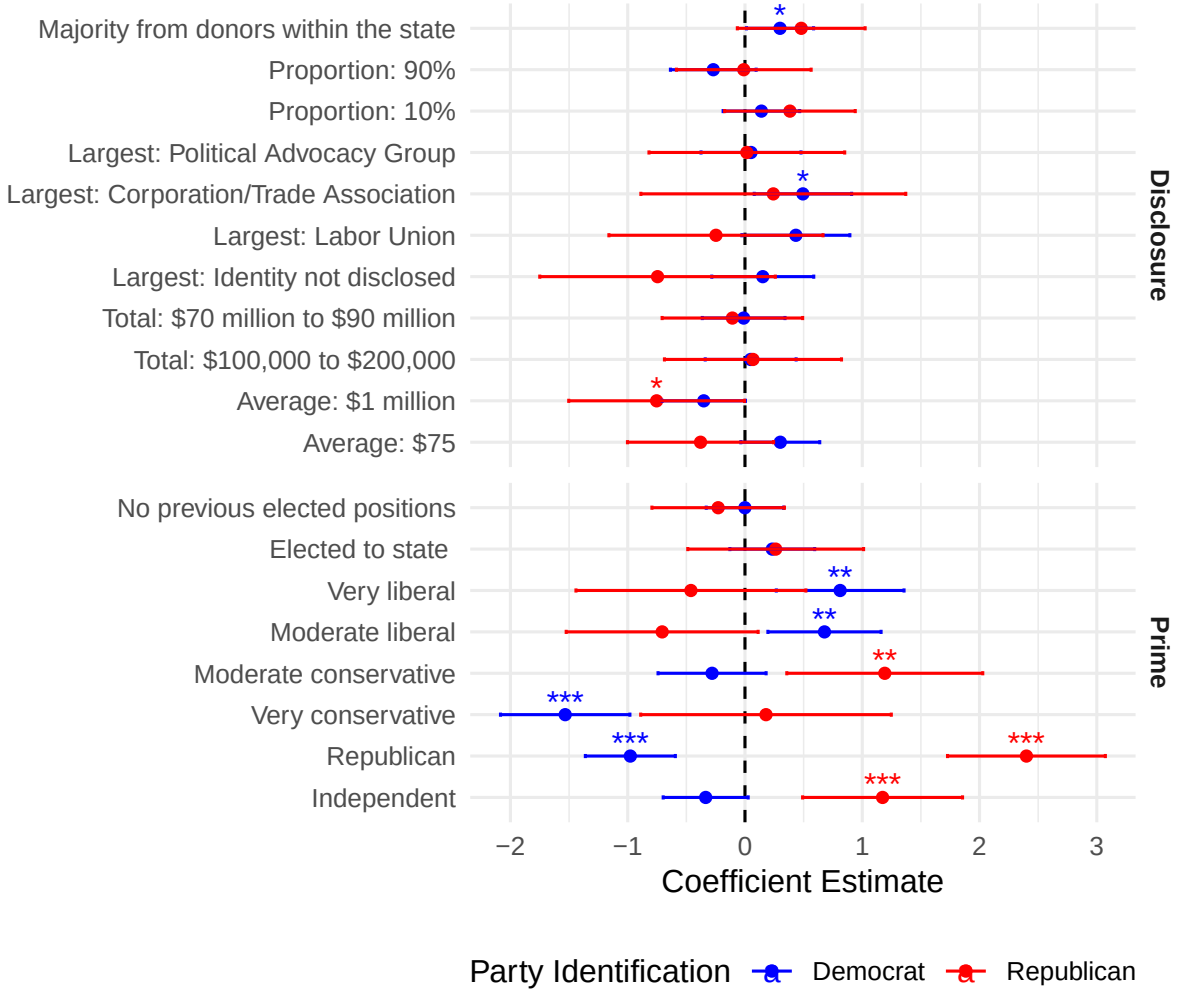


Figure C.5: Comparison of causal effects between Democratic and Republican respondents in the sample. Coefficients are shown with 95 percent confidence intervals. Stars above coefficients indicate statistical significance at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

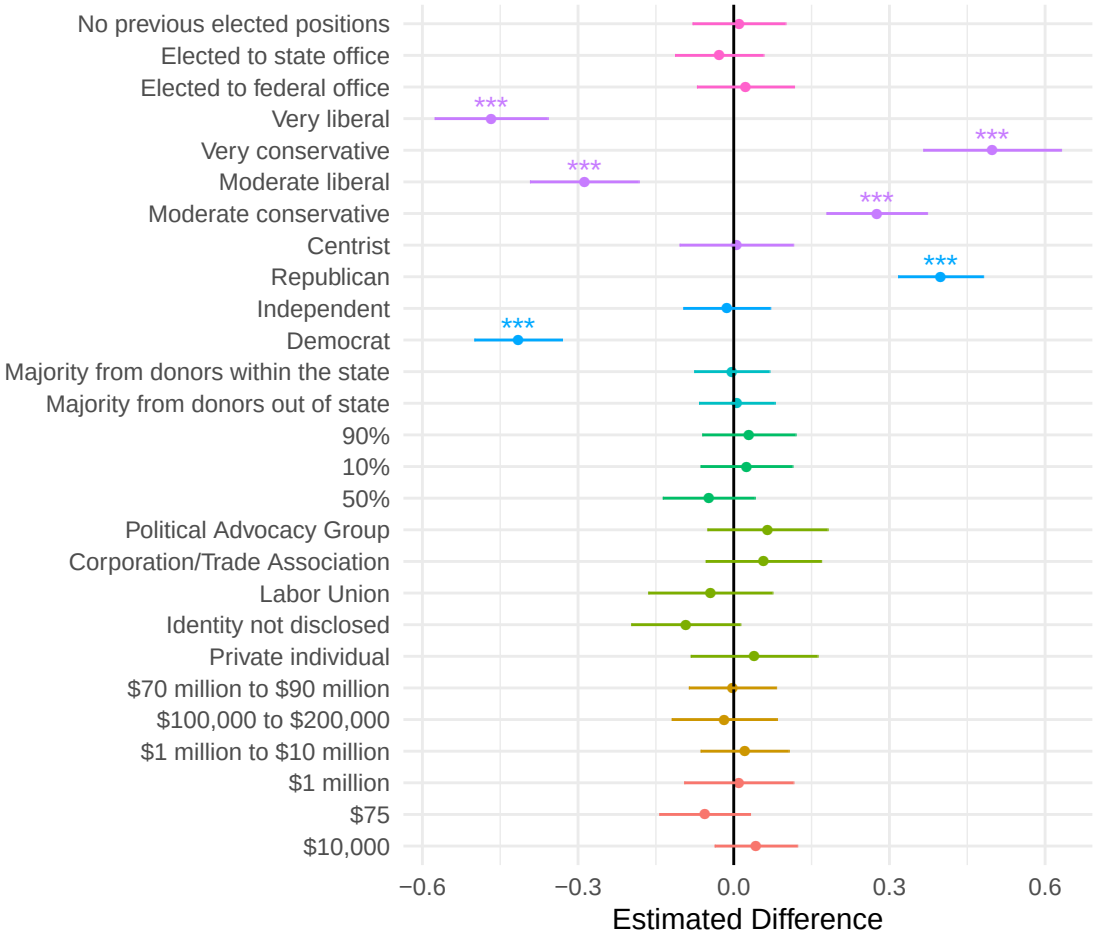


Figure C.6: Difference in marginal means for each attribute level by respondents' party identification, for subjects exposed to additional party, ideology, and valence attributes. Results for comparison between Republican and Democratic respondents only. Estimated differences are shown with 95 percent confidence intervals. Stars indicate statistically significant differences at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

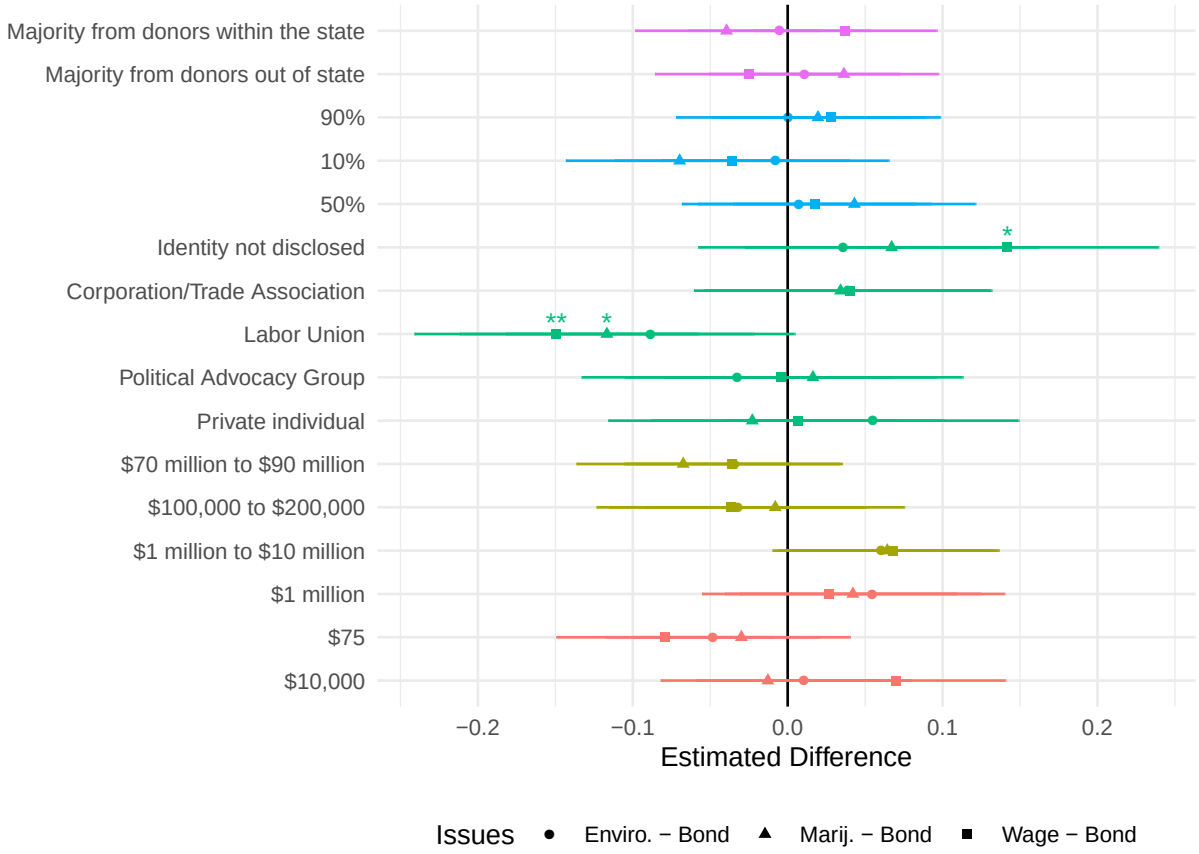


Figure C.7: Difference in marginal means by issue for the initiative conjoint. Estimated differences are shown with 95 percent confidence intervals. Stars indicate statistically significant differences at $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

References

Briffault, Richard. 2011. "Nonprofits and Disclosure in the Wake of Citizens United." *Election Law Journal: Rules, Politics, and Policy* 10(4):337–361.

Citizens United v. Federal Election Commission, 558 U.S. 310. 2010.

Jiang, Lear. 2018. Disclosure's Last Stand? The Need to Clarify the 'Informational Interest' Advanced by Campaign Finance Disclosure. SSRN Scholarly Paper ID 3262349.

Mayer, Jane. 2016. *Dark Money: The Hidden History of the Billionaires behind the Rise of the Radical Right*. First edition. ed. New York: Doubleday.

Oklobdzija, Stan. 2019. "Public Positions, Private Giving: Dark Money and Political Donors in the Digital Age." *Research and Politics* .

Wood, Abby K. 2017. "Show Me the Money: 'Dark Money' and the Informational Benefit of Campaign Finance Disclosure." *SSRN Electronic Journal* .

Wood, Abby K. and Douglas M. Spencer. 2016. "In the Shadows of Sunlight: The Effects of Transparency on State Political Campaigns." *Election Law Journal: Rules, Politics, and Policy* 15(4):302–329.