Do Primaries Work?
Constituent Ideology and Congressional Nominations

By

Michael G. DeCrescenzo

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Political Science)

at the

University of Wisconsin–Madison,

2020

Approved by the thesis committee on the oral defense date, October 16, 2020

Barry C. Burden (Chair), Professor of Political Science

Kenneth R. Mayer, Professor of Political Science

Eleanor Neff Powell, Associate Professor of Political Science

Alexander M. Tahk, Associate Professor of Political Science

Michael W. Wagner, Professor of Journalism and Mass Communications

for Tina:

I became a worse political scientist
so I could be a better person.

Abstract

In contemporary electoral politics in the U.S., primary elections are widely believed to play a crucial role. Many scholars believe that primary election competition is the standout reason why classic predictions from formal models of electoral competition—that candidates take ideological positions near the median voter—fail to manifest in the real world. The general election context provides incentives for candidates to take centrist policy positions, but candidates must win their party's nomination before advancing to the general election. Because primary elections take place predominantly among voters of one political party affiliation, and because those voters tend to hold strongly partisan beliefs about political issues, candidates feel more acute incentives to take strong partisan stances on issues rather than moderate stances even amid stiff general election competition.

This story of primary elections and representation is widely believed, but is it true? Despite its prominence, the empirical evidence is unclear. The theory rests on a notion that voters make informed choices in primary elections by consulting their policy preferences and choosing the candidate with the closest policy platform. Past research has been unable to operationalize key constructs in this prediction, or it has operationalized the wrong constructs. Candidates should take more extreme positions when the primary constituency has a stronger preference for ideologically extreme policy, but studies have not directly measured the policy preferences of partisans within a candidate's district. Further, districts where partisans hold more extreme preferences should nominate candidates with more extreme campaign positions as well, but methods for estimating candidates' ideological positions have been incompletely applied to the study of primaries. Moreover, because primary elections are characterized by low levels of voter information and the partisanship of candidates is held largely constant, non-policy forces such as candidate valence and campaign spending may be more powerful than in general elections. For these reasons, the proposition that primary

elections advance the ideological interest of local partisan voters is theoretically contestable.

This dissertation develops and applies new Bayesian approaches for estimating both constructs that have yet eluded the study of primary politics: the preferences of partisan voters as a group and the campaign positioning of primary candidates. With these estimates in hand, I explore the relationship between local partisan preferences and primary candidate positions. Do primary candidates position themselves relative to partisan primary voters, and is the relative extremism of partisan constituencies related to the ideological positions of the candidates they nominate?

Contents

List of Tables

List of Figures

Acknowledgments

This project is better thanks to the advice and suggestions of my committee, Scott Straus and the students in the 2017 prospectus workshop, other Burden advisees (Levi Bankston, Jordan Hsu, Matt Shor, and Rochelle Snyder), David Canon, Devin Judge-Lord, Evan Morier, Blake Reynolds, and Marcy Shieh. Researchers at other institutions provided suggestions and moral support about the ideas in this project as well: Devin Caughey, David Doherty, Andrew Heiss, Seth Hill, Shiro Kuriwaku, TJ Mahr, Steve Miller, Jacob Montgomery, Arman Oganisian, John Poe, Rachel Porter, Jonathan Robinson, Sarah Treul, and Chris Warshaw. Extra special thanks to researchers whose freely-provided data made the project possible: Robert Boatright, Adam Bonica, Ella Foster-Molina, Seth Masket, Nolan McCarty, Eric McGhee, Vincent Moscardelli, Steven Rogers, Boris Shor, Clifford Vickrey, and the teams who make the ANES and CCES surveys possible. I received financial support from the Elections Research Center at the University of Wisconsin–Madison, and technical support from UW's Social Science Computing Cooperative, Garrick Aden-Buie, and Matthew Kay. Thanks also to Deb McFarlane for her help stewarding other students and myself through the bureaucracy of dissertating.

Thanks also to my parents, who kept so many pieces of my life together during this process. Thanks to Ricky O'Connor, who connected me to my first job after graduate school that gave me the push to finish the project and move on to the next thing. Thanks finally to some close friends who made my time in Madison better: Shaan Amin, Josh Cruz, Micah Dillard, José Luis Enríquez Chiñas, Caileigh Glenn, Ryan Hinchey, Jordan Hsu (again), Devin Judge-Lord (again), Amy Kawleski, Richard Loeza, Anna Meier, Evan Morier (again), Erin Nelson, Anna Oltman, Madeline Vogt, and Erin Zwick.

— 1 —

Introduction: Policy Ideology and Congressional Primaries

Elections are the foremost venue for citizens to influence government actors and public policy. Classic theories of voting suggest that citizens weigh the policy positions of alternative candidates and vote for the candidate whose platform most closely aligns with their own preferences (Downs 1957). Political parties simplify the voter's calculations by providing a powerful heuristic in the form of the party label, enabling voters to infer candidates' values and issue positions without expending the effort to thoroughly appraise each campaign (Campbell et al. 1960; Green, Palmquist, and Schickler 2002; Rahn 1993).

The rise of partisan polarization, however, has complicated the role of parties in U.S. politics. Although citizens, journalists, pundits, and even elected leaders frequently bemoan the bitter rhetoric and legislative gridlock that has accompanied the widening partisan divide, political scientists have noted several positive consequences to polarization. Compared to the parties of the early- and mid-1900s that political scientists believed were too similar to provide voters with meaningful choices (American Political Science Association 1950), the Democratic and Republican Parties of recent decades have taken divergent and oppositional stances across a greater number of policy issues. As a result, voters can more easily differentiate the policy platforms of the two parties in order to vote consistently with their political values. Voters in turn became more thoroughly sorted into partisan groups that represent

distinct ideological viewpoints in American politics, holds beliefs across multiple issues that are more ideologically consistent, think more abstractly about the ideological underpinnings of issue stances, and participate more in politics than they did in the past (Abramowitz and Saunders 1998; Fiorina, Abrams, and Pope 2005a; Layman and Carsey 2002; Levendusky 2009).

Even as polarization has strengthened many aspects of political representation between the two parties, it may have troubling effects on representation within the two major parties. The typical voter is a partisan who intends to cast her ballot for her preferred party, whoever that candidate may be (Bartels 2000; Petrocik 2009). As party-line voting increases, voters are more thoroughly captured by their loyalties. A partisan voter's choices are locked in long before Election Day. Candidates from her preferred party have already been selected through a nomination process, and she may be more likely to abstain from voting when faced with an undesirable candidate than she is to vote for a different party (Hall and Thompson 2018). Recent research supports this notion of capture amid polarization—when voters must choose between polarized candidates, they become less responsive to candidates' actual platforms and instead are more influenced by motivated reasoning and partisan teamsmanship (Rogowski 2016). Voters relax their substantive scrutiny of candidates to cast low-cost votes for their own party, weakening the influence of *policy* as a separate consideration from partisanship.

This presents an important problem for our understanding of how elections contribute to the representation of voter preferences in government. Elections are intended to be a voter's choice over alternative political values to be expressed in government, but if the choice of candidates does not present the average partisan voter with realistic alternatives, how should we think about the "representation" of these voters' actual policy preferences? If general elections provide an ever-coarsening choice over policy priorities, does the U.S. electoral system incorporate voters policy preferences in other ways?

When the choice before voters in the general election does not present realistic alterna-

tives, political scientists naturally shift their focus to the nomination of partisan candidates. V.O. Key, for example, studied Democratic Party dominance in the American South, asking if competition within the party could provide a quality of representation similar to two-party competition (Key 1949). Although scholars are right to examine within-party competition, focusing on contexts of single-party dominance is a serious limitation. Even in races between viable candidates from both major parties, within-party competition plays a crucial role simply due to the fact that partisan voters almost certainly cast a vote for their own party. Rank-and-file partisan constituents are all but captured. If they are to express their policy preferences through the act of voting, their voices may register as relatively weak because they present little electoral risk to their party in the general election. The nomination stage—the primary election in particular—remains an important venue for the representation of partisans' policy views, whether the general election is closely contested or not.

1.1 Policy Preferences and the Strategic Positioning Dilemma

This dissertation is chiefly concerned with the policy preferences of partisan voters and their role in electoral representation through Congressional primary elections. The study of American electoral politics has not ignored the representational function of primary elections (Aldrich 2011; Cohen et al. 2009; Geer 1988; Norrander 1989; Sides et al. 2020), but as I discuss below, the quantifiable impact of primary voters' policy preferences in government is a startlingly open question. Several existing studies have examined other aspects of representation through House primaries, such as the introduction of the direct primary (Ansolabehere et al. 2010), how candidates position themselves in response to the presence or threat of primary challenges (Brady, Han, and Pope 2007; Burden 2004; Hirano et al. 2010), and how primary nomination rules affect elite polarization (Hirano et al. 2010; McGhee et al. 2014; Rogowski and Langella 2015). Though these studies address interesting

aspects of electoral representation and party competition, they cannot speak directly to the influence of voter's policy preferences on (1) the positioning of House primary candidates and (2) the outcomes of House primary elections.

The absence of voter preferences from the empirical study of primaries is troubling because they play a crucial role in the dominant theory that relates representation to primary politics. Although the Downsian model of candidate positioning explains the incentives for candidates to stake out moderate policy positions to cater to the ideological "median voter" (Downs 1957), candidates behave differently in the real world. Instead, candidates engage in highly partisan behavior and take divergent issue stances even on salient local issues and in closely competitive districts (Ansolabehere, Snyder, and Stewart 2001; Fowler and Hall 2016). But why? Scholars and political observers have argued that because competing in the general election requires each candidate to clinch their party's nomination contest, these candidates face a combination of convergence-promoting and divergence-promoting incentives. Primary elections tend to be dominated by partisan voters who are more attentive to politics, hold more non-centrist issue preferences, and "weight" candidates' issue positions more heavily than the average voter in the general election.[1] As a result, the risk that a candidate is defeated in the primary for being too moderate may outweigh the risk of losing the general election for being too partisan. The conflicting incentives imposed by partisan constituency and the general election constituency creates a "strategic-positioning dilemma" that leads candidates to take divergent issue stances rather than targeting a district median voter (Aldrich 1983; Brady, Han, and Pope 2007; Burden 2001; Hill 2015).

The strategic positioning dilemma (SPD) is a central theoretical feature of this project,

[1]Primary elections are not *entirely* partisan affairs. States vary in their regulations that primaries be "closed" to partisan voters only, that voters must preregister with their preferred party to vote in the primary, and even whether primaries are partisan at all (see McGhee et al. 2014 for a thorough and contemporary review of these regulations). Although many observers suspect that regulations on primary openness greatly influence the ideological extremity of the primary electorate, recent survey research finds that these regulations do little to affect the policy preferences of primary voters on average (Hill 2015).

and tests of the SPD are key empirical contributions in the following chapters. The sections that follow introduce key terms for understanding my critique of the existing research and my contribution to it in this project.

1.1.1 Key concept: policy ideology

If we had an ideal test of the SPD's implications, the policy preferences of partisan primary voters are an essential ingredient. Primary voters are one of the key constituencies that a candidate must please in the SPD view of primary elections. When partisan voters in a district are more conservative, the SPD claims that the candidate experiences a pressure to stake out a more conservative campaign position, especially in the primary. This section briefly discusses this project's terminology around voter ideology, the groups in the electorate for whom these concepts are at play, and how relate to other political science research.

When this project discusses voter "preferences" or voter "ideology," it specifically refers to a notion of *policy ideology*. An individual's policy ideology is a summary of their policy views in a left–right ideological space. Policy views are naturally complex and multidimensional, and it is possible for individuals to hold beliefs across policy areas that would strike many political scientists as being "ideologically inconsistent" (e.g. Campbell et al. 1960). Policy ideology distills this complexity into average tendencies; voters who hold a greater number of progressive preferences about policy are more ideologically progressive, and vice versa for voters with more conservative policy preferences. Voters who hold a mixture of progressive and conservative beliefs are ideologically moderate.

Policy ideology is different from policy *mood*, since mood measures voter preferences for the government to do more or less than an ever-shifting baseline, while ideology meant to be directly comparable using only issue information (Enns and Koch 2013; McGann 2014; Stimson 1991). Policy ideology is thus a similar concept to any method that measures a hidden ideological summary from one-off issue-based stimuli. This includes ideal point scores

for members of Congress, Supreme Court justices, and even individual citizens (Clinton, Jackman, and Rivers 2004; Martin and Quinn 2002; Poole and Rosenthal 1997; Tausanovitch and Warshaw 2013; Treier and Hillygus 2009). Other researchers have called this concept "policy liberalism" (Caughey and Warshaw 2015), which orients the concept so that "larger" values represent "more liberalism." For this project, I prefer to orient the construct as policy *conservatism*, which orients a scale so that larger/more conservative values correspond to "rightward" movements on a number line. I try to be conscious of the difference between *consistent* issue beliefs and *extreme* issue beliefs throughout this project. Consistently conservative issue beliefs do not necessarily imply that an actor is "extremely" conservative (Fiorina, Abrams, and Pope 2005b), and an actor may appear "moderate" even if they hold a mixture of non-moderate progressive and conservative issue beliefs (Broockman 2016).

This project views policy ideology in a measurement modeling context, which we return to in Chapter 2. Policy ideology affects voters' issue beliefs, and while issue beliefs can be measured using a survey, policy ideology itself is not observable. Instead, policy ideology exists in a latent space, an survey items on specific issues reveal only limited information about voters' locations in the latent space. This is different from summarizing policy views by adding or averaging policy responses, which implicitly assumes that all items about all issues are equally informative about ideology. Modern measurement approaches relax this assumption, instead viewing survey items as sources of correlated measurement error across respondents, leading to more careful modeling approaches for estimating a latent signal from noisy survey data (Ansolabehere, Rodden, and Snyder 2008). Following this modeling tradition, I refer to an individual's location in policy-ideological space as their "ideal point," the point at which their expected utility of a policy is maximized with respect to their ideological preferences.

1.1.2 Key concept: district-party groups

I argue that another key construct at work in the SPD is the notion of *groups* in the electorate. For a given district, the general election is a contest among all voters, so we consider this constituency as a group. We sometimes refer to this group as the "general election constituency," since it contains anybody who is eligible to vote in the general election. It does not specifically refer to voters only, but contains any citizen who could potentially be a voter in the general election. This ambiguity of who among the general election constituency actually votes is important to understanding a candidate's incentives during the campaign, since the candidate is uncertain whether certain campaign tactics will galvanize some constituents while alienating others.

Another important grouping for this the partisan constituency within a district. Each congressional district contains constituents who are aligned with the Democratic Party or the Republican Party. I call these two groups of constituents *district-party groups*. All 435 congressional districts contain voters from the two major parties, totaling 870 district-party groups. For brevity, I sometimes refer to district-party groups as "party groups" or "partisan groups." A district-party group contains any voting-eligible citizen who resides in a given district and identifies with a given party. As with the general election constituency, membership in a party group is no guarantee that the constituent votes either in the primary or in the general election. The important fact is that they are nominally aligned with one party's voter base over the other. As I discuss below, decomposing a district's voters into separate party groups is the key theoretical innovation in this project. To the best of my knowledge, an empirical study of primary representation that decomposes the voter preferences into district-party groups has never been done, even though it is crucial for testing the implications of the SPD theory.

One important distinction about district-party groups is that they are made of constituents,

not organizations. For this reason, it is sometimes helpful to refer to district-party groups as district party "publics," which emphasizes that the groups are composed of ordinary citizens (Caughey and Warshaw 2018). There is no formal registration requirement to be a member of a party group, only a partisan identification. This construction of district-party publics aligns most closely with Key's "party in the electorate" rather than "party as organization" (Key 1955). This distinguishes party publics from interest groups, policy groups, "intense policy demanders," or the "extended party network," which are concepts that describe organizations or maneuvers by political elites rather than rank-and-file constituents (Cohen et al. 2009; Koger, Masket, and Noel 2009; Masket 2009). Although recent research has underscored the importance of elite actors in shaping party nominations, this project focuses specifically on testing the SPD, which is a voter-centric view of primary representation. We bring in important concepts from elite-driven stories of primaries as they apply to particular claims being tested in later chapters.

1.1.3 Key concept: district-party ideology

It is important to define both "policy ideology" and "district-party publics" because they combine to form a key concept that anchors the substantive contributions of this project. This concept is *district-party ideology*: policy ideology aggregated to the level of the district-party group. Just as any individual might have a policy ideology ideal point, and any individual might affiliate with a party, district-party ideology averages the ideological variation within a district-party group into one group-level ideal point. By aggregating policy ideology within groups in this way, this project summarizes how policy ideology differs between Democrats and Republicans in the same district, and it shows how Democratic and Republican party groups vary across congressional districts. This enables us to consider how candidates are responsive to partisan sub-constituencies that together make up a shared general election constituencies (see also Clinton 2006).

1.1.4 Key concept: candidate campaign positioning

As with individual voters, we can imagine that candidates for Congress have campaign platforms, or at least promises and stated issue positions, that are located in ideological space as well. The study of United States politics most commonly places elite political actors in ideological space using their voting records, including members of Congress, Supreme Court justices, federal judges, and state legislators (Clinton, Jackman, and Rivers 2004; Epstein et al. 2007; Martin and Quinn 2002; Poole and Rosenthal 1997; Shor and McCarty 2011a). Researchers have extended the modeling intuitions to estimate ideal points from unconventional sources of data, including surveys of congressional candidates, campaign finance transactions, interest group ratings, text from political advertisements, and even Twitter activity (Ansolabehere, Snyder, and Stewart 2001; Barberá 2015; Bonica 2013; Burden 2004; Burden, Caldeira, and Groseclose 2000; Henderson 2016).

This project is interested in the ideological locations of candidates for office as measured through their campaigns. The positioning of campaigns is more directly related to the strategic positioning dilemma than any other concepts that we might scale in ideological space: candidates compete against one another by positioning themselves to appeal to a partisan base of voters, and partisan constituents consult use these campaign positions to nominate the candidate of their liking. To be sure, campaign positions are influenced by other activities that researchers have used to scale candidates for office. Incumbent legislators cast votes to form a defensible record in office, for instance, which both bolsters and constrains their campaign messages (Canes-Wrone, Brady, and Cogan 2002; Mayhew 1974). Not every primary candidate has a roll-call voting record to compare, however, so this project requires an ideal point measure that places incumbents, candidates challenging incumbents, and candidates running for open seats in a comparable ideological space.

This project measures primary candidates' campaign positioning using CF scores from

Bonica's (2019b) *Database on Ideology, Money in Politics, and Elections* (DIME) database. CF scores use campaign contributions to measure the political ideologies of contributors and recipients of campaign contributions. Because a wide variety of political actors engage in the contribution and receipt of campaign funds, the DIME contains CF score estimates for political candidates, party organizations, PACs, and individual donors. Unlike interest group ratings, another source of ideology scores for non-incumbent political candidates, CF scores are not constructed with a political agenda that implicitly "weights" issues according to the interest group's priorities (Fowler 1982; Snyder Jr 1992). CF scores assume that a donor makes financial contributions to political actors to maximize their utility over all potential contribution they could make, where utility decreases with greater ideological distance between donors and candidates. CF scores are the estimated ideal points for contributors and recipients that maximize this utility (Bonica 2013, 2014). These scores have been used in other studies of primary candidate ideology by Thomsen (2014), Thomsen (2020), Rogowski and Langella (2015), Ahler, Citrin, and Lenz (2016), and Porter and Treul (2020), and similar donation-based ideal point measures by Hall and Snyder (2015) have been used by Hall (2015) and Hall and Thompson (2018). As I discuss in future chapters, CF score are not without controversy as indicators of elite ideology, especially when comparing members of the same party (Hill and Huber 2017; Tausanovitch and Warshaw 2017), but other research shows that donors differentiate moderate and ideological candidates within the same party (Barber, Canes-Wrone, and Thrower 2016), the ideology component of CF scores outperforms a party-only model of giving (Bonica 2014), and CF scores predict future votes by members of Congress to a similar degree of accuracy as roll-call based scores do (Bonica 2019a).

1.1.5 The strategic positioning dilemma, implications, and research questions

Now that we have defined some key terms, we can see how they relate to previous research on the strategic positioning dilemma. The theory states that candidates balance two competing

constituencies during their campaign for office. Candidates face incentives to cater to the median voter in the general election, but they do not progress to the general election without first catering to partisan voters in the primary election. As a result, their campaign position is tailored to split the difference between the two constituencies, perhaps leaning more to the partisan base in safe districts and to the median voter in competitive districts. This section unpacks this intuition in detail and argues that existing research does not test the key claims.

First, how does district-party ideology affect the way candidates position themselves in a campaign? The logic of the SPD suggests that, at minimum, district-party conservatism should be positively correlated to the conservatism of a candidate's campaign position. At maximum, more conservative partisan voters exert a positive causal effect on the conservatism of a candidate's campaign position. This implies that candidates can perceive the conservatism of their partisan constituents, reflecting the relative variation in actual constituents' views if not the absolute level (Broockman and Skovron 2018).

Second, if candidates anticipate partisan voters' policy views and position themselves accordingly, this suggests that candidates believe partisan voters are capable of voting in accordance with their policy views. If this is true, we should expect that district-party groups that are more conservative should be more likely to nominate conservative nominees in primary elections.

These two predictions are the core empirical implications of the "strategic positioning dilemma" theory of representation in primaries. Crucially, testing each prediction requires a researcher to observe the policy ideologies of partisan constituents within a district, which is a separate group from the general election constituency or the location of the median voter. This project argues that district-party policy preferences are either absent from existing research or thoroughly misconstrued—an important theoretical and methodological point that I unpack in Section 1.2.3. As a result, U.S. elections research has been unable to empirically evaluate a widely held theory of representation in primaries.

Stated differently, this dissertation asks if primaries "work" the way the SPD claims they do. It is widely believed that primaries are effective means for voters to inject their sincere preferences into the selection of candidates and, in turn, the priorities of elected officials. Is this *actually* true? The two empirical research questions underlying this project are:

1. Do candidates position themselves to win the favor of primary voters?
2. Do primary voters select the candidate who best represents their issue beliefs?

1.2 Does the Strategic Positioning Dilemma Describe Primary Representation?

1.2.1 Theoretical concerns

The strategic positioning dilemma view of U.S. primaries has reasonable intuitions, but there are reasons to doubt some of its theoretical premises. First, the SPD is put forth as a theory to explain divergent candidate platforms across parties, but there are numerous theories that explain candidate divergence that do not rely on bottom-up pressures from primary voters. And second, the SPD requires voters and candidates to be highly sophisticated actors. Candidates must be capable of perceiving the relative extremity of their constituents, and voters capable of learning about candidate platforms, differentiating between candidates, and acting on sincerely-held preferences over candidate platforms.

The notion of the SPD emerges from a clash between idealized candidate positioning in formal models and the candidate positioning we observe in the real world. Classic formal models highlight a strategic logic for candidates to position themselves by "converging" to the location of the median voter: if constituents vote primarily with policy-based or ideological considerations, then candidates maximize the probability of electoral victory by positioning

themselves as closely to the median constituent as possible (Black 1948; Downs 1957).[2]

Empirical work finds evidence in partial support of both convergent and divergent candidate incentives. Candidates who run in electorally competitive districts are more moderate than co-partisans who are running in districts that run in electorally "safe" districts (Ansolabehere, Snyder, and Stewart 2001; Burden 2004), and even candidates who run in safe districts are marginally rewarded for taking more moderate issue positions than a typical party member would (Canes-Wrone, Brady, and Cogan 2002). Extremist candidates, meanwhile, earn fewer votes and are less likely to win in Congressional elections, and this tendency is stronger in competitive districts than in safe districts (Hall 2015). Despite these incentives to take moderate campaign positions, candidates nonetheless take divergent rather than convergent stances by and large. Republican and Democratic members of Congress vote very differently from one another, and this partisan divergence increased in recent years (McCarty and Poole 2006; Poole and Rosenthal 1997). The difference in legislative voting behavior across parties does not arise simply because Republicans and Democrats represent different districts, since Republicans and Democrats who represent similar districts (or the same state, in the case of U.S. Senators) nonetheless vote differently from one another (Brunell 2006; Brunell, Grofman, and Merrill 2016; McCarty, Poole, and Rosenthal 2009). Even among Congressional races in the exact same district, there is a sizable gap between Republican and Democratic candidate positions (Rogowski and Langella 2015). And although qualitative evidence from decades past suggests that candidates take careful positions on issues of local concern (Fenno 1978), more recent systematic tests find mixed evidence of localized, particu-

[2]Some empirical studies of candidate positioning (e.g. Ansolabehere, Snyder, and Stewart 2001; Brady, Han, and Pope 2007) claim that these formal models "predict" candidate convergence at the median voter. In my opinion, this misrepresents the formal work. Downs (1957) in particular explains the logic of candidate convergence, but he also explores many circumstances that would prevent the convergent equilibrium from appearing in the real world. This is important to clarify because, although it is common to describe candidate convergence as a "Downsian result" or a "Downsian prediction," we should recognize that the convergent equilibrium is an oversimplification. Understanding the theoretical incentives that promote candidate moderation is more important than the whether we observe perfect candidate convergence empirically.

laristic position-taking. (Canes-Wrone, Minozzi, and Reveley 2011; Fowler and Hall 2016). In total, even though there is some evidence that candidates benefit by positioning themselves as marginally more moderate or more in line with local public opinion, the dominant finding is that candidates take divergent positions that are more closely aligned with a national party platform than with a set of local issue priorities.

The Downsian logic is a strong "centripetal" force that promotes moderation among candidates, but what "centrifugal" forces explain the non-moderate stances (Cox 1990)? Political scientists have explored several theories whose underlying mechanisms are distinct from the SPD notion of competing constituencies. Parties are interested in cultivating long-term reputations for pursuing certain policy priorities (Downs 1957; Stokes 1963). It benefits both major parties for these reputations to be distinct from one another, since parties have office-seeking motivations to mutually divide districts into geographic bases that tend to support one party platform consistently over time (Snyder 1994). Party leaders maintain these party reputations by constructing brand-consistent legislative agendas and pressuring legislators to support reputation-boosting legislation (Butler and Powell 2014; Cox and McCubbins 2005; Lebo, McGlynn, and Koger 2007). In turn, non-median party platforms are more appealing to constituents with ideologically consistent issue beliefs. Candidates benefit by rewarding these constituents in particular because they are more likely to be influence election outcomes in favor of the candidate (Hirano and Ting 2015). These voters are more likely to turn out in general elections than moderate voters are, so it is more efficient for candidates to cater to these constituents. Partisan constituents are also more likely to engage in pro-party activism, such as staffing campaigns, contributing financially to campaigns, and attending party conventions (Aldrich 1983; Barber 2016; La Raja and Schaffner 2015; Layman et al. 2010; McClosky, Hoffmann, and O'Hara 1960).

These incentives for candidates to diverge from median positions are possible without considering primary elections whatsoever. Even if we introduce primary elections into the

theoretical story, many plausible explanations for divergence do not rely on outward pressures from ideological primary voters either. Many scholars of political parties maintain that parties retained their gatekeeping roles over party nominations even as the direct primary ostensibly removed their formal powers over candidate selection. Although primary campaigns take place, these scholars argue that an informal network of party actors wields enormous influence behind the scenes, controlling which candidates obtain access to the party's resources, donor lists, and partisan campaign labor (Cohen et al. 2009; Masket 2009). Through these mechanisms, candidates can live or die by the nomination process long before primary *voters* ever enter the picture.

One reason to doubt the SPD on theoretical grounds is that it has high demands of voter sophistication in primary elections. It is well understood that learning about the characteristics and issue positions of political candidates is costly for voters, particularly in non-presidential elections. Party labels on the ballot are valuable heuristics for voters to differentiate the issue positions of Republican and Democratic candidates likely hold (Hill 2015). Primary elections, however, occur most of the time between candidates in the same party,[3] which denies voters' the informational shortcut of a candidate's party affiliation (Norrander 1989). Primary elections often occur during months when voters are paying less attention to politics, and the press cover primary campaigns less closely than general election campaigns. Primary voters have a reputation for being more attentive and sophisticated consumers of political information, but in these lower-information environments, they may cast their ballots for non-policy reasons by prioritizing "Washington outsiders" or identity-based candidate features such as gender or race (Porter and Treul 2020; Thomsen 2020). They may also vote for the familiar candidate instead of the ideologically proximate one, in which case asymmetric campaign expenditures or news coverage may advantage one candidate over the other.

[3]There are a few exceptions to this institutional configuration of intra-party nominations. Some states hold blanket primaries, top-two primaries, or "jungle" primaries, where candidates from all parties compete on one ballot to be included in a runoff general election.

For example, Bonica (2020) attributes lawyers' numerical prominence in Congress to their ability to raise early money from their wealthy social networks. Furthermore, despite the disproportionate news coverage received by primary candidates who challenge incumbents on ideological grounds, the absolute number of explicitly ideological primary challenges in a given election cycle is low (Boatright 2013), so primary voters are unlikely to experience a deluge of policy-focused campaign messages even if they are attentive and sophisticated to receive and process those messages. In short, the claim that voters' policy preferences affect their choices in primary campaigns sounds straightforward, but the information environment of primary campaigns makes it difficult for constituents to vote foremost with their policy ideologies.

The SPD also requires candidates to perceive the policy ideologies of their partisan constituencies accurately in order to position their candidacies in relation to the partisan base and the median voter. Broockman and Skovron (2018) lend contradictory evidence to this notion by measuring the degree to which politicians "misperceive" their constituency's policy views. The authors find that elected politicians believe that their constituents are much more conservative on many issues than they actually are, which could affect how accurately candidates position themselves in relation to constituent views.

1.2.2 Empirical ambiguity

Empirical support for the strategic positioning dilemma is as unclear as the theoretical underpinning. When researchers conduct empirical tests of the SPD or the narrower premises of primary representation and competition on which it rests, the results are ambiguous and often contradictory of the SPD story. This section reviews existing research in this area to review the outstanding questions and preview the substantive innovations in this project.

Much of the interest in primary elections and representation comes from a focus on candidate divergence and partisan polarization. Why do candidates who stand for general

election take divergent stances from one another, and do the competitive dynamics of primary elections increase this divergence? Prominent studies of candidate positioning in general elections initially found conflicting evidence about the influence of stiff primary competition on candidate extremity. Using survey data from congressional candidates during the 2000 campaign, Burden (2004) finds that general election candidates take more extreme policy positions in their campaigns if they also faced stronger primary competition. This makes sense especially if primary candidates care more about the candidate's ideological positioning than general election voters do, the latter of whom are also receptive to non-policy appeals. Ansolabehere, Snyder, and Stewart (2001) find the reverse pattern using 1996 survey data. The gap between major party candidates was actually smaller when one of the candidates faced stiffer primary competition. This counter-intuitive finding makes sense if the presence of a primary challenger is itself a consequence of candidate positioning. If an incumbent maintains a partisan reputation, this may fend off credible primary challengers who have less room to wage an ideological campaign against the incumbent. As a result, the *threat* of a primary challenge exerts a centrifugal force on candidate positioning, even if a primary challenger never actually appears (Hacker and Pierson 2005). Hirano et al. (2010) study this threat-based hypothesis by measuring potential primary threat as the average presence of primary competitors in down-ballot races. In district with high levels of latent primary threat, we might expect the incumbent to take more extreme stances in Congress. Although the idea that incumbents vote as party faithfuls to preempt opportunistic challengers is intuitive and supported by other research (e.g. Mann 1978), this measure was not meaningfully related to the extremity of an incumbent's voting record in Congress (Hirano et al. 2010). In short, the evidence of the polarizing effects of primary challenges is mixed and unclear.

Researchers interested in the polarizing effects of primaries on candidates and legislators has also examined primary "rules." Political parties are private organizations, and nominees are intended to represent the parties' priorities and governing values, but participation in

primary elections is not always restricted to party members only. Primary "openness" rules that govern who can participate in a partisan primary are managed by state election law, with some allowances for parties to set rules within those limits. States with "closed" primaries restrict participation in primaries only to individuals who are registered as Republicans or Democrats in their state registration records. States that allow third-party or non-partisan voters to participate in partisan primaries are "partially" open, and states where any voter can participate in any primary are regarded as "open" primaries. I discuss finer details of primary rules in later chapters. Researchers seeking to exploit state-level variation in primary rules hypothesize that states with more restrictive participation criteria might select more ideologically extreme primary nominees, and states with more relaxed rules might select relatively moderate nominees. This is because primary voters are commonly believed to hold more ideologically consistent policy views than other constituents, so candidate polarization will respond to the polarization among the voting public (Jacobson 2012). However, the consensus among recent studies finds little evidence supporting the hypothesis that primary rules affect polarization in congress or candidate divergence more broadly. This is because there is little consensus in public opinion research that partisans who participate in primaries are much different from partisans who do not participate in primaries, either demographically or ideologically (Geer 1988; Hill 2015; Jacobson 2012; Norrander 1989; Sides et al. 2020), though these studies cover many years, and the dynamics of primary voting might have changed. And even recent studies that find that primary voters hold more ideologically consistent views find no evidence that closed primaries nominate candidates that are more ideologically off-center (Hill 2015). This finding appears to hold for the House, Senate, and state legislatures through the past several decades (Hirano et al. 2010; McGhee et al. 2014; Rogowski and Langella 2015). Even reforms that drastically change the primary rules, such as California's recent shift to a blanket primary where candidates in all parties compete for the same limited number of positions on the general election ballot, do not nominate legislators

whose voting records are much more moderate than before (Bullock and Clinton 2011).

These studies are incomplete in important ways that bear on the key substantive questions underlying this project. Most of these studies evaluate primaries' effects on representation by examining roll-call votes only. Since roll-call votes are only observable for incumbents, many of these analyses cannot measure candidate *divergence* because they cannot compare incumbents to non-incumbents nor two open-seat candidates. Some notable studies examine non-incumbent candidates for general election using candidate surveys (Ansolabehere, Snyder, and Stewart 2001; Burden 2004), but these studies are also limited because they do not observe the positions of candidates who lose the primary nomination. Without observing primary losers, we have no way of knowing if the general election candidate was relatively moderate or ideological in comparison to other primary candidates. It is much rarer for a study to measure primary candidate positioning as the key outcome variable using a method that covers incumbents, challengers, and open-seat candidates (Rogowski and Langella 2015).

1.2.3 Vote shares do not identify policy ideology

Another important drawback of the existing research on primaries and ideological representation is the way these studies handle voters' policy preferences. The strategic positioning dilemma pits two constituencies in a district against each other: the nominating constituency (district-party group) that contains constituents from one party's base, and the general election constituency that contains constituents from both major parties and with no party affiliation. The former is theorized to prefer ideologically faithful candidates who adhere closely to a partisan policy platform, while the latter prefers moderate candidates in the general election. Studies routinely acknowledge this distinction in theory, but they often abandon the distinction between the two groups in applied studies, instead operationalizing the preferences of all three constituencies—the general constituency and two partisan primary constituencies—using the same measure: the district-level presidential vote.

This project argues that the presidential vote is not a suitable for the study of primary representation for the simple reason that votes are not equivalent to policy preferences or policy ideology. Votes are choices that voters make under constraints, namely, the distance between the voter and the presidential candidates. Even in simple models where ideology is the only factor influencing vote choice, observing a voter's choice of candidate contains very little information about their ideological location. In the aggregate, Republican voters in a district may be ideological moderates or ideological conservatives, and the fact that they vote Republican does not inform us on the ideological distribution of Republican voters. Similarly, a district's vote outcome captures how all of its constituents vote *on average*, but because partisans tend to vote foremost for their preferred party even in the face of strong policy disagreements with the candidate (e.g. Barber and Pope 2019), aggregate vote shares for a district could easily be more affected by the *number* of Republicans and Democrats in a district rather than the exact location of their ideological preferences. Using the terminology by Tomz and Van Houweling (2008), studying vote shares rarely presents a "critical test" of theories of voting because the same observable vote outcome can arise from many underlying voter preference configurations.

Stated differently, the observed vote share in a district does not uniquely identify any important features of the underlying preferences of voters. Figure 1.1 demonstrates the problem using a simple theoretical model of ideological voting for president. We begin by demonstrating the the basic mechanics of the scenario in the two left-side panels. In this scenario, we consider one congressional district that contains many constituents. Every constituent has a policy ideal point represented on the real number line, with larger values indicating greater policy conservatism. Every constituent also identifies with either the Republican Party or the Democratic Party. The top-left panel breaks voters into Democratic and Republican Party affiliations and shows the probability distribution of ideal points within each partisan base, which in this example are both Normal distributions with a scale of 1.

Republican-identifying constituents hold policy preferences that are more conservative than Democratic constituents on average: the median Republican and Democrat are respectively located at 1 and -1.[4] There is enough within-party variation that some Democratic constituents are more conservative than some Republican constituents, despite their party affiliation. The bottom-left panel combines the two partisan distributions into one distribution for the entire constituency. We assume at first that both partisan constituencies are equally sized, so the composite distribution is a simple finite mixture of the two distributions.[5] The midpoint between two presidential candidates is shown at policy location 0. Assuming all constituents vote according to single-peaked and symmetric utility functions over policy space, constituents are indifferent between candidates if they have ideal points equal to 0, vote for the Democratic candidate if they have ideal points less than 0 (shown in darker gray), and vote for the Republican candidate if they have ideal points greater than 0 (shown in lighter gray). The aggregate election result, therefore, is equal to the cumulative distribution function of the combined distribution evaluated at the candidate midpoint. In the bottom-left panel, the vote share for the Democrat is 50%, with some Democrats voting for the Republican candidate, and some Republicans voting for the Democratic candidate.

The panels on the right side of Figure 1.1 show how slight changes to one party's preference distribution affects the aggregate distribution of preferences in the combined constituency and, as a result, the presidential vote share in the district. The composite distribution is again shown in gray, with dark and light shades indicating vote choice as in the bottom-left panel. The underlying partisan distributions are outlined only with red and blue lines to reduce

[4] Because these are Normal distributions, the median and the mean are equivalent. I refer to the median instead of the mean because medians are more directly relevant to spatial models of voting.

[5] Analytically, if $f_p(x)$ is the probability density of ideal points x in party p, then the composite density $f_m(x)$ is a weighted sum of the component densities: $f_m(x) = \sum_p w_p f_p(x)$, where w_p is a mixture weight representing the proportion of the total distribution contributed by party p, with weights constrained to sum to 1. In this first example, both partisan constituencies are equally populous, so both parties have weight $w_p = \frac{1}{2}$. If parties had different population sizes within the same district, w_p would take values in proportion to those population sizes.

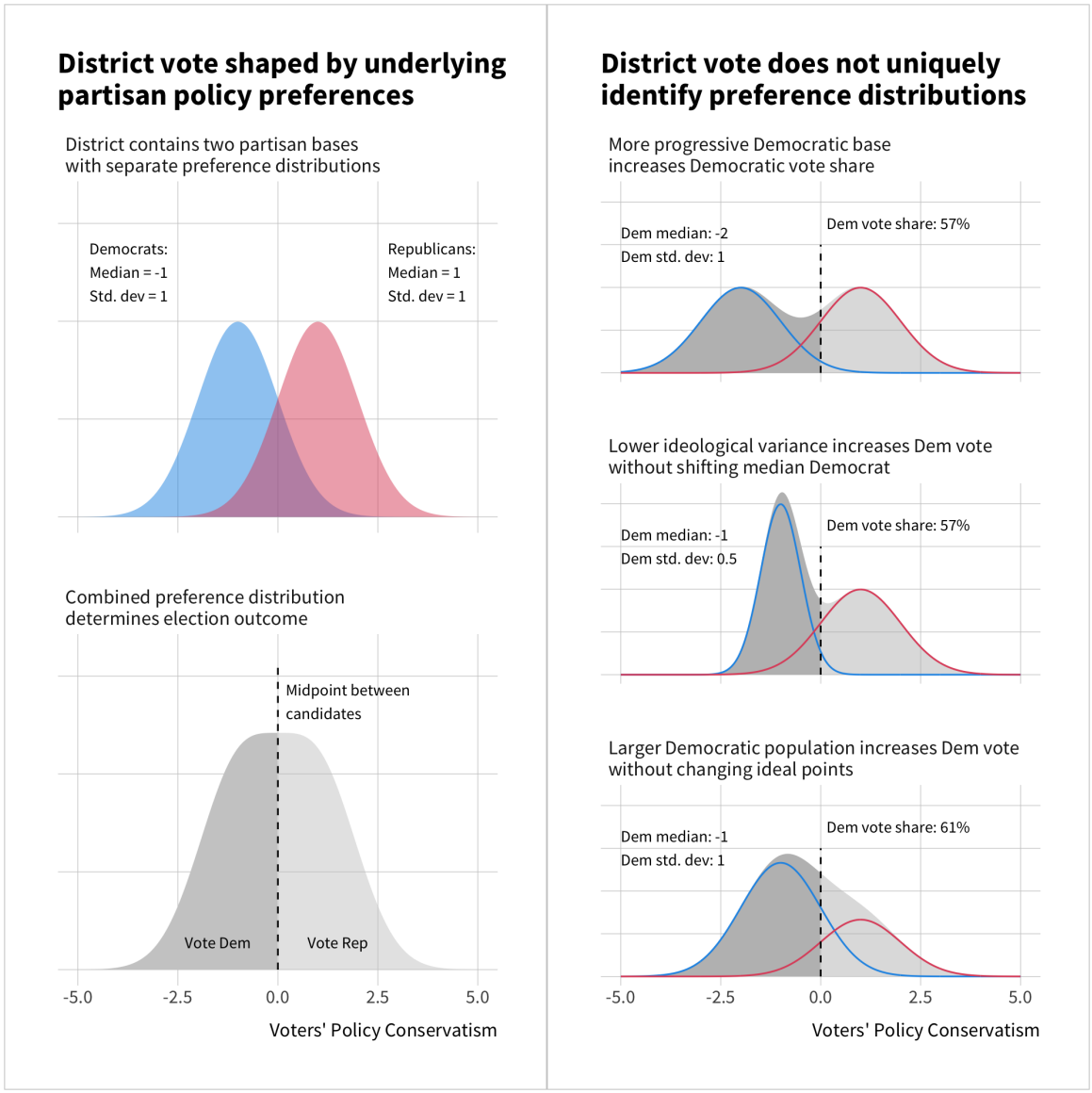


Figure 1.1: Demonstrating how district vote shares from a single election are insufficient to identify underlying policy-ideological features of the district. The left side shows how the policy preference distributions for two parties in a district (top panel) combine to form an aggregate preference distribution for the district as a whole (bottom panel). The right side shows how the Democratic vote share is affected by changes to either the locations, the scales, or the population sizes of the underlying partisan distributions.

visual clutter. The modifications to the underlying partisan preferences are simple, but even these simple changes reveal the fundamental problem with using district voting as a proxy for policy ideology in the voting population. In each panel, I intervene on only one feature of the Democratic Party ideal point distribution, leaving the Republican distribution untouched (median of 1, standard deviation of 1). Intervening on just one component of one party's distribution is meant to keep the demonstration simple, bearing in mind that the problem is much more complex in the real world, where we can imagine multiple simultaneous changes to both parties at once. The interventions highlight two classes of problems. First, we can perform multiple modifications of the underlying partisan distribution that obtain the same aggregate vote share. This proves that the district vote does not uniquely identify the characteristics of the underlying voter distributions. And second, we can alter the district vote outcome by changing party *sizes* without any change to the ideal point distribution within either party. This proves that vote shares may vary across districts even if partisan ideal points distributions are the same.

In the top-right panel, I shift the location of the Democratic ideal point distribution to the left, from a median of -1 to -2. This location shift results in a greater number of Democratic constituents with ideal points left of the candidate midpoint, increasing the Democratic vote share in the district from 50% to 57%. In the middle-right panel, I shrink the scale of the Democratic ideal point distribution from a standard deviation of 1 to a standard deviation of 0.5. Lower ideal point variance within the Democratic base has the exact same effect on the vote as shifting the location: more Democratic voters left of the midpoint, which increases the Democratic vote share to 57%. This means that compared to a district with a 50% presidential vote split, we would not be able to attribute the increased Democratic vote to a constituency that is *more progressive on average* (location) or simply *less heterogeneous* in its policy preferences (scale). The bottom-right panel in the figure shows how we obtain a different district vote without changing the underlying ideological distribution in either

party whatsoever, instead changing only the relative population size of each partisan base. The Democratic base in the final panel is unchanged compare to the original distribution laid out in the top-left: median of -1 and standard deviation of 1. The only difference is that the district contains an unequal balance of partisan voters, two Democratic constituents to every one Republican constituent. This results in an increased Democratic vote from 50% to 61%—ironically, the largest impact on the overall district vote despite not changing the ideological distribution of either party.

To review the lessons of Figure 1.1, observing a Democratic vote share greater than 50% reveals very little about the underlying distribution of voters. In every panel, we observe an increase in the Democratic vote compared to our baseline scenario, but the the median voter in either party does not need to change in order for vote shares to be affected. Since the Republican distribution is identical in every panel, inferring that Republicans are less conservative in districts with greater Democratic voting would be incorrect in every case. For the Democratic constituents, inferring a more progressive Democratic median voter from greater Democratic voting would be wrong in two of the three cases.

It is worth repeating that the scenario laid out in Figure 1.1 is a vast oversimplification of the real electorate. This is intentional, as it shows how intractable the problem becomes even in an artificial setting where we can take many variables as given. This scenario contains no complicating elements such as non-partisan or third-party identifiers, non-policy voting, random sources of utility or utility function heterogeneity across different voters, differential turnout between partisan bases, and so on, that we might incorporate directly into a formal model. It also does not take into account the inconveniences of real election data, where short-term forces impose additional shocks to vote shares that are unrelated to underlying voter preferences.

The conceptual difference between district vote shares and aggregate ideology appears in real data as well, as shown in Figure 1.2. The figure shows ideological self-placement responses

to the Cooperative Congressional Election Study (CCES) as an approximate measure of a citizen policy ideology. I calculate the average self-placement for all respondents in each congressional district, as well as the average self-placement of Republican and Democratic identifiers as separate subgroups within each district. The first two panels use 2018 data to show that the district vote captures variation in ideological self-placement reasonably well when examining congressional districts as a whole, but it does a poorer job capturing variation in self-placement within each party. The first panel shows that districts that voted more strongly for Democratic presidential candidate in 2016 were more liberal on average, and districts that voted more strongly for the Republican candidate in were more conservative, indicated by a positively sloped loess fit line. The middle panel shows that this pattern does not hold as strongly within parties. Among Republican identifiers within each district, a weaker but still positive relationship holds overall, with more conservative Republicans in districts that voted more Republican. Among Democratic identifiers, however, ideological self-placement is not as strongly related to aggregate voting, with a loess fit that is flatter and even negative at several points. The final panel of loess fits is included to show that this pattern appears in all CCES years and is not particular to 2018 CCES responses: a strong relationship between vote shares and self-placement *on average*, and weak or non-relationships within each party.

The substantive takeaway from Figure 1.2 is further evidence that we should doubt the use of aggregate voting in a district is a reliable proxy of ideological variation within partisan primaries. Because the presidential candidates are the same in each district in each year, we know that this mismatch isn't due to different candidates with different campaign positions in each district. Instead, the observed pattern suggests that any aggregate relationship between ideological self-placement and district voting is driven at least in part by the partisan *composition* of a district—more Republicans or more Democrats—rather than cross-district ideological variation within either party. As a result, studies that use the presidential vote

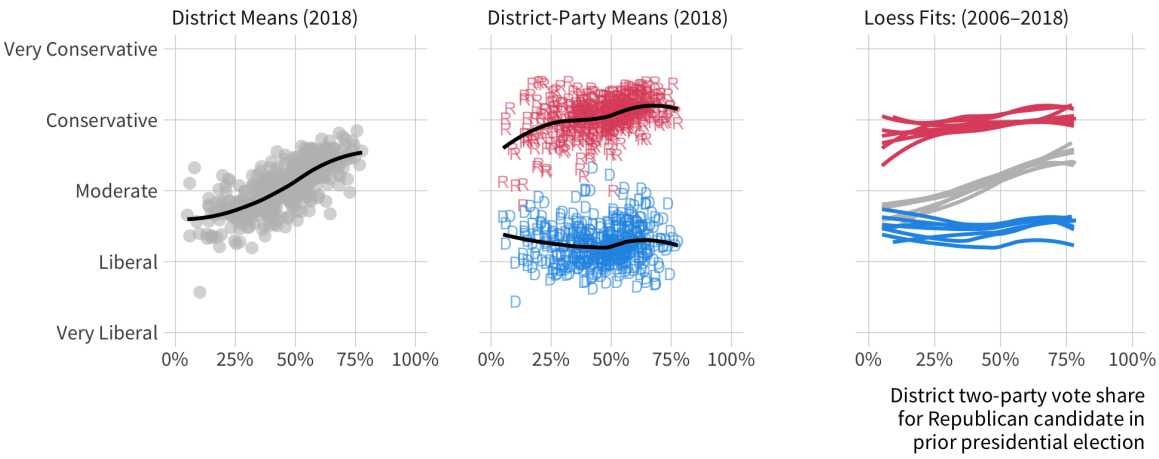


Figure 1.2: Average ideological self placement (vertical axis) and Republican vote share (horizontal axis) in all 435 congressional districts. Mean self-placement is calculated by numerically coding CCES ideological self-placement responses before averaging. The first panel plots average self-placement among all CCES respondents in each congressional districts. The middle panel breaks respondents in each congressional district into Republican and Democratic subgroups before averaging. The final panel plots loess fits for the same relationship measured over all CCES years.

to proxy within-party ideology may simply be measuring the *size* of a partisan group in a district instead of its ideological makeup.

Some researchers have recognized the identifiability problems with district presidential vote shares as a measure of district preferences. Levendusky, Pope, and Jackman (2008) specify a Bayesian structural model to subtract short-term forces on election results and isolate latent partisanship. Kernell (2009) formally proves that using a single election to cardinally place district ideal point medians is never possible, but that estimating the mean and variance of ideal point distributions is possible under distributional assumptions and a formal model of voting. Although these methods are promising innovations over the

common practice of using votes as a proxy for policy preferences, I have uncovered no studies of primary representation in the intervening years that have incorporated these methods. Furthermore, the methods estimate the median policy preference for a district as a whole. They do not describe separate partisan constituencies within a district, which is the essential missing ingredient.

I stress that this measurement problem is more than methodological nitpicking. The theoretical consequences are systemic. The literature's dependence on the presidential vote as a proxy for district preferences has prevented scholars from incorporating key theoretical constructs into empirical studies of primaries: the ideological preferences of partisan voters. Without serviceable measures of partisan policy preferences, we can say very little about the role of primary elections in the broader democratic order of U.S. politics. This affects our knowledges of topics beyond party nominations as well. To study how politicians weigh the opinions of various subconstituencies, which the study of U.S. politics is obviously interested in (Bartels 2009; Clinton 2006; Cohen et al. 2009; Fenno 1978; Gilens and Page 2014; Grossman and Hopkins 2016; Phillips 1995; Pitkin 1967), research must be able to measure the policy preferences of subconstituencies directly. The technology to estimate subconstituency preferences using survey data is admittedly quite new, and this district intends to continue this effort by extending existing models, highlighting important methodological considerations for model building and computation, and demonstrating how to use these measures for observational causal inference.

1.3 Project Outline and Contributions

1.3.1 Measuring district-party ideology

This chapter has so far identified a shortcoming in the study of primaries that subconstituency preferences are rarely measured. This project rectifies this shortcoming by measuring district-

party ideology for Republican and Democratic party groups in Chapter 2. This allows the project to carry out direct tests of SPD hypotheses that were previously impossible in Chapters 4 and 5.

I estimate district-party ideology this using an item response theory (IRT) approach to ideal point modeling. The model estimates the policy ideology for a typical Democrat and a typical Republican in each congressional district over time. I employ recent innovations in hierarchical modeling to measure individual traits at subnational units of aggregation using geographic and temporal smoothing (Caughey and Warshaw 2015; Lax and Phillips 2009; Pacheco 2011; Park, Gelman, and Bafumi 2004; Tausanovitch and Warshaw 2013; Warshaw and Rodden 2012). The model I build extends these technologies by specifying a more complete hierarchical structure for the bespoke parties-within-districts data context, a more flexible predictive model for geographic smoothing, and advances in Bayesian modeling best-practices from beyond the boundaries of political science (see also Section 1.3.5).

1.3.2 Empirical tests: how district-party ideology matters

After estimating the ideal point model for district-party groups, I apply these estimates in two critical tests of the strategic positioning dilemma.

Chapter 4 studies how district-party ideology affects candidate positioning in primary elections. If the primary constituency exerts a meaningful centrifugal force on candidate positioning, we should expect candidates with more ideological partisan constituencies to take more ideological stances, all else equal.

Chapter 5 studies how district-party ideology affects candidate selection in primary elections. If the primary constituency exerts a credible threat against candidates taking overly moderate campaign positions, we should expect more ideological constituencies to select more ideological candidates, all else equal.

An important institutional factor at play in each of these empirical settings is the mod-

erating effects of primary openness. Past studies have explored whether primaries that are closed to non-partisan and cross-partisan participation lead to the election of more extreme party nominees. District-party ideology is missing from these studies, but it matters for our theoretical expectations about the effects of primary rules. For instance, we should not expect a relatively partisan constituency to nominate an extremist candidate solely because the primary is closed to non-party members. Past studies have either ignored ideological variation across districts or used unsuitable proxy measures that do not measure district-party ideology. Including primary rules in Chapters 4 and 5 will provide a more faithful test of the primary rules hypothesis.

This project is not rooting for or against the veracity of the strategic positioning dilemma as a model of primary representation. The theory is intuitive and reasonable in its predictions for rational elite behavior, but its assumptions about voter competence and its empirical track record are less supportive of the theory. I wish for the empirical components of this project to be theory *testing* rather than advocacy for or against an idea in current political thought.

1.3.3 Causal inference with structural models

The strategic positioning dilemma is a story about the causal effects of district-party ideology on candidate positioning and candidate selection. Testing the theory requires a serious engagement with causal inference methods. Unfortunately, the observational data at work are difficult to manipulate in support of causal claims. District-party ideology is not randomly assigned, so we require methods for identifying unconfounded variation by design or adjusting for confounding with careful modeling.

One inherent limitation of the district-party ideology estimates is that they come from a measurement model. The measurement model smooths estimates with a hierarchical regression, where partial pooling improves the estimate for one unit "borrowing information" from other units. This shrinks estimates toward one another, imposing correlations between

estimates that share a common cause. To leverage exogenous variation for design-based causal inference, this variation would likely have to come predominantly through exogenous shocks to raw survey data, which is challenging to conceive of considering that many surveys must be pooled to achieve feasible estimates at the district-party level.

Given these data limitations, this project turns to causal identification through a conditional independence assumption (Rubin 2005), also known as "selection on observables." Although selection on observables is a common approach to quantitative research, many analyses are not careful about their modeling choices, controlling for variables that do not improve causal identification or using modeling approaches that impose fragile or implausible functional assumptions on the data. One guiding ethic for the methodological contributions in this project is to take observational causal modeling more seriously than the existing research on primary representation by setting up empirical analyses that aspire to do the following:

- clearly state the potential outcomes model that links treatments, outcomes, and confounders.
- clearly state the causal estimand implied by a causal structure.
- clearly state the assumptions required to identify estimands and how modeling approaches relate to identification assumptions.
- use modeling approaches that are flexible enough to absorb confounding effects without too much dependence on strict functional forms.

I hope to satisfy these aims by invoking more explicit causal models of potential outcomes (Rubin 2005) and using "structural causal models" (SCMs) to guide model specification choices (e.g. Pearl 1995). The SCM approach makes heavy use of causal diagrams, or "directed acyclic graphs" (DAGs), to visualize a causal structure and identify causal claims. Causal diagrams as heuristic devices for causal inference are not new to political science in general

(Gerring 2001), but combining causal diagrams with the formal exactitude of the current causal inference tradition is less common in political science. Furthermore, SCMs and causal diagrams are less common in the literature on primaries and representation, which has not been as explicit about causal assumptions and empirical designs, with some notable exceptions (Fowler and Hall 2016; Hall 2015).

This project's approach to causal inference has two stand-out contributions to the study of primary representation that would be impossible but for this approach. First, Chapter 4 contains a detailed discussion of the causal effect of district party ideology on candidate positioning *as mediated by* aggregate district partisanship. I lay out the causal structure in causal graphs, discuss identification assumptions required to estimate the causal quantity of interest, and implement a sequential-*g* modeling approach to estimate it (Acharya, Blackwell, and Sen 2016). Chapter 5 explores flexible modeling with machine learning (ML) as a way to reduce dependence on fragile model assumptions. The chapter discusses *regularization-induced confounding*, a statistical bias in a treatment effect estimate that arises when regularized estimators, such as those used in common ML methods, under-correct for strong confounding by injecting too much shrinkage into a statistical model. I show how to correct this bias using Neyman orthogonalization, a two-stage modeling approach that de-biases causal estimates by reparameterizing the structural causal model (Hahn et al. 2018). Regularization-induced confounding is a serious problem for high-dimensional causal inference, but it has been discussed almost nowhere in political science (Ratkovic 2019).

Selection on observables is a fragile assumption for causal identification, which leads many researchers to speak in "scientific euphemisms" about causality instead of invoking explicit causal language (Hernán 2018). I adopt the position that this "taboo against explicit causal inference" is harmful to the larger aims of a research program because it obscures the dependence of research findings on causal assumptions, whose transparency is essential for credible causal inference, and leads work to be misinterpreted by future audiences who tend to

interpret findings as causal regardless of author intent (Grosz, Rohrer, and Thoemmes 2020). No study will ever prove the existence of a causal effect. Researchers should be transparent about causal assumptions so that future readers and researchers have clearer ideas about how to improve previous work. As such, this work will invoke causal language, highlight identification assumptions, and discuss threats to identification assumptions openly.

1.3.4 Bayesian causal modeling

Another important methodological contribution in its Bayesian approach to causal inference. The key independent variable of interest, district-party ideology, is estimated using a Bayesian measurement model. It is not observed exactly, but it is estimated up to a probability distribution. Using those estimates in subsequent analysis requires some accounting for the uncertainty in those estimates. I do this by propagating the Bayesian framework from the measurement model forward into the causal models. Operationally, this is done by taking the posterior distribution from the measurement model and using it as a prior distribution in subsequent models, recovering a joint probability distribution that captures uncertainty in causal effects and its relationship to the uncertainty in the underlying data.

Although the Bayesian view of causal inference is not new (Rubin 1978), it appears almost nowhere in political science. Political scientists occasionally use Bayesian technology for analytical or computational convenience (e.g. Horiuchi, Imai, and Taniguchi 2007; Carlson 2020; Ornstein and Duck-Mayr 2020; Ratkovic and Tingley 2017a), but rarely are the epistemic contours of Bayesian analysis explicitly credited for adding value to a causal analysis (Green et al. 2016; in economics, see Meager 2019).

Chapter 3 explores a Bayesian approach to causal inference in political science at length. It lays out a probabilistic model of potential outcomes adapted from Rubin (1978) and discusses how to interpret causal inference research designs through a Bayesian updating framework. I give pragmatic guidance for thinking about priors and specifying Bayesian causal models,

and I demonstrate the modeling approaching by replicating and extending a few published analyses in political science, noting where the Bayesian approach leads to different conclusions and interpretations about the findings.

I apply Bayesian approaches to causal modeling in Chapters 4 and 5 by combining multistage models into one posterior distribution, which is natural for applied Bayesian modeling where causal effects can be summarized by marginalizing over "design-stage" parameters (Liao and Zigler 2020). Bayesian estimation is also valuable in Chapter 5 to quantify uncertainty in machine learning methods. This is done using a Bayesian neural network model, which automatically penalizes model complexity using prior distributions and quantifies treatment effect uncertainty in the posterior distribution (Beck, King, and Zeng 2004; MacKay 1992).

1.3.5 Bayesian best practices

Another important contribution of the modeling exercise is the detailed discussion of Bayesian modeling and computational implementation it contains. Classic Bayesian texts for political and social sciences are written for an outdated computational landscape where Metropolis-Hastings and Gibbs sampling algorithms were state-of-the-art estimation approaches (Gill 2014; Jackman 2009). Recent years have seen rapid progress in the development and understanding of Hamiltonian Monte Carlo algorithms, which are faster, more statistically reliable, and easier to diagnose (Betancourt 2017, 2019; Duane et al. 1987; Neal 2012), but they also require renewed attention to the way researchers specify and implement Bayesian models (Betancourt and Girolami 2015; Bürkner and others 2017; Carpenter et al. 2016). Furthermore, this new generation of applied Bayesian modeling has updated best practices for specifying priors, modeling workflow, and model evaluation that (to my knowledge) have no precedent in the current political science awareness (Betancourt 2018; Gabry et al. 2019; Gelman, Simpson, and Betancourt 2017; Lewandowski, Kurowicka, and Joe 2009; Vehtari,

Gelman, and Gabry 2017; Vehtari et al. 2020). One contribution of this project is to highlight the evolving landscape for Bayesian thinking and Bayesian workflow, which has not received its due attention as a new generation of political scientists explores Bayesian analysis.

— 2 —

Hierarchical IRT Model for District-Party Ideology

To study how partisan constituencies are represented in primary elections, we require a measure of the partisan constituency's policy preferences. This chapter presents the statistical model that I use to estimate the policy ideal points of district-party publics.

This chapter proceeds in three major steps. First, I review the theoretical basis for ideal point models, rooted in spatial models of policy choice. I connect these formal models to statistical models of policy ideal points (in a style that follows Clinton, Jackman, and Rivers 2004) as well as their connection to Item Response Theory (IRT) models from psychometrics and education testing (e.g. Fox 2010).

Second, I specify and test the group-level model that I build and employ in my analysis of district-party publics. This discussion includes details that are relevant to Bayesian estimation, including identification restrictions on the latent policy space, specification and simulation of prior distributions, and model parameterizations that expedite estimation with Markov chain Monte Carlo (MCMC).

Lastly, I describe how I fit the model to real data. This section describes data collection, data processing, and model performance, and a descriptive analysis of the estimates.

2.1 Spatial Models and Ideal Points

Ideal points are constructs from spatial models of political choice. These models exist under formal theory—they simplify scenarios in the political world into sets of actors whose behaviors obey utility functions that conform to mathematical assumptions. Spatial models invoke a concept of "policy space," where actor preferences and potential policy outcomes are represented as locations along a number line. A canonical example is a left-right continuum, where progressive or "liberal" policies occupy locations on the left side of the continuum, while conservative policies are on the right side. Actors are at least partially motivated by their policy preferences, so they strive to achieve policy outcomes that are closest to their own locations in policy space. Figure 2.1 plots a simple example of an Actor's choice over two policies in one-dimensional policy space. The "Left" outcome is a more progressive policy outcome than the "Right" outcome. The Actor has a location, which corresponds to their most-desired policy outcome or "ideal point." Formal models like these encode assumptions that structure policy choice: whether the policy space is one-dimensional or multidimensional, the functional form of the Actor's policy utility (linear, quadratic, Gaussian…), and distributional assumptions about error terms in the utility function.

Figure 2.1: A left–right policy continuum featuring an Actor, a progressive policy outcome, and a conservative policy outcome.

Formal models of ideal points are distinct from statistical models of ideal points. Formal models are primarily theoretical exercises; they explore the incentives and likely actions of Actors in specific choice contexts, building theoretical intuitions that can be applied in the study of real-world politics with real data. Statistical models, on the other hand, explicitly or implicitly assume an underlying formal model of policy choice and estimate its parameters

using data. Data could come from legislators casting voting on bills, judges ruling on case outcomes, survey respondents stating their policy preferences (as in this project), and other situations. The statistical model in this chapter begins with a formal model where actors are partisan voters, choices are statements of policy preferences in surveys, and ideal points are summarized on a single dimension.[1] An Actor $i \in \{1, \ldots, n\}$ has an ideal point in policy space represented by θ_i. The Actor is confronted with survey item $j \in \{1, \ldots, J\}$ and chooses a Left alternative located at position L_j or a Right alternative located at R_j.

The Actor's choice is a function of the distance between their ideal point and the respective choice locations. Utility is maximized if the Actor can choose a policy located at exactly their ideal point, and utility is "lost" for choices farther from their ideal point. Like Clinton, Jackman, and Rivers (2004), I assume quadratic utility loss with increasing ideological distance, which implies a utility function over the *squared distance* between an Actor and a choice location. Let $U_i\left(L_j\right)$ and $U_i\left(R_j\right)$ be utility functions for i's choice of Left or Right on item j,

$$\begin{aligned} U_i\left(R_j\right) &= -\left(\theta_i - R_j\right)^2 + u_{ij}^{\mathsf{R}} \\ U_i\left(L_j\right) &= -\left(\theta_i - L_j\right)^2 + u_{ij}^{\mathsf{L}}, \end{aligned} \tag{2.1}$$

where u_{ij}^{R} and u_{ij}^{L} are the idiosyncratic error terms for the Right and Left alternatives, respectively. I sometimes refer to the quadratic utility loss as the "deterministic" component of the Actor's utility function, while the idiosyncratic error terms are "stochastic" components.

Let $y_{ij} = 1$ indicate the Actor chooses Right and $y_{ij} = 0$ indicate that they choose Left. The Actor chooses Left or Right based on which choice gives greater utility:

$$y_{ij} = 1 \iff U_i\left(R_j\right) > U_i\left(L_j\right) \tag{2.2}$$

[1]Researchers disagree about the appropriateness of a one-dimensional model for the mass public (Ansolabehere, Rodden, and Snyder 2008; Treier and Hillygus 2009). Past work on the strategic positioning dilemma theory guides our choice of a single-dimensional model, since the theory operates under a model of policy positioning where candidates take more progressive or conservative positions to target the policy preferences of their voters.

To visualize this choice, I plot the deterministic components of Equation (2.2) in Figure 2.2, omitting the stochastic utility terms. The parabola represents *i*'s utility loss for any choice along the ideological continuum, owed to their distance from that policy choice. The vertex of the parabola is at the Actor's ideal point, where policy utility is maximized. Dashed lines below the Left and Right alternatives represent the utility loss owed to the Actor's distance from those specific choices. In the current example, the Actor is closer to Right than to Left, so they receives greater utility (or lose *less* utility) by choosing Right instead of Left.

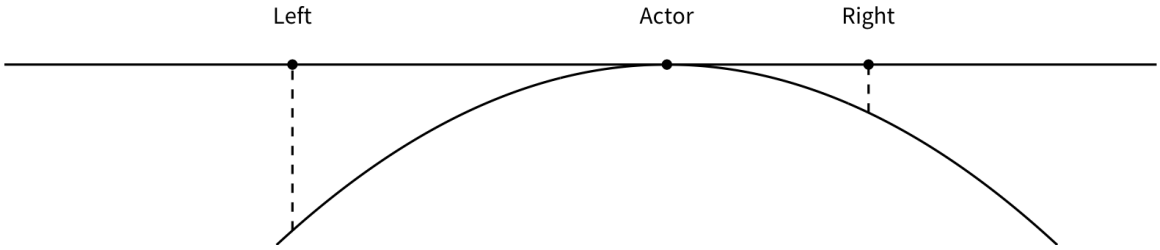


Figure 2.2: A representation of quadratic utility loss over policy choices

The Actor's choice is affected by stochastic utility components u_{ij}^{R} and u_{ij}^{L} in addition to the deterministic utility loss. This means that even if the Actor's distance to Right is smaller than their distance to Left, there remains a nonzero probability that *i* chooses Left. This probability depends on the values of the stochastic error terms for each choice. These error terms represent the accumulation of several possible, non-ideological shocks to utility: systematic decision factors that are not summarized by ideology, issue-specific considerations that do not apply broadly across all issues, random misperceptions about the policy locations, and so on. Supposing that these idiosyncratic terms follow some probability distribution, Equation (2.2) can be represented probabilistically:

$$\begin{aligned}
\Pr\left(y_{ij}=1\right) &= \Pr\left(U_i\left(R_j\right) > U_i\left(L_j\right)\right) \\
&= \Pr\left(-\left(\theta_i - R_j\right)^2 + u_{ij}^{\mathsf{R}} > -\left(\theta_i - L_j\right)^2 + u_{ij}^{\mathsf{L}}\right) \\
&= \Pr\left(\left(\theta_i - L_j\right)^2 - \left(\theta_i - R_j\right)^2 > u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right)
\end{aligned} \tag{2.3}$$

which states that the Actor will choose the closest policy alternative *unless* idiosyncratic factors overcome ideological considerations.

Equation (2.3) can be rearranged to reveal an appealing functional form for i's choice probability. By expanding the polynomial term and then factoring…

$$\begin{aligned}
\Pr\left(y_{ij}=1\right) &= \Pr\left(\left(\theta_i - L_j\right)^2 - \left(\theta_i - R_j\right)^2 > u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right) \\
&= \Pr\left(2\theta_i R_j - 2\theta_i L_j + L_j^2 - R_j^2 > u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right) \\
&= \Pr\left(2\theta_i R_j - 2\theta_i L_j + \left(R_j - L_j\right)\left(R_j + L_j\right) > u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right) \\
&= \Pr\left(2\left(R_j - L_j\right)\left(\theta_i - \frac{R_j + L_j}{2}\right) > u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right)
\end{aligned} \tag{2.4}$$

we reveal two interpretable expressions for important constructs in the model. First, note that $\frac{R_j+L_j}{2}$ is a formula for the midpoint between the Left and Right locations. Midpoints are significant for spatial models because they represent the threshold where Actor prefers one alternative over the other on average.[2] The term $\theta_i - \frac{R_j+L_j}{2}$ conveys which policy alternative is closer to the Actor; the expression is positive if the Actor is closer to Right and negative if closer to Left. Second, $2\left(R_j - L_j\right)$ captures how far apart the policy alternatives are from one another. Together, these capture whether the Actor is closer to the Left or Right policy and by how much.

The final manipulation is to simplify the terms above, resulting in a convenient parameterization for statistical estimation.

$$\Pr\left(y_{ij}=1\right) = \Pr\left(\iota_j\left(\theta_i - \kappa_j\right) > \varepsilon_{ij}\right) \tag{2.5}$$

This setup contains a "discrimination parameter" $\iota_j = 2\left(R_j - L_j\right)$, the "midpoint" or "cutpoint" parameter $\kappa_j = \frac{R_j + L_j}{2}$, and a joint error term $\varepsilon_{ij} = u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}$.[3] Similar to Equation (2.4),

[2] Ansolabehere, Snyder, and Stewart (2001) and Burden (2004) use candidate midpoints as predictors in regression analyses to estimate the impact of candidate ideology in House elections.

[3] The names for these parameters are adapted from item-response theory (IRT), an area of psychometrics that is similarly interested in inferring latent traits from observed response data.

$\theta_i - \kappa_j$ shows how far the Actor is from the midpoint between Left and Right and in which direction, and ι_j behaves as a "slope" on this distance: the distance from the midpoint has a *stronger influence* when the policy alternatives are farther from one another, since more utility is lost over larger spatial distances. I explore the intuitions of this functional form more thoroughly in Section 2.2.

A complete statistical model is obtained by a distributional assumption for ε_{ij}. A common assumption in ideal point modeling is to assume that ε_{ij} is Normal, which assumes that utility error represents a *sum* of unrelated utility shocks.[4] This assumption results in a probit regression model for the probability that Actor i chooses Right on choice j:

$$\begin{aligned} \Pr\left(y_{ij} = 1\right) &= \Pr\left(\iota_j\left(\theta_i - \kappa_j\right) > \varepsilon_{ij}\right) \\ &= \Pr\left(\iota_j\left(\theta_i - \kappa_j\right) - \varepsilon_{ij} > 0\right) \\ &= \Phi\left(\iota_j\left(\theta_i - \kappa_j\right)\right), \end{aligned} \tag{2.6}$$

where $\Phi(\cdot)$ is the cumulative Normal distribution function.[5] The probit model simply states that i is closer to Right than Left is a function of ideal point distances plus Normal error.

2.2 Item Response Theory

Ideal point models have a similar construction to models developed under item response theory (IRT) in psychometrics. IRT models have a similar aim as ideal point models: measuring latent features in the data given individuals' observable responses to stimuli. The canonical psychometric example is in education testing, where several test questions are

[4] Many IRT models assume that ε_{ij} follows a Logistic distribution, resulting in a logistic regression model (for example Londregan 1999). Logistic models are computationally convenient but harder to justify in the ideal point context, especially as we add sources of variance to the choice problem in the group-level model.

[5] The probit model implies a scale restriction that $\text{Var}\left(u^{\text{L}}_{ij} - u^{\text{R}}_{ij}\right) = 1$. A scale restriction for a single choice is not problematic because the ideological scale is latent and can be arbitrarily stretched with the discrimination parameter ι_j. The important implication for estimation is that the error variance is equal *across individuals*.

used to measure a student's latent academic "ability" level. This section connects ideal point models to IRT to help explain the theoretical and mathematical intuitions at work.

2.2.1 Latent traits

IRT models are *measurement models*. The goal of a measurement model is to use observed data $\mathbf{y}$ to estimate some latent construct of theoretical interest $\boldsymbol{\theta}$. The observed data $\mathbf{y}$ are affected by $\boldsymbol{\theta}$, but there is no guarantee of a one-to-one correspondence between the two because $\boldsymbol{\theta}$ is not observable. A measurement model provides assumptions that structure a data-generating process, allowing the researcher to infer latent traits from observed data. We can represent a measurement model with general notation $\mathbf{y} = f(\boldsymbol{\theta}, \boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ represents auxiliary model parameters that are estimated in addition to $\boldsymbol{\theta}$. In the education context, a student's observable pattern of answers to test questions reveals their underlying ability level. In a context of policy choice, it is impossible to observe an individual's political ideology directly, but the model provides a structure to infer ideology from a pattern of policy choice questions on a survey.

An important point about measurement models is that they require assumptions to infer reconstruct a signal about latent traits from observed data, otherwise no inferences can be draws. In this sense, the estimates are always sensitive to the model's assumptions. While this is always important to acknowledge, it is an ever-present consideration even for simpler measurement strategies such as additive indices that estimate ideology by summing or averaging responses to a battery of policy questions. In fact, additive indices are special cases of measurement models where key parameters are assumed to be known with certainty and fixed across all policy items, which is problematic if there is any reason to suspect that item responses are correlated across individuals.[6] In this way, measurement models actually

[6] Additive indices are represent a model where all policy choices have the same Left and Right locations, which is a highly restrictive assumption that IRT models do not make.

relax the assumptions of simpler measurement strategies, even if the underlying mathematics are more intensive.

IRT models have been used in political science to measure latent constructs in many different political contexts. The predominant context is ideological scaling, which has been performed for members of Congress (Clinton, Jackman, and Rivers 2004; Poole and Rosenthal 1997), Supreme Court justices (Martin and Quinn 2002), contributors and recipients of campaign funds (Bonica 2013), Twitter users (Barberá 2015), individuals in the mass public (Treier and Hillygus 2009) and groups of individuals (Tausanovitch and Warshaw 2013). These models have been used in non-ideological contexts as well to model regime type (Treier and Jackman 2008), UN voting (Voeten 2000), comparative judicial independence (Linzer and Staton 2015), and more.

2.2.2 Item characteristics and item parameters

Measurement models relax assumptions about the data's functional dependence on the latent trait by modeling features of the items to which subjects respond. Different items reveal different information about the latent construct (Fox 2010). Consider a simple model where an student i is more likely to answer test questions j correctly if she has greater academic ability θ_i. Analogously, a citizen who is more conservative is more likely to express conservative preferences for policy question j. Keeping the probit functional form from above, we can represent this simple model with the equation:

$$\Pr\left(y_{ij} = 1\right) = \Phi\left(\theta_i\right), \tag{2.7}$$

where the probability of a correct/conservative response is 0.5 at $\theta_i = 0$. This model asserts that knowing θ_i is sufficient to produce exchangeable response data; there are no remaining systematic differences across questions that lead to systematically different answers. This implicit assumption is often unrealistic: some test questions are more difficult than others,

and some policy questions present more extreme or lopsided choices than others. Even so, political science is replete with measurement approaches that omit all item-level variation, such as the racial resentment scale, the additive index of of political participation, and more (Henry and Sears 2002; Verba and Nie 1972).

IRT models assume that items reveal systematically different information from one another and model these item characteristics using *item parameters*. IRT models have different behaviors based on the parameterization of the item effects in the model. The simplest IRT model is a "one-parameter" model, which includes an item-specific intercept κ_j.

$$\Pr\left(y_{ij} = 1\right) = \Phi\left(\theta_i - \kappa_j\right) \tag{2.8}$$

IRT parlance refers to κ_j as the idem "difficulty" parameter. In the testing context, if a student has greater ability than the difficulty level of the question, they will most likely answer the item correctly. This probability goes up for easier questions and down for more difficult questions. In policy choice, the difficulty parameter represents the "midpoint" parameter between two policy alternatives. The chooser prefers Left or Right depending on their ideal point location relative to the midpoint. The "two-parameter" IRT model is more common, especially in the ideal point context. The two-parameter model introduces the "discrimination" parameter ι_j, which behaves as a slope on the difference between θ_i and κ_j.

$$\Pr\left(y_{ij} = 1\right) = \Phi\left(\iota_j\left(\theta_i - \kappa_j\right)\right), \tag{2.9}$$

Intuitively, the discrimination parameter captures how well a test item differentiates between the responses of high- and low-ability students, with greater values meaning more divergence in responses. In the ideal point context, it captures how strongly a policy question divides liberal and conservative respondents.

Figure 2.3 shows how response probabilities are affected by the parameterization of item effects. Each panel plots how increases in subject ability or conservatism (the horizontal

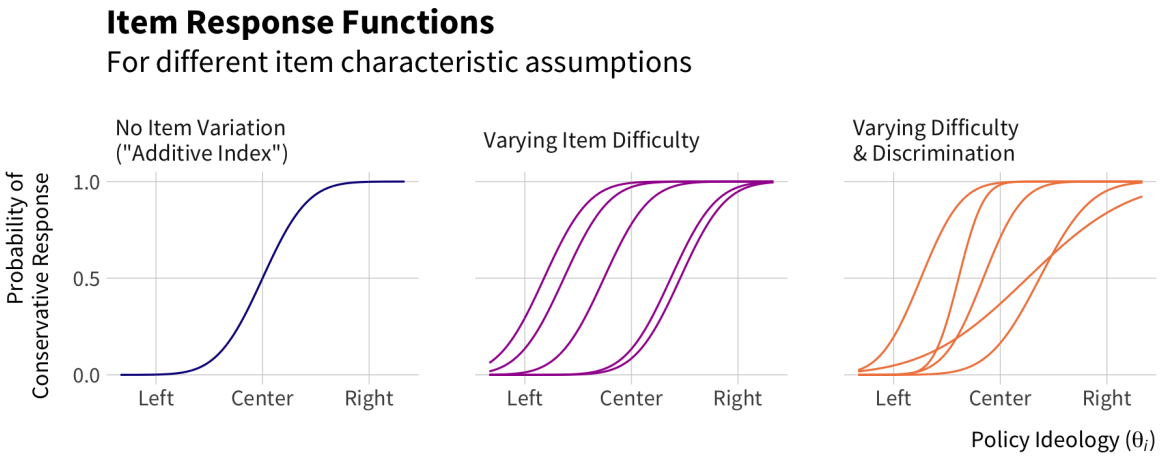


Figure 2.3: Examples of item characteristic curves under different item parameter assumptions

axis) result in increased response probability (the vertical axis), where the shape of the curve is set by values of the item parameters. These curves are commonly referred to as *item characteristic curves* (ICCs) or *item response functions* (IRFs). The leftmost panel shows a model with no item effects whatsoever; any item is theorized to behave identically to any other item, and response probabilities are affected only by the subject's ability (ideology). The middle panel shows a one-parameter model where item difficulties (cutpoints) are allowed to vary systematically at the item level. Difficulty parameters behave as intercept shifts, so they change the value of θ predicts a correct response with probability 0.5, but they do not affect the slope of the item response function. The final panel shows item response functions from the two-parameter IRT model, where item difficulties (intercepts) and discriminations (slopes) are allowed to vary across items.

How do we interpret our statistical model of ideal points (Equation (2.6)) in light of item response theory? The cutpoint parameter $\kappa_j = \frac{R_j + L_j}{2}$ behaves like a difficulty parameter, capturing intercept shifts in the item response function. If $\theta_i - \kappa_j = 0$, the item cutpoint falls directly on an Actor's ideal point, this the Actor is indifferent (in expectation) between Left and Right and thus chooses Right with probability 0.5. The value of κ_j increases by moving

either the Right or Left alternatives to the right (increasing R_j or L_j), subject to the constraint that $R_j \geq L_j$. Larger values of the item cutpoint imply a lower probability that the Actor chooses Right, since κ_j has a non-positive effect on the conservative response probability.[7] The opposite intuition holds as the Left position becomes increasingly progressive, resulting in larger values of κ_j that imply a higher probability of choosing Right, all else equal.

The discrimination parameter $\iota_j = 2(R_j - L_j)$ grows when the distance between the Right and Left alternatives grows larger, either because R_j increases or L_j decreases. This parameter captures the "slope" on the distance between the Actor ideal point and the cutpoint, meaning that the Actor's choice is more elastic to their policy preferences as ι_j increases.[8] Substantively, this means that progressives and conservatives vote more differently from one another when the policy alternatives present a starker ideological contrast.[9]

2.3 Modeling Party-Group Ideology in Congressional Districts

Having laid out the basics of the individual ideal point model and item response theory, this section outlines the group-level ideal point model for district-party publics. It connects the individual- and group-level response models and lays out a hierarchical model for smoothing district-party ideal points. I discuss several technical features of model implementation in the Stan software for Bayesian estimation (Carpenter et al. 2016), including the structure

[7] Formally we can show this by taking the derivative of the link function with respect to the cutpoint: $\frac{\partial \iota_j \left(\theta_i - \kappa_j\right)}{\partial \kappa_j} = -\iota_j$, where the polarity of the ideal point space (conservative means "right") constrains ι_j to be positive.

[8] Again we can demonstrate this by noticing that the derivative of the link function with respect to the discrimination parameter is $\frac{\partial \iota_j \left(\theta_i - \kappa_j\right)}{\partial \iota_j} = \left(\theta_i - \kappa_j\right)$. The derivative's magnitude depends on the absolute value of this distance, and its sign depends on the sign of the difference.

[9] In a special case that Right and Left alternatives are located in exactly the same location, the result is $\kappa_j = \iota_j = 0$, leading all Actors to choose Right with probability 0.5. This result represents a situation where policy preferences are not systematically related to the choice whatsoever, and only idiosyncratic error affects the choice of Right or Left. Although the model implies that this result is *mathematically* possible, it is not realistic to expect any of the policy choices in this project to induce this behavior.

of the data, model parameterization, identifiability restrictions, and prior distributions for Bayesian estimation.

Unlike the ideal point model discussed so far, this project is concerned with the average ideal point for a *group* of individuals. These groups are district-party publics: groups of major-party identifiers nested within congressional districts (435 districts × 2 parties per district = 870 district-party groups). The group model assumes that individual ideal points are distributed around some group average. Let g index groups, which are intersections of congressional districts d and major political party affiliations p. As before, a probit model describes the probability that an individual i gives a conservative response to survey item j.

$$y_{ij} \sim \text{Bernoulli}(\pi_{ij}) \tag{2.10}$$

$$\pi_{ij} = \Phi\left(\iota_j\left(\theta_i - \kappa_j\right)\right) \tag{2.11}$$

Following Caughey and Warshaw (2015), it is helpful to reparameterize the IRT model to accommodate a group-level extension. This parameterization replaces item "discrimination" with item "dispersion" using the parameter $\sigma_j = \iota_j^{-1}$,

$$\pi_{ij} = \Phi\left(\frac{\theta_i - \kappa_j}{\sigma_j}\right). \tag{2.12}$$

which controls the additional "measurement error" introduced by a survey question beyond the standard Normal utility error in the probit model (Caughey and Warshaw 2015).

The group model assumes that individual ideal points are Normally distributed within a group. In other words, individual ideal point deviations from the mean ideal point in the group are the result of a sum of random forces.

$$\theta_i \sim \text{Normal}\left(\bar{\theta}_{g[i]}, \sigma_{g[i]}\right) \tag{2.13}$$

where $\bar{\theta}_{g[i]}$ and $\sigma_{g[i]}$ are the mean and standard deviation of ideal points within i's group g.[10]

[10]Notation for Normal distributions will always describe the scale parameter in terms of standard deviation σ instead of variance σ^2.

While it is possible to continue building the model hierarchically from (2.13), it would be far too computationally expensive to estimate every individual's ideal point in additional to the group-level parameters—every individual ideal point is essentially a nuisance parameter. Instead, we aggregate individual-level responses to the group level and marginalize over the distribution of individual ideal points. Let $s_{gj} = \sum_{i \in g}^{n_{gj}} y_{ij}$ be the number of conservative responses from group g to item j, where n_{gj} is the total number of responses (trials) to item j by members of group g. If trials are conducted independently across groups and items (an assumption that is relaxed later), we could model the grouped outcome as a binomial random variable,

$$\begin{aligned} s_{gj} &\sim \text{Binomial}\left(n_{gj}, \bar{\pi}_{gj}\right) \\ \bar{\pi}_{gj} &= \Phi\left(\frac{\bar{\theta}_g - \kappa_j}{\sqrt{\sigma_g^2 + \sigma_j^2}}\right), \end{aligned} \tag{2.14}$$

where $\bar{\pi}_{gj}$ is the average conservative response probability for item j in group g, or the probability that a randomly selected individual from group g gives a conservative response to item j. Our uncertainty about the item response now contains two sources of variance: the uncertainty introduced by the item itself and the variance of the individual ideal points around the group mean. Because individual ideal points are assumed to be Normal within their group, the within-group variance can simply be added to the probit model as another source of measurement error. Larger within-group variances attenuate $\bar{\pi}_{ij}$ toward 0.5.[11]

The current setup assumes that every item response is independent, conditional on the group and the item. This assumption is violated if the same individuals in a group answer multiple items—one individual who answers 20 items is less informative about the group average than 20 individuals who answer one item apiece. While this too could be addressed by explicitly modeling each individual's ideal point (extending the model directly from Equation

[11] Caughey and Warshaw (2015) derive this result in the appendix to their article.

(2.13)), I implement a weighting routine that downweights responses from subjects who answer multiple items, described in Section 2.3.3.

2.3.1 Hierarchical model for group parameters

The group model described so far can be estimated straightforwardly if there are enough responses from enough individuals in enough district-party groups. In practice, however, it would take dozens more surveys than are practically available to achieve this level of precision with survey data alone. Instead, I build a hierarchical model for the group parameters that regularizes group ideal point estimates in a principled way. The hierarchical model estimates how group ideal points are related to features of the districts and states in which they are located, borrowing strength from data-rich groups to stabilize the ideal point estimates for data-sparse groups. This section describes the multilevel structure using traditional notation for hierarchical models; later in Section 2.3.4 I describe how I reparameterize the model for computational implementation.

I posit a hierarchical structure where groups g are "cross-classified" within districts d and parties p. This means that groups are nested within districts and within parties, but districts and parties have no nesting relationship to one another. Districts are further nested within states s. I represent this notationally by referring to group g's district as $d[g]$. Similarly, g's party is $p[g]$. For higher levels such as g's state, $s[g]$ is shorthand for the more-specific but more-tedious $s[d[g]]$. In some settings, I drop the g from the subscript altogether to achieve simpler notation.

I use this hierarchical structure to model the probability distribution of group ideal points $\bar{\theta}_g$. I consider the group ideal point as a Normal draw from a regression on geographic-level data with parameters that vary across political party. This regression takes the form

$$\bar{\theta}_g \sim \text{Normal}\left(\mu_{p[g]} + \mathbf{x}_{d[g]}^{\top}\beta_{p[g]} + \alpha_{s[g]p[g]}^{\texttt{state}}, \sigma_p^{\texttt{group}}\right) \tag{2.15}$$

where μ_p is a constant specific to party p,[12] and $\mathbf{x}_d$ is a vector of congressional district-level covariates with party-specific coefficients β_p. State effects $\alpha_{sp}^{\texttt{state}}$ are also specific to each party.

The benefit of specifying separate parameters for each party is that geographic features may be related to ideology in ways that are not identical across all parties. This is an important departure from the structure laid out by Caughey and Warshaw (2015), who estimate the same set of geographic effects for all groups in the data. This flexibility is especially important if certain contextual factors influence ideology in opposite ways: for instance, racial heterogeneity within a congressional district may make Democrats more progressive and Republicans more conservative.

The state effects are regressions on state features as well,

$$\alpha_{sp}^{\texttt{state}} \sim \text{Normal}\left(\mathbf{z}_s^\top \gamma_p + \alpha_{r[s]p}^{\texttt{region}}, \sigma_p^{\texttt{state}}\right), \tag{2.16}$$

where state-level covariates $\mathbf{z}_s$ have party-specific coefficients γ_p. Each state effect contains a party-specific region effect $\alpha_{r[s]p}^{\texttt{region}}$ for Census regions indexed r, which is a modeled mean-zero effect to capture region-level correlations across the state effects.

$$\alpha_{rp}^{\texttt{region}} \sim \text{Normal}\left(0, \sigma_p^{\texttt{region}}\right) \tag{2.17}$$

This model is estimated in a Bayesian framework using Markov chain Monte Carlo (MCMC) to sample the posterior distribution. This means that all above parameters have prior distributions or are functions of parameters with prior distributions. I discuss the parameterization of the model and priors below.

I use this model to scale the ideal points for groups of major party identifiers in all congressional districts for the 2010s districting cycle. Researchers wishing to extend the district-party ideal point model could construct a dynamic model that scales congressional districts across election cycles or across 10-year districting cycles. Because congressional

[12] Or "grand mean," since all covariates are eventually be centered at their means.

districts change boundaries across districting cycles—sometimes falling out of or coming into existence, depending on decennial apportionment of districts to states—the district-party groups may not be considered as the "same group over time." The ideal point space, however, could be identified over time either by specifying dynamic priors on hierarchical regression parameters or by fixing item parameters over time (Caughey and Warshaw 2015).

2.3.2 Identifying the latent policy space

Ideal point models, as with all latent space models, are unidentified without restrictions on the latent space. The likelihood model as written can rationalize many possible estimates for the unknown parameters, with no prior basis for deciding which estimates are best. These identifiability problems affect the location, scale, and polarity of the latent space.

- Location: the latent scale can be arbitrarily shifted right or left. We could add some constant to every ideal point, and the response probability would be unaffected if we also add the same constant to every item cutpoint.
- Scale: the latent scale can be arbitrarily stretched or compressed. We could multiply the latent space by some scale factor, and the response probability would be unaffected if we also multiply the discrimination parameter by the inverse scale factor.
- Polarity: the progressive and conservative poles of the scale could be reversed by flipping the sign of every ideal point and item parameter.

All statistical models have the potential to be non-identified in this way, but fixed covariate data typically provide the restrictions necessary to identify a model.[13] Because the response probability is a function of latent-space parameters, however, data alone do not identify a unique solution. As a result, I must provide my own identifiability restrictions to the ideal point space.

[13] We could imagine shifting, stretching, or reversing the sign of a model covariate to reveal the same identifiability issues.

The polarity of the space is fixed by coding all survey items such that conservative responses are 1 and liberal responses are 0. This implies a restriction that all discrimination parameters are positive, which ensures that moving to the *right* in the latent ideological space always leads to an increased probability of a *conservative* item response, all other parameters held equal.

The location of the space is set by restricting the sum of the J item cutpoints to be 0. If $\tilde{\kappa}_j$ is a latent cutpoint in an unrestricted space, the cutpoint value in the restricted space κ_j would be defined as

$$\kappa_j = \tilde{\kappa}_j - \frac{\sum_{j=1}^{J} \tilde{\kappa}_j}{J}, \tag{2.18}$$

which is performed in every iteration of the estimation routine. This restriction on the sum of the cutpoint parameters implies that the mean of the cutpoints is zero as well.

Lastly, I set the scale of the latent space by restricting the product of the J discrimination parameters to be equal to 1. In practice, I implement this restraint on the log scale by restricting the log discrimination parameters to sum to 0, which achieves an equivalent transformation. If $\tilde{\iota}_j$ is a discrimination parameter in an unrestricted space, the restricted ι_j value is defined as follows.

$$\log\left(\iota_j\right) = \log\left(\tilde{\iota}_j\right) - \frac{1}{J}\sum_{j=1}^{J} \log\left(\tilde{\iota}_j\right) \tag{2.19}$$

$$\iota_j = \exp\left(\log\left(\iota_j\right)\right) \tag{2.20}$$

Item discrimination is then reparameterized as dispersion, $\sigma_j = \iota_j^{-1}$. These restrictions on the item parameters are sufficient to identify $\bar{\theta}_g$.

2.3.3 Weighted outcome data

The group-level model learns about group ideal points by surveying individuals within groups, but the model currently assumes that all y_{gj} are independent conditional on the item. If the

same individuals answer multiple items, this assumption is violated. Additionally, we cannot assume that responses are independent in the presence of nonrepresentative survey designs that oversample and undersample certain populations. This section describes an approach for weighting group-level data that adjusts for both issues, based on Ghitza and Gelman (2013) and Caughey and Warshaw (2015). The recipe contains three adjustments: a "design effect" to account for nonequal response probabilities within a given group g, an adjusted sample size for every group-item combination, and an adjusted number of successes for each group-item.

First, the sample size in each group-item "cell" $g\,j$ must is adjusted for survey design and multiple responses per individual. Let $n^*_{g[i]j}$ be the adjusted sample size for i's group-item cell, defined as

$$n^*_{g[i]j} = \sum_{i=1}^{n_{g[i]j}} \frac{1}{r_i d_{g[i]}}, \tag{2.21}$$

where r_i is the number of responses given by individual i, and $d_{g[i]}$ is a survey design correction for i's group. The effective sample size decreases when respondents answer multiple questions ($r_i > 1$) or in the presence of a sample design correction ($d_g > 1$). The design correction, originally specified by Ghitza and Gelman (2013), penalizes information collected from groups that contain greater variation in their survey design weights. It is defined as

$$d_{g[i]} = 1 + \left(\frac{\text{sd}_{g[i]}\,(w_i)}{\text{mean}_{g[i]}\,(w_i)} \right)^2, \tag{2.22}$$

where $\text{sd}(\cdot)$ and $\text{mean}(\cdot)$ are the within-group standard deviation and mean of respondent weights w_i. If all weights within a cell are identical, their standard deviation will be 0, resulting in a design correction equal to 1 (meaning, no correction). Larger within-cell variation in weights increases the value of d_g, thus decreasing the effective sample size within a cell. The intuition of this correction is to account for increased variance of weighted statistics compared to unweighted statistics, given a fixed number of observations (Ghitza and Gelman 2013, 765).

To obtain the weighted number of successes in cell gj, I multiply the cell's weighted sample size by its weighted mean. The weighted mean $\bar{y}^*_{gj}$ adjusts for the respondent's survey weight w_i and number of responses r_i and is defined as

$$\bar{y}^*_{g[i]j} = \frac{\sum_{i=1}^{n_{g[i]j}} \frac{w_i y_{ij}}{r_i}}{\sum_{i=1}^{n_{g[i]j}} \frac{w_i}{r_i}}. \tag{2.23}$$

The weighted number of successes in each cell, in turn, is

$$s^*_{gj} = \min\left(n^*_{gj}\bar{y}^*_{gj}, n^*_{gj}\right). \tag{2.24}$$

where I take the minimum to ensure that the number of successes does not exceed the adjusted sample size.

It is likely that many values of n^*_{gj} and s^*_{gj} will be non-integers. This is a problem for modeling a Binomial random variable, which is an interger-valued count of successes out of an interger-valued number of trials. To rectify this, I follow Caughey and Warshaw (2015) and round the weighted number of trials up (to ensure no trail counts of zero) and round the weighted number of successes to the nearest integer. These integer results are then used in the IRT likelihood as follows.[14]

$$\lfloor s^*_{gj} \rceil \sim \text{Binomial}\left(\lceil n^*_{gj} \rceil, \bar{\pi}_{gj}\right) \tag{2.25}$$

2.3.4 Hierarchical model parameterization

I estimate the model using Stan, a probabilistic programming language for constructing Bayesian analysis implements estimation using compiled `C++` programs (Carpenter et al.

[14] Another possible tactic would be to use a quasi-likelihood approach (Ghitza and Gelman 2013), supplying the weighted number of successes and trials to the Binomial probability mass function directly instead of using a built-in computer function that will fail for non-integers. This is more complex to implement in Bayesian estimation software because it requires the researcher to add log probability expressions directly to the *log density accumulator* construct in a Stan program (Carpenter et al. 2016), though other research has demonstrated the routine for a Binomial quasi-likelihood problem (DeCrescenzo and Mayer 2019, 350).

2016). Stan implements an adaptive variant of Hamiltonian Monte Carlo (HMC) sampling, an algorithm that efficiently proposes Markov transitions by "surfing" a proposal trajectory along the gradient of the posterior distribution instead of walking randomly in parameter space (Betancourt 2017; Neal 2012). Although these Hamiltonian mechanics are highly effective for fast and reliable Bayesian estimation, the model must be parameterized to ensure that the geometry of the posterior distribution is favorable to the HMC proposal trajectories. This section describes how the parameterization of the model *as programmed in Stan* departs from the model *as written* above.

It is important to discuss these implementation details because they are essential to valid MCMC estimation. Models that are not parameterized for stable computation are likely to return severely biased estimates, which is especially worrisome for Bayesian models with complex, high-dimensional posterior distributions. Furthermore, recent work in political science that employs similar models (Caughey and Warshaw 2015) does not utilize best-practice parameterizations, so this discussion provides an important corrective for other researchers estimating similar models. Broadly speaking, there are no reference texts for doing contemporary Bayesian analysis in political science that reflect recent innovations in Hamiltonian Monte Carlo methods. This is despite the fact that HMC has been the state of the art MCMC method for most applied Bayesian problems for several years (Betancourt 2017; Bürkner and others 2017; Carpenter et al. 2016; Patil, Huard, and Fonnesbeck 2010).[15]

The group-level IRT model regularizes parameters estimates using several hierarchical regression models: group ideal points are regularized using a district-level regression, state effects are regularized using a state-level regression, and region effects are pooled toward a hierarchical prior with a mean of zero. Hierarchical models like these have posterior distributions whose curvatures are known present difficulties for sampling algorithms (Betancourt

[15] Gibbs sampling methods are still common for problems with categorical parameters and other models with discontinuous posterior distributions.

and Girolami 2015; Papaspiliopoulos, Roberts, and Sköld 2007). To improve the estimation in Stan, the best practice for most situations is to parameterize hierarchical models using a "non-centered" model parameterization rather than a "centered" parameterization. Whereas the centered parameterization considers $\bar{\theta}_g$ as a random draw from a hierarchical distribution (Equation (2.15) above), the non-centered parameterization defines $\bar{\theta}_g$ as a deterministic function of its conditional hypermean and a random variable.

$$\bar{\theta}_g = \mu_p + \mathbf{x}_d^\top \beta_p + \alpha_{sp}^{\texttt{state}} + u_{dp[d]}\tau_p, \tag{2.26}$$

$$u_{dp} \sim \text{Normal}\,(0, 1) \tag{2.27}$$

$$\tau_p \sim \ldots \tag{2.28}$$

where $u_{dp}\tau_p$ behaves as a group-level error term. The group-level error term is decomposed into a standard Normal variate u_{dp} and a scalar parameter τ_p that controls the scale of the error term. The scale parameter τ_p then is given a separate prior.

The non-centered model is algebraically equivalent to centered model as it pertains to the likelihood of the data, but the non-centered parameterization "unnests" the mean and the scale from the hierarchical distribution. This parameterization improves MCMC sampling by de-correlating random variables in the posterior distribution, leading to more efficient MCMC exploration. The centered parameterization, meanwhile, contains regions of highly correlated parameter space that trap Markov chains in "funnels" that are difficult to escape (Betancourt and Girolami 2015; Papaspiliopoulos, Roberts, and Sköld 2007). These problematic regions in the posterior lead to "divergent transitions" in the Markov chain: transitions where the posterior curvature is so high that the Hamiltonian algorithm cannot accurately estimate the trajectory of its own proposals. Markov chains with many divergent transitions have a high risk of being severely biased because they indicate that the chain is failing to navigate the parameter space effectively. The non-centered parameterization

smooths out these problematic regions of posterior density, safeguarding against biased MCMC estimates.

Equation (2.28) is an incomplete implementation of the non-centered form; to complete the parameterization, I apply it to all hierarchical components in the regression, including the state and region effects,

$$\begin{aligned} \bar{\theta}_g &= \mu_p + \mathbf{x}_d^\top \beta_p + \mathbf{z}_s^\top \gamma_p \\ &\quad + u_{dp}^{\texttt{district}} \tau_p^{\texttt{district}} + u_{sp}^{\texttt{state}} \tau_p^{\texttt{state}} + + u_{rp}^{\texttt{region}} \tau_p^{\texttt{region}} \\ u^{\texttt{district}}, u^{\texttt{state}}, u^{\texttt{region}} &\sim \text{Normal}(0, 1) \\ \tau^{\texttt{district}}, \tau^{\texttt{state}}, \tau^{\texttt{region}} &\sim \dots \end{aligned} \tag{2.29}$$

where $u^{\texttt{district}}$, $u^{\texttt{state}}$, and $u^{\texttt{region}}$ are all standard Normal variables, and the scale components τ for each level have their own prior distributions that I specify below.

2.3.5 Prior Distribution

Bayesian models require a prior probability distribution over the model parameters, which can be a benefit and a drawback of the approach. The primary benefit is the ability to encode external information into a model, allowing the researcher to downweight unreasonable estimates, regularize against overfitting, stabilize weakly identified quantities, and pool estimates across many groups of data (Gelman and Hill 2006). Bayesian estimation is especially valuable for ideal point models because MCMC generates posterior samples of latent variables as if they were any other model parameter. The drawback of Bayesian modeling is that prior specification is additional work for the researcher, which can be complicated. The following discussion describes and justifies priors used in the group ideal point model. The discussion here is more detailed than a typical paper describing a Bayesian ideal point model for a few reasons. First, the norms of the typical Bayesian workflow are evolving toward more rigorous checking of prior distributions and their implications (Betancourt 2018; Gabry

et al. 2019), allowing researchers to explore and demonstrate the consequences of priors. Bayesian models in political science tend to lack explicit prior simulations, which can make prior specifications feel opaque or arbitrary to non-Bayesian readers. Additionally, and more specifically to this project, the nonlinearities in a probit model present particular challenges for specifying priors. In a few instances, I choose priors that depart from the priors used in previous Bayesian ideal point work for theoretical and practical reasons that I detail below.

As a general orientation, this model favors "weakly informative priors" (Gelman, Simpson, and Betancourt 2017): priors that use structural information about a model to rule out extreme model configurations. Weakly informative priors are stronger than flat priors, which can have problematic implications especially for nonlinear models like the probit model.[16] At the same time, weakly informative priors are weaker than fully informative priors that represent substantive beliefs about model parameters (for a political science example, see Western and Jackman 1994). Priors are pragmatic devices that scale parameters to match the order of the variation in the data. For instance, most predictors are scaled to have a standard deviation of 1, so effects on the order of one standard deviation would actually be quite large.

2.3.6 The model is prior information

One important source of pragmatic prior information is the model itself. Probit models translate latent scale quantities into probabilities using the standard Normal cumulative distribution function (CDF)—the latent scale represents z-scores (Normal quantile values) for predicted success probabilities. This relationship between the probability scale and the Normal quantile scale means that the researcher can reliably anticipate that their model will generate latent-scale estimates that map to reasonable probability values. For most social science problems, reasonable probability values map to a very narrow region of the Normal quantile values. This is a principle that holds for many nonlinear models with latent scales:

[16] For more on flat priors, see Chapter 3.

latent scales are often described as "uninterpretable," but the researcher usually only needs a few heuristic rules to derive sensible priors.

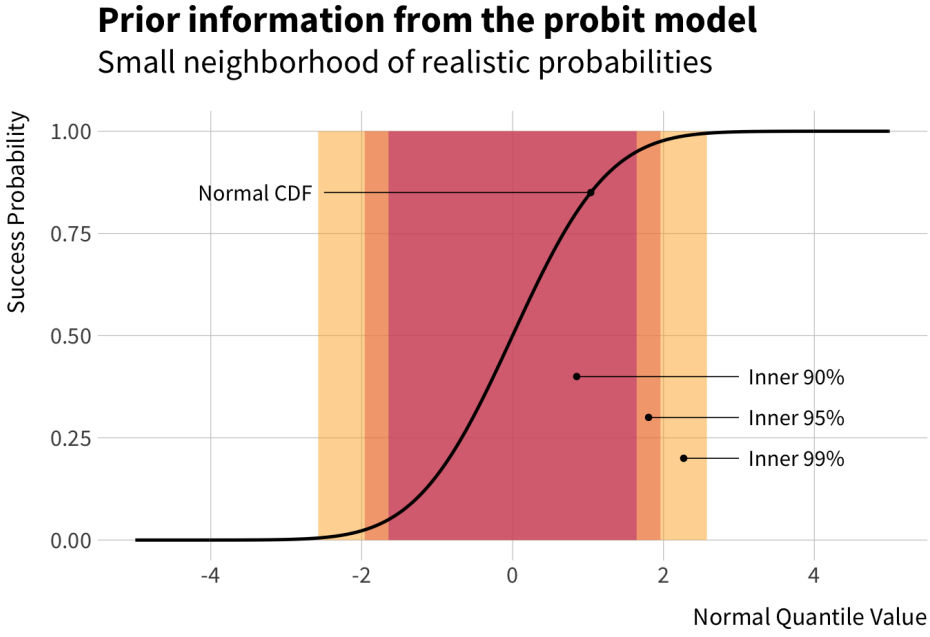


Figure 2.4: The region of the probit model's latent index that maps to response probabilities between 1 and 99 percent.

For a probit model, one helpful heuristic is knowing that 95% of the probability scale falls between approximately -2 and 2. This means that if as long as a model is not expected to give predicted probabilities that are extremely high or extremely low, priors that keep model predictions in the neighborhood of -2 and 2 provide weak information to constrain the space of possible model configurations. Figure 2.4 provides a graphical depiction of this point, with different highlighted regions to indicate how probabilities map to the latent quantile scale.

A more general takeaway from the relationship between the probit model and the standard Normal distribution is that prior ignorance on the probability scale is reasonably approximated by priors scales of order 1 on the quantile scale. For instance, if a model quantile value has a Normal $(0, 1)$ distribution, this implies a flat distribution over the probability scale. Priors of this sort are a simple example of weak information: enough regularization to

downweight extreme model configurations that upweight very high or very low probabilities, but weak enough that the data adjudicate between model configurations in relevant regions of the probability scale. For this reason, the priors for regression parameters are concentrated around these scale values. Constants are given Normal $(0, 1)$ priors. The hierarchical scale parameters in the group, state, and region random effects (τ) are given Half-Normal $(0, 1)$ priors, and regression coefficients are regularized slightly more with Normal $(0, 0.5)$ priors. Because the covariate data for the regressions are all scaled to be standard deviation 1, a prior scale of 0.5 means that approximately 98% of prior mass is concentrated on effect sizes of less than 1 on the quantile scale.[17]

Allowing the likelihood model to inform prior scales is common among Bayesian statisticians but rarer among political scientists. The widespread understanding that flatter priors always represent more principled prior ignorance is not generally true, but instead depends on the parameter space—an important point that I explore further in Chapter 3. This is especially relevant for modeling problems like this, where nonlinear parameter spaces that concentrate most plausible model configurations on certain orders of magnitude. Early ideal point models for the mass public (e.g. Treier and Hillygus 2009) tend to use highly diffuse priors, such as Normal $\left(0, \sqrt{1000}\right)$, which place the majority of prior probability on model configurations that are virtually impossible to be true. More recent Bayesian ideal point models (e.g. Tausanovitch and Warshaw 2013; Caughey and Warshaw 2015) use more reasonable prior scales such as 5 or 2.5, which are still broad but at least on an order of magnitude that fits the probit model.

[17] Normal priors for coefficients are similar to a ridge penalty (Bishop 2006), a commonplace regularization choice in predictive modeling.

2.3.7 Priors for item parameters

I specify priors on the unscaled cutpoint and discrimination parameters that are Normal and LogNormal, respectively. Whereas Caughey and Warshaw (2015) specify independent priors for all item cutpoint and discrimination parameters separately, my hierarchical model partially pools the item parameters toward a common multivariate distribution. This allows estimates to borrow precision from one another rather than "forgetting" the information learned from one item when updating the prior for the next item. This joint priors is a multivariate Normal distribution for the cutpoint and logged discrimination parameter,

$$\begin{bmatrix} \tilde{\kappa}_j \\ \log\left(\tilde{\iota}_j\right) \end{bmatrix} \sim \text{Normal}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{2.30}$$

where $\boldsymbol{\mu}$ is a 2-vector of means and $\boldsymbol{\Sigma}$ is a 2×2 variance–covariance matrix. The discrimination parameter, which has a product of 1 when scaled, is logged so that it has a mean of 0 on the log scale. This simplifies the prior specification of the mean vector $\boldsymbol{\mu}$, which is a standard multivariate Normal with no off-diagonal elements.[18]

$$\boldsymbol{\mu} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \tag{2.31}$$

I build a prior for the variance–covariance matrix $\boldsymbol{\Sigma}$ using a "separation strategy" for covariance matrix priors (Barnard, McCulloch, and Meng 2000). The separation strategy decomposes the covariance matrix into a diagonal matrix of scale terms and a unit-diagonal

[18] Although I use a joint prior, the assumptions about the parameters' marginal distributions are similar to Caughey and Warshaw (2015). Their choice to restrict discrimination parameter to have a product of 1 and a LogNormal distribution is identical to my choice to restrict log discrimination parameters to have a sum of 0 and a Normal prior. The benefit of my parameterization is that, by specifying the Normal family directly on the logged discrimination parameter, it is much simpler to build the joint hierarchical prior for all item parameters simultaneously.

correlation matrix,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\tilde{\kappa}}^2 & \rho\sigma_{\tilde{\kappa}}\sigma_{\tilde{\iota}} \\ \rho\sigma_{\tilde{\kappa}}\sigma_{\tilde{\iota}} & \sigma_{\tilde{\iota}}^2 \end{bmatrix} \tag{2.32}$$

$$= \begin{bmatrix} \sigma_{\tilde{\kappa}} & 0 \\ 0 & \sigma_{\tilde{\iota}} \end{bmatrix} \boldsymbol{S} \begin{bmatrix} \sigma_{\tilde{\kappa}} & 0 \\ 0 & \sigma_{\tilde{\iota}} \end{bmatrix} \tag{2.33}$$

$$\boldsymbol{S} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{2.34}$$

where ρ captures the correlation between item cutpoints and log discrimination parameters. I then specify priors for the scale terms, $\sigma_{\tilde{\kappa}}$ and $\sigma_{\tilde{\iota}}$, and the correlation matrix $\boldsymbol{S}$ separately. The scale terms $\sigma_{\tilde{\kappa}}$ and $\sigma_{\tilde{\iota}}$ are given weakly informative Half-Normal $(0, 1)$ priors, which provide weak regularization toward zero. The correlation matrix $\boldsymbol{S}$ gets a prior from the LKJ distribution, which is a generalization of the Beta distribution defined over the space of symmetric, positive-definite, unit-diagonal matrices, such as a correlation matrix (Lewandowski, Kurowicka, and Joe 2009). This is an increasingly approach to covariance matrix priors that provides more flexibility that inverse-Wishart distributions.[19]

$$\sigma_{\tilde{\kappa}}, \sigma_{\tilde{\iota}} \sim \text{Half-Normal}\,(0, 1) \tag{2.35}$$

$$\boldsymbol{S} \sim \text{LKJcorr}(\eta = 2) \tag{2.36}$$

The LKJ distribution has one shape parameter η that controls the prior distribution of ρ with a similar intuition as a Beta distribution: setting $\eta = 1$ produces a flat prior over correlation matrices, and creater values of $\eta > 1$ concentrate the prior for ρ toward a mode of 0. In

[19] Inverse-Wishart priors are often chosen for covariance matrices because they ensure conjugacy of the multivariate Normal distribution. Unlike with Gibbs sampling, conjugate priors provide no computational benefit for models estimated with HMC. Furthermore, the separation strategy weakens the dependence between prior scales and prior covariances, which is an advantage over inverse-Wishart priors (Akinc and Vandebroek 2018; Alvarez, Niemi, and Simpson 2014).

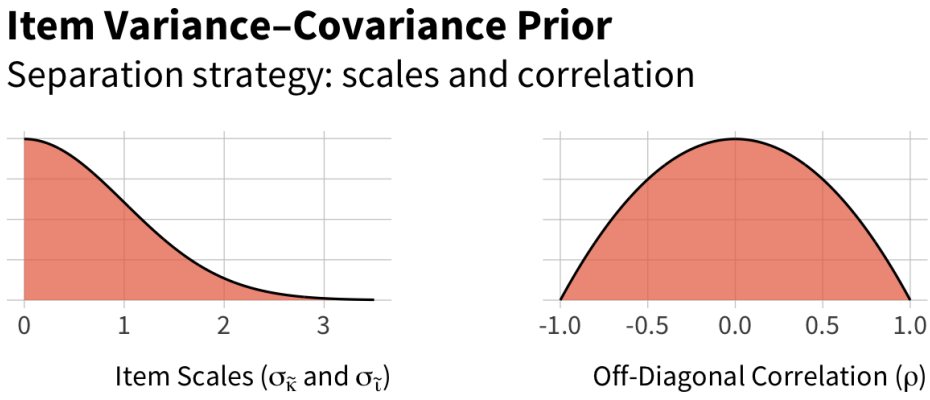
the limit, the prior for $\boldsymbol{S}$ is an identity matrix. The chosen value of $\eta = 2$ provides weak regularization against extreme correlations near ±1.

Figure 2.5 visualizes these details of the joint item prior. The top row shows the prior densities for the terms in the decomposed variance–covariance matrix $\boldsymbol{\Sigma}$. The left panel shows the Half-Normal prior density for the scale terms. The right panel shows the marginal distribution of ρ, the correlation term generated from the LKJ prior. The bottom panel plots 3,000 draws from the joint item prior. Each point represents a simulated item as a combination of unrestricted cutpoint values (on the horizontal axis) and unrestricted log-discrimination values (on the vertical axis). Brighter points indicate greater density among the random draws. As explained above, these parameters are restricted to sum to zero in each MCMC iteration to identify the ideological space.

2.4 Testing the Model with Simulated Data

Because the model is custom-built using Stan, it comes with no off-the-shelf quality assurances. To test the model's ability to recover unknown parameters, I estimate the model on simulated data. I intentionally stress-test the model by building a dataset that is smaller than the real dataset that I use for estimation. The fake data contain just 20 with 5 districts per state. With two parties per district, this totals 200 groups instead of the $435 \times 2 = 870$ groups in the real data. Individuals in each group offer responses to 40 items. For simplicity, the simulation assumes that each individual in each group answers only one question. When the model is estimated on real data, this assumption is relaxed by the weighting scheme laid out in Section 2.3.3. Because the weighting scheme downweights respondents who answer multiple survey items, I generate a small number of item responses for each item in each group: just 5 independent responses per item.

The model is designed to estimate group-level parameters from an individual-level data



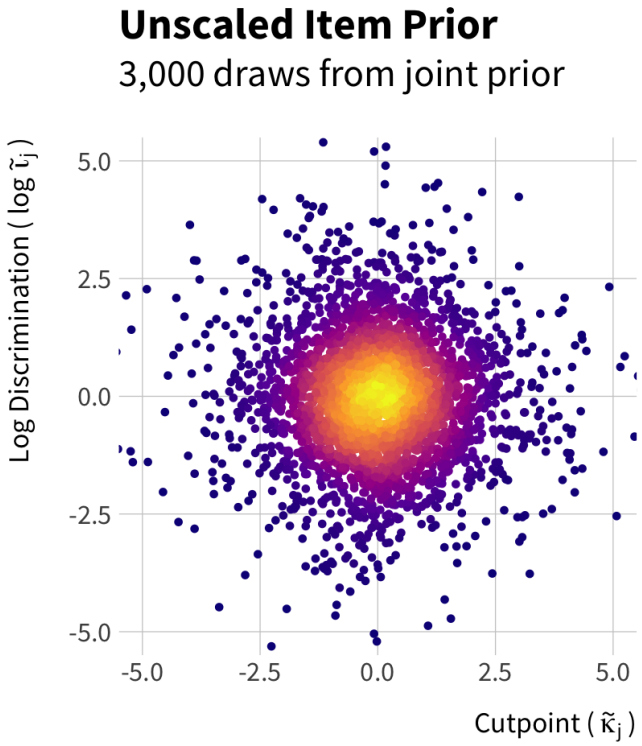


Figure 2.5: Components of the joint hierarchical prior for the unscaled item parameters. Left panel shows prior values for unscaled item parameters from the joint prior. Remaining panels show priors for decomposed covariance matrix components: including the standard deviation that form the matrix diagonal (middle) and the off-diagonal correlation from the LKJ prior (right).

generating process. As such, I begin by simulating individual-level item response data before aggregating the data to the group level for estimation. I simulate a universe containing 20 states nested within 5 regions, allocated so each region contains 4 states. Each state contains 5 districts containing voters belonging to 2 parties, totaling 200 groups across 100 districts in the whole "country." Data for each district-party group contain responses to 40 items that are answered by 5 unique individuals apiece.[20]

I draw item parameters from independent Normal distributions: Normal$(0, 1)$ for the unadjusted midpoint parameters and Normal $(0, 0.5)$ for the unadjusted log discrimination parameters. The hierarchical model does not contain a lot of data: I generate just 2 district-level covariates and 1 state covariate. District and state covariates are random draws from a Normal distribution with standard deviation 0.25, with separate coefficients for "Democrats" and "Republicans." Intercept values for Democrats and Republicans are -1 and 1, respectively. The random effects for districts, states, and 5 regions are Normal draws with standard deviations of 1, 0.1, and 0.1.

I estimate the model with MCMC, running 3 for 2000, using the first 1000 iterations for an adaptive warmup period. This results in 3,000 samples for each parameter in total.

Figure 2.6 compares each group's "true" ideal point to its estimated ideal point from the IRT model. Because the latent ideological space is not identified, I facilitate the comparison of ideal points by standardizing both sets to be mean zero and variance of 1. The model recovers a strong correspondence between the true and estimated ideal point values, underscoring the computational fidelity of the model and the statistical precision to detect meaningful variation in the underlying data.

The other important result from the fake data simulation is the item response functions (IRFs). Figure 2.7 plots IRFs from 10 randomly chosen items, showing 50 MCMC draws

[20] While a single survey will not contain 40 policy items, this study will pool items from several surveys into a final dataset.

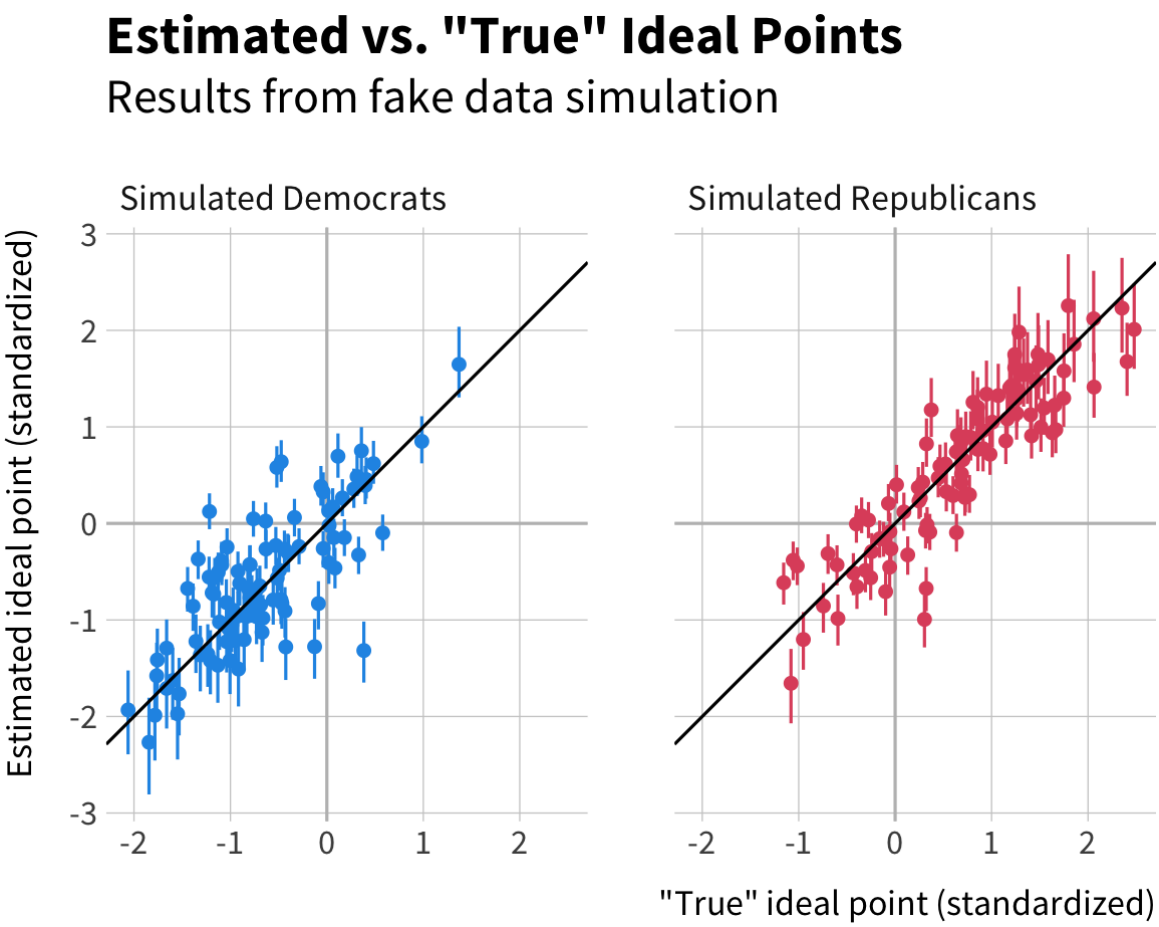


Figure 2.6: Ideal point estimates from fake data simulation. Estimates are plotted with posterior means and 90 percent compatibility intervals.

of the estimated IRFs (light orange) against the true IRF (black). These results also testify to the model's ability to recover key parameters even under weak data settings. In the few cases where the estimated IRF departs slightly from the true IRF, this appears to be due to partial pooling: the hierarchical item prior pooling an item parameter toward the estimated distribution of item parameters. In particular, the model appears to pooling high-magnitude midpoint parameters toward the mean of the distribution, which is restricted to be zero.

It is important to remember that the bias introduced by partial pooling is intentional. Partial pooling, like all forms of regularization, is intended to stabilize parameters that are estimated with weak signals from data. Without partial pooling, the model must estimate

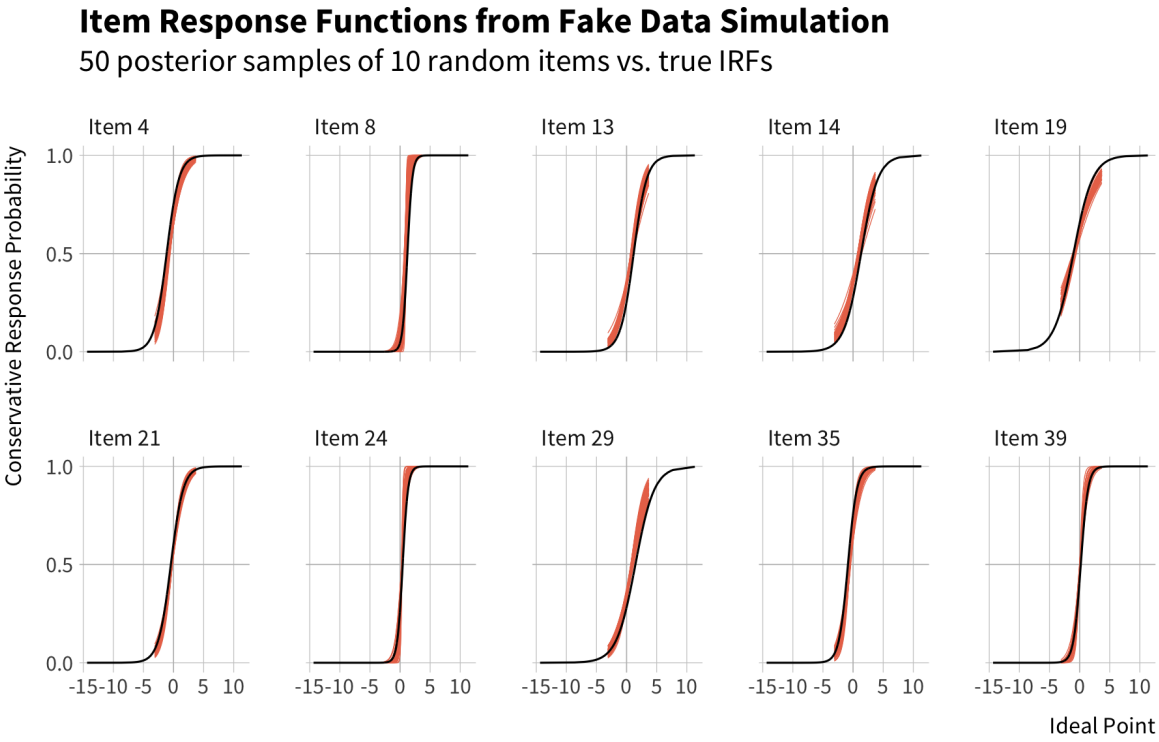


Figure 2.7: Item response functions from fake data simulation. Comparison between estimated (yellow) and true (black) item parameters for a random sample of items.

each item parameter with no "memory" of the other item parameters in the model. Although less regularization leads to less bias (all else equal), estimates will have higher variance especially in measurement models that have many parameters. Pooling toward hierarchical distributions can be particularly valuable for item response models, for which each additional observation introduces either two additional item parameters or an additional ideal point parameter (Bailey 2001).

2.5 Data Sources

2.5.1 Response data

Survey response data are drawn from the Cooperative Congressional Election Study (CCES) waves 2012, 2014, 2016, and 2018, and the 2016 wave of the American National Election Study

(ANES). I restrict the data to include only respondents who identify with the Republican and Democratic Parties and who reside in one of the 435 congressional districts.

I combine the responses to policy questions in these surveys into a single dataset, commonly called a "megapoll" (Kastellec, Lax, and Phillips 2010). The megapoll contains 95 policy items across all of the surveys it contains. Some items in different surveys are either worded identically or similarly enough to be considered as the same item, resulting in 61 unique items used for scaling. There are 15 items about economics, the federal budget, and health care; 7 items about the environment and climate change; 10 items on sex/gender equality and reproductive rights; 13 on immigration and national defense; and 16 items on other social issues such as race, policing, drug laws, and gun rights. If an item contained a response scale instead of a binary response option, I collapse responses into a binary coding to facilitate estimation with the probit model. A table describing the text and data sources of all items is included in Appendix 5.5.

Each survey contained responses from between 3,643 and 40,035 respondents, totaling 3,507,682 individual item responses across all surveys, respondents, and items. After aggregating respondents into district-party groups, the data contain 50,692 observations at the group-item level, which averages to approximately 58 item responses per group, or 831 responses per item.

2.5.2 Covariates

I draw district-level demographic data from Foster-Molina (2016), whose study of district demographics and congressional voting includes a publicly released time-series of congressional district demographics. I use observations from the 2012 election cycle, which represents is the first general election of the districting cycle for which I estimate ideal points. I measure the median income, median age, Gini coefficient, and the percent of the district population that was White, had a college degree, was born outside the U.S., and was unemployed.

State covariate data were included from the Correlates of State Policy Project (Jordan and Grossmann 2017). I include a state measure of the percent of the population that is Evangelical Christian, a measure that has been shown to predict aggregate opinion well (Buttice and Highton 2013; Lax and Phillips 2009), but that had no district-level measure for. I also measure the percent of the population that is non-White and per capita income at the state level, which are similar to variables I include at the district level but may capture different patterns at different levels of aggregation (e.g. Gelman et al. 2007).

The common worry that correlation among covariates inflates the variance of coefficient estimates is not of particular concern for this model, because covariates are included to improve the prediction of ideal points, not for inference on the regression coefficients themselves. Furthermore, prior distributions for the regression parameters serve a similar function as a ridge penalty, regularizing coefficients against noisy or weakly-identified solutions (Bishop 2006). To facilitate the specification of weakly regularizing priors across all covariates, I scale all covariates to have mean zero and variance 1 before estimation.

2.6 Model Results

I estimate the ideal point model using Stan's "No-U-Turn" sampling algorithm, an adaptive variant of Hamiltonian Monte Carlo. I draw posterior samples using 5 Markov chains that were each run for 2,000 iterations, discarding the first 1,000 iterations that are used for an adaptive warmup period.[21] I initialize the algorithm with an `adapt_delta` parameter of 0.9 and a maximum proposal tree-depth of 15. Following the advice of Link and Eaton (2011), I stored every post-warmup sample with no thinning of chains, resulting in a total of 5,000

[21] Given the size of the data, a chain of this length can run on a 2014 Macbook Pro in approximately three hours. To generate more chains, however, I use an external computing cluster affiliated with the Social Science Computing Cooperative at the University of Wisconsin–Madison.

samples per parameter across all chains.[22] Just 3 out of 5,000 iterations (0.06%) encountered a divergent transition, which indicates no systematic issues with model parameterization, and 0 iterations exceeded the maximum tree-depth.

The energy metrics that monitor the model's Hamiltonian mechanics also detect no problematic model behavior.

I present posterior summaries of the ideal point estimates in Figure 2.8. The figure features data from 870 district-party groups, including a posterior mean (black dot), a 50% compatibility interval (dark band) and a 90% compatibility interval (light band). The horizontal axis shows these estimates along the restricted ideological spectrum, which is oriented so that larger values indicate greater ideological conservatism. The vertical axis ranks the ideal points in ascending order, so the lowest value (the most progressive district-party group) gets the rank of 1.

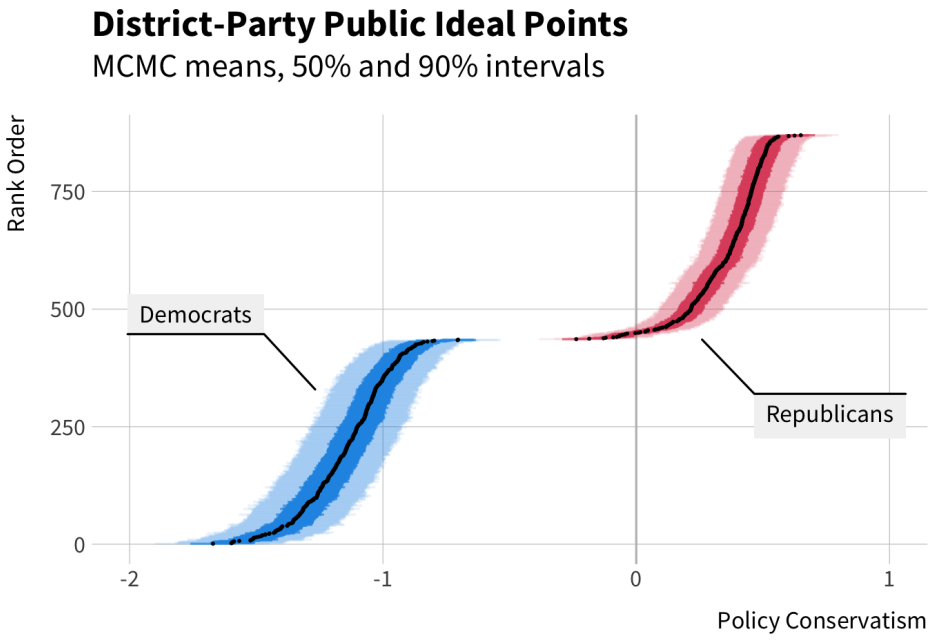


Figure 2.8: District-party ideal points and rank-ordering. Points are posterior means. Error bars are 50 and 90 percent compatibility intervals.

[22]The chains mix well and exhibit little autocorrelation, which is a credit to the model parameterization and the fact that the No-U-Turn algorithm is efficient at proposing transitions that explore the parameter space.

Unlike ideal point estimates for individual citizens, which may feature a great deal of overlap between the distribution of Republican and Democratic partisans, district-party ideal point estimates have no overlap whatsoever. This is because district-party ideal points are estimates of the *mean* ideology for Republican and Democratic groups, so they mask the ideological idiosyncrasies of individual citizens.

It is also important to note that the ideological space appears "asymmetrical"—the clusters of Republican and Democratic ideal points are not equidistant from zero. Instead, the Republican cluster is located much closer to an ideal point value of zero, with some Republican groups estimates to have negative ideal points. The appearance of asymmetry results from the way I restrict the item parameters to identify the latent ideal point space. In particular, the item midpoint parameters are restricted to sum to be mean zero, which means that the typical item presents policy alternatives that are equidistant from zero. The fact that Republican ideal points are clustered closer to zero suggests that Republicans are more likely to offer progressive item responses than Democrats are to offer conservative item responses, consistent with earlier research suggesting a general tendency of US citizens to hold progressive views on policy even if their symbolic worldview is more conservative (Ellis and Stimson 2012).

We can feel confident that these ideal point estimates capture real ideological variation by comparing them to other survey-based measures of ideology. Figure 2.9 compares the IRT model's district-party ideology estimates (posterior means) to a measure of ideological self-placement. Self-placement data are drawn from the Cooperative Congressional Election Study (CCES) for years 2012 through 2018, coding the 5-category item numerically and averaging partisan responses within each district-party. The figure shows a high degree of correlation between the IRT scores and the self-placement scores. The overall correlation between the ideal points is 0.98, and the within-party correlations are 0.64 among Democrats and 0.67 among Republicans. These within-party correlations are as strong as the within-party correlations between CF scores and DW-NOMINATE scores for incumbent House

members (Bonica 2013, 2014).

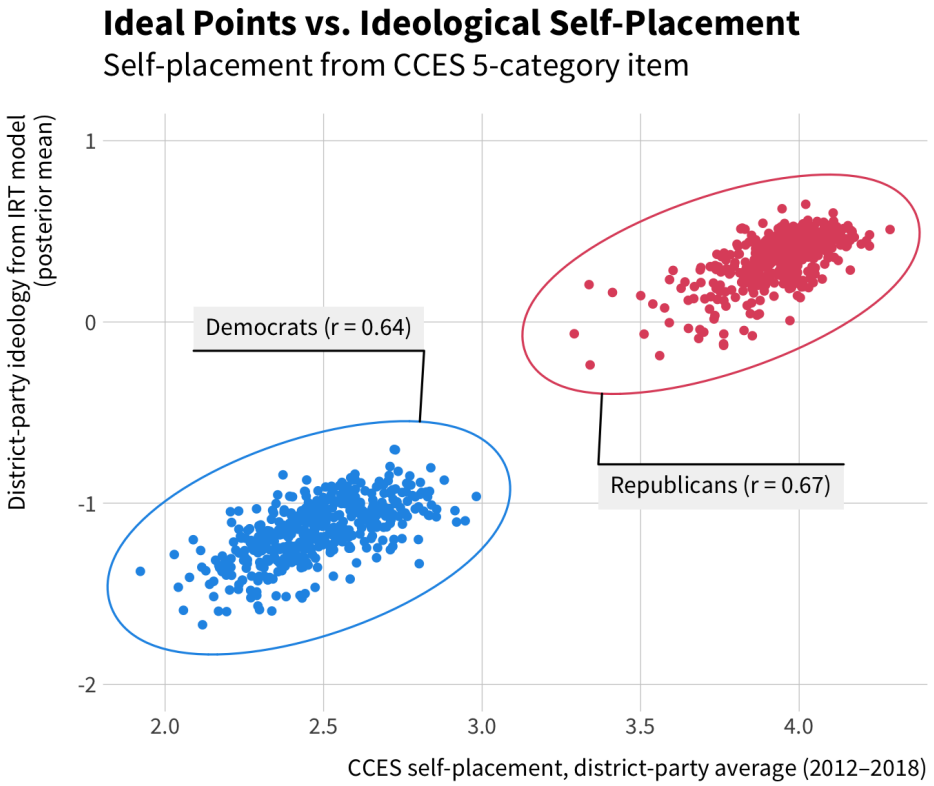


Figure 2.9: Comparison of district-party ideology and ideological self-placement. Average ideological self-placement in each district-party is measured using the CCES 5-category item, combining survey waves 2012 through 2018.

2.6.1 Ideological variation across districts

Recent scholarship on party coalitions has highlighted the ideological cohesion of the Republican Party compared to the Democratic Party. Scholars have proposed that Republicans are more ideologically cohesive than Democrats, owing to the fact that the Democratic Party is a big-tent assemblage of social groups with policy priorities that may conflict (Feldman and Zaller 1992; Grossman and Hopkins 2016; Lelkes and Sniderman 2016). Party differences in ideological cohesion may appear across districts as well, especially if the demographic composition of Democratic constituents is more heterogeneous from one district to the

next than the demographic composition of Republican constituents. Figure 2.10 provides some aggregate evidence that Republican constituencies are more ideologically similar to one another than Democratic constituencies are. The histograms in the figure show that the distribution of Democratic district-party ideal points is approximately symmetric around their mean, while Republican district-party groups are more tightly distributed around a modal ideology that is on the conservative edge of the ideal point scores.

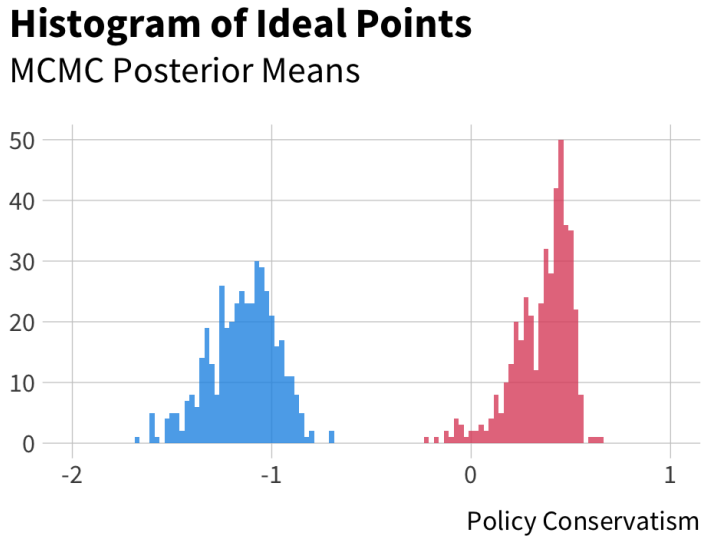


Figure 2.10: Histogram of ideal point means in both parties.

Figure 2.11 verifies this intuition by comparing the standard deviation of district-party ideal points in each party. The first panel shows standard deviation estimates for Democratic and Republican groups. The histogram represents the distribution of estimates across all MCMC sample iterations. Democratic groups tend to be more variable across districts, with a distribution of standard deviation estimates centered on a mean of 0.19. Republican groups are less variable across districts, with a distribution of standard deviations centered on a mean of 0.16. The second panel of Figure 2.11 plots the difference between the Democratic and Republican standard deviation estimates, again with a distribution representing all MCMC sample iterations. The histogram shows that almost all MCMC iterations (98%) contain ideal

point estimates that are higher variance for Democrats than for Republicans.

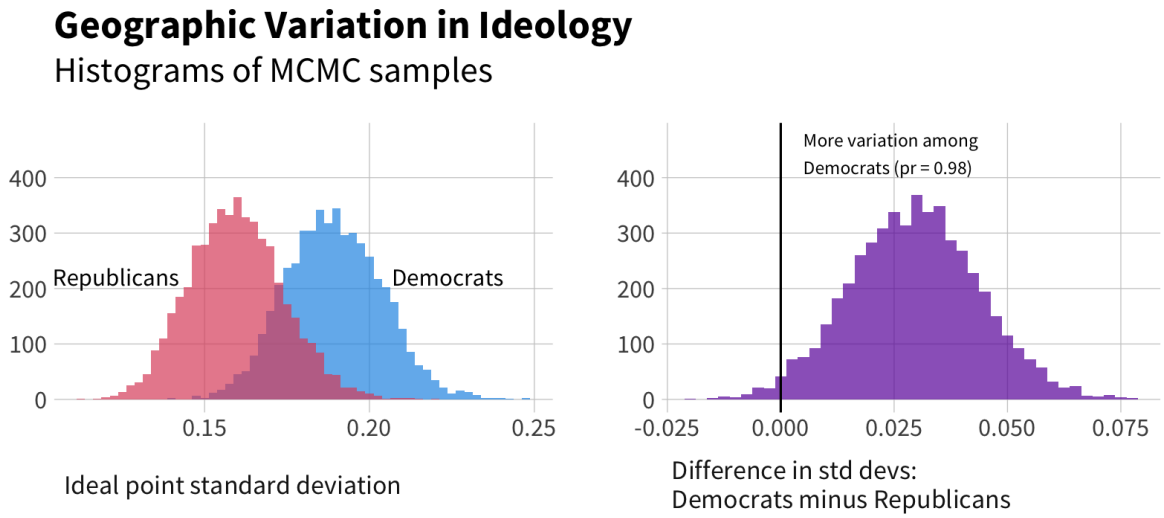


Figure 2.11: Geographic heterogeneity in ideal points. Left panel: standard deviation of Republican and Democratic district-party ideologies. Right panel: difference in standard deviations (Democrats minus Republicans). Distributions reflect MCMC draws.

Given the geographic variation in ideology, it is natural to wonder if Republican and Democratic ideal points are related. Do the districts containing more conservative Republicans also contain more conservative Democrats? Perhaps the pattern is reversed, where local conditions that reinforce the conservatism of Republicans actually reinforce progressivism among Democrats? Figure 2.12 explores this possibility. The left panel plots the Republican group ideal point in a district (vertical axis) against the Democratic group ideal point in the same district, with points representing posterior means. The posterior means do not exhibit much correlation to one another. The right panel plots a histogram of correlation estimates (Pearson's r) from all MCMC draws. The distribution of correlations suggests that there is a slight correlation between Republican and Democratic ideal points ($r > 0$ in 92% of MCMC draws), the correlation is quite small, with a posterior mean of 0.08 and a standard deviation of 0.06. The predominant takeaway is that most of the variation in Republican ideal points is unrelated to Democrat ideal points. Although this project does not explore the correlates or

possible causes of local ideological convergence or divergence with much detail, this project enables this research agenda by measuring local partisan ideology.

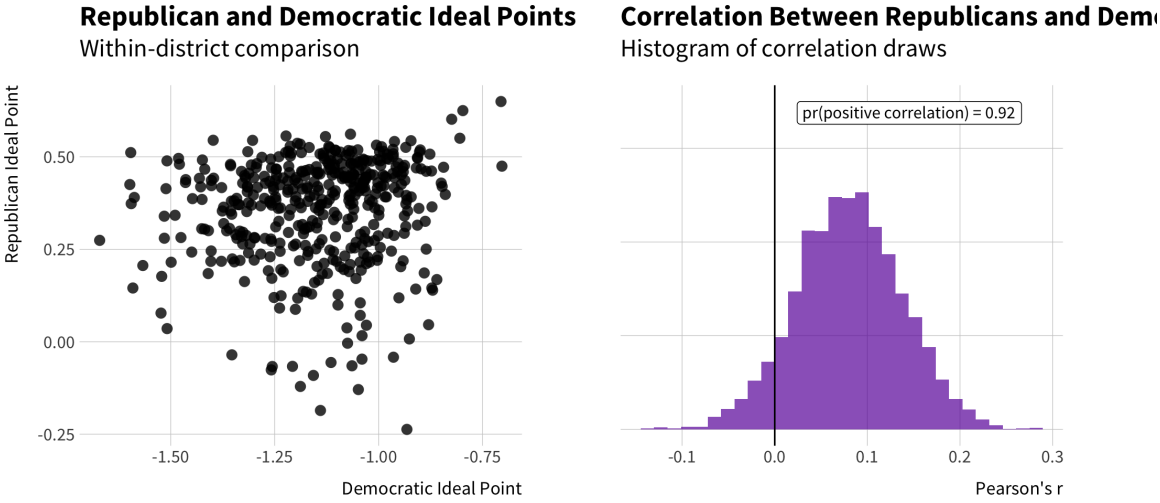


Figure 2.12: Correlation between Republican and Democratic ideal points in the same district. Left: scatterplot of Republican versus Democratic ideal points. Left: posterior distribution of the correlation (Pearson's r) between Democratic and Republican ideal points.

2.6.2 Aggregate correlates of district-party ideology

If Republican and Democratic ideal points are largely unrelated, it may be because district and state characteristics are related to partisan ideology in diverging ways. One final piece of descriptive analysis is to explore the relationship between district-party ideal points and the hierarchical covariates used to smooth the ideal point estimates in the model. Figure 2.13 visualizes the relationship between ideal points and aggregate covariates in two ways. The left side of the figure displays coefficient estimates from the IRT model for district- and state-level covariates included in the hierarchical regression. These coefficients capture the linear relationship between the covariates and ideal points, holding other covariates fixed. The right side of the figure plots the *bivariate* relationship between ideal points (posterior means) and a selection of district-level covariates, with no additional statistical adjustments. The bivariate relationships convey information about the "types" of districts that contain more

conservative or progressive partisans, and the coefficients convey whether these relationships may be statistical artifacts of other confounding relationships.

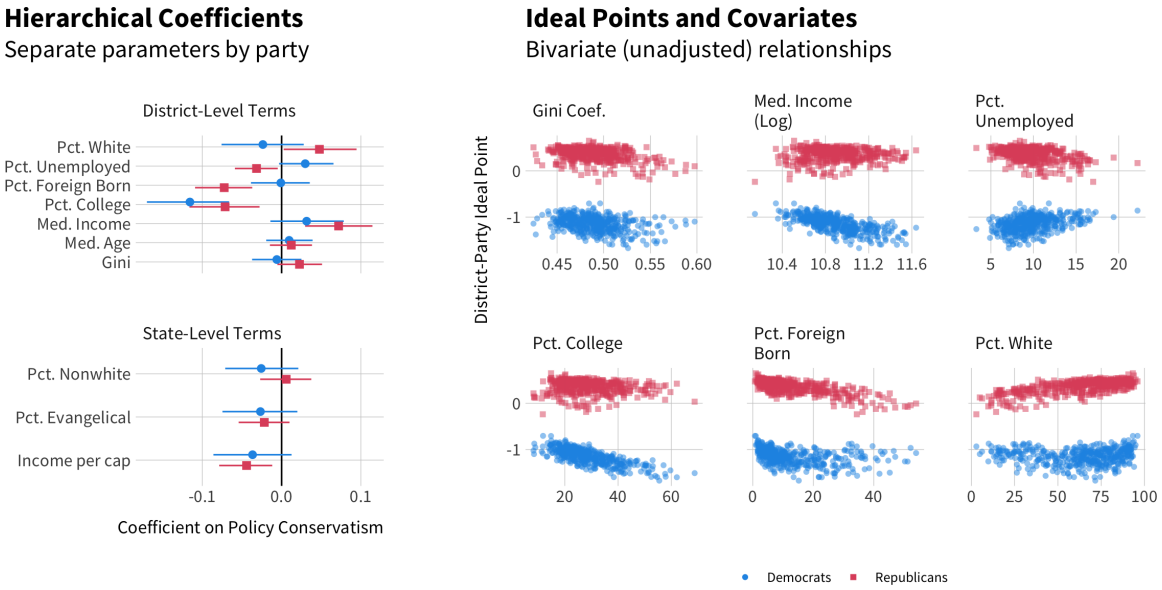


Figure 2.13: How ideal points relate to hierarchical covariates. Left: coefficients from hierarchical regression on district covariates (top left) and state covariates (bottom left), with 90 percent compatibility intervals. Right: Bivariate relationships between ideal points and a selection of district-level covariates with no statistical controls.

Larger values of the ideal point space are associated with greater conservatism, so increasing covariate values are related to increasing conversatism if the coefficient is positive and progressivism if the coefficient is negative. Some covariates have similar "effects" in both parties: districts with higher median income levels and median ages are estimated to be more conservative on average, and districts with greater college attainment are estimated to be more progressive. Other covariates have diverging effects for each party. Whiter districts are associated with greater conservatism among Republicans and (with less statistical confidence) greater progressivism among Democrats. Higher unemployment rates are related to decreased conservatism among Republicans and increased conservatism among Democrats. Districts with greater numbers of foreign born residents contain more progressive Republicans, which could be related to the cosmopolitanism of the district culture, and districts

with greater income inequality (measured by the Gini coefficient) contain more conservative Republicans. The uncertainty of the coefficients at the state level indicate weaker relationships to state-level covariates, but it is worth noting that Republicans and Democrats in wealthier *states* are more progressive than in less wealthy states, which is a pattern that differs from the district-level relationship of increasing conservatism greater wealth and is similar to the findings of Gelman et al. (2007) regarding state-level wealth. Counter-intuitively, I also find that larger evangelical populations appear to be related to greater progressivism among both Republicans and Democrats, and larger White populations are weakly related to greater progressivism among Democrats. Because these aggregate demographic features are likely to be causally related, it is important not to give these coefficients a causal interpretation. Furthermore, the correlations themselves may not be most straightforward if collider bias is introduced by conditioning on mediating variables.

The bivariate relationships between ideal points and district covariates convey what types of districts are more progressive or conservative, setting aside statistical adjustments. These relationships sometimes contrast with the model coefficients in interesting ways. For instance, although the coefficient for median income is positive in both parties, the scatterplot shows that higher-income districts contain more progressive Democrats. This suggests that other factors dominate the effect of income when determining the progressivism of Democrats. This factor could be college education, which is strongly related to progressivism among Democrats. The scatterplot also captures relationships that appear in the coefficients: the strongest relationships to Republican ideology are the White share of the population (positively related to conservatism) and the foreign-born share of the population (negatively related to conservatism). Among Democrats, greater unemployment is related to conservatism, and ideology is weakly related to racial composition, income inequality, and the foreign-born share of the population.

2.7 Improving the ideal point model

The results from this ideal point model are promising and informative, but there are some modeling approaches that could improve this model and others like it.

The model's predictive ability could be improved by creating more flexible hierarchical regressions. Because the hierarchical regression is included to smooth the ideal point estimates and not for causal interpretation, the linear specifications have no particular benefit over more flexible modeling approaches. The use of nonlinear models and "machine learning" approaches is new to hierarchical measurement modeling in political science. Examples include the use of regression tree methods for multilevel regression and poststratification (Bisbee 2019) and Gaussian process priors for IRT methods (Duck-Mayr, Garnett, and Montgomery 2020). Work outside of political science has also explored the use of spline regression for IRT models (Woods and Thissen 2006), which may not be as flexible as Gaussian processes but are computationally less intensive.

The substance of the model could also be extended in several ways aside from its predictive capacity. As mentioned above, Caughey and Warshaw (2015) lay out a dynamic linear model approach to bridge the ideal point space across multiple time periods. The parameterization of the model exposes a parameter for the variance of individual ideal points within a district, which could itself be modeled as a function of covariates (Lauderdale 2010). Lastly, because survey items vary in their response design, researchers have explored the use of ordinal response models to scale survey items that ask respondents to choose from more than two policy alternatives (e.g. Hill and Tausanovitch 2015).

— 3 —

Bayesian Causal Inference

I use the estimates of district-party public ideology from Chapter 2 to conduct two causal studies later in this project that, like the ideal point model, use a Bayesian modeling framework. While Bayesian methods are commonplace in ideal point modeling, the approach is almost entirely absent from causal inference work in political science. The purpose of this chapter is to explain and justify Bayesian causal modeling.

The discussion in this chapter highlights three primary contributions of Bayesian modeling for this project. First, I argue that causal inference is best understood as a problem of posterior predictive inference. Causal models are models for missing data: what we would observe if a treatment variable were set to a different value. Bayesian causal inference describes the plausible values of unobserved potential outcomes—or, more generally, the probability distribution of any causal estimand—given the data. This is how researchers think about causal inference, even if implicitly, almost all of the time.

Second, the Bayesian framework is a coherent method for quantifying uncertainty, which has several benefits for this thesis. District-party ideology, a key variable in this project, is not fully observed. It is only estimated up to a probability distribution using the measurement model in Chapter 2. Uncertainty about the causal effects of district-party ideology therefore

contain two sources of uncertainty: statistical uncertainty about counterfactual data, and *measurement uncertainty* about the observed values of district-party ideology before causal interventions. Bayesian analysis quantifies uncertainty in causal effects as if they were any other posterior quantity: by marginalizing the posterior distribution over all uncertain model parameters. This unified method of uncertainty quantification is also valuable for multi-stage causal analyses and flexible models with many correlated parameters, both of which appear in the causal analyses to follow.

Third, prior information often improves the estimation of causal effects. The empirical analyses in this project use priors for regularization: penalizing the complexity of a flexible model against overfitting. This is a common concern in the search for heterogeneous treatment effects, where the search for interactions or nonlinearities increases the number of potential false positive findings. Priors can encode other types of prior information, including structural information about possible data and modeling assumptions. I show how these sorts of priors can improve the precision of causal estimates and clarify how estimates are sensitive to prior assumptions.

This chapter unpacks these issues according to the following outline. I begin by reviewing the notation and terminology for causal modeling in empirical research, where data and causal estimands are posed in terms of "potential outcomes" or "counterfactual" observations. I then describe a Bayesian reinterpretation of these models, which uses probability distributions to quantify uncertainty about causal effects and counterfactual data. Bayesian methods are not heavily used in political science, so I spend much of the chapter explaining what a Bayesian approach to causal inference means with theoretical and practical justifications: how priors are inescapable for many causal claims, how priors provide valuable structure to improve the estimation of causal effects, and practical advice for constructing and evaluating Bayesian causal models. I provide examples of Bayesian causal modeling by replicating and extending published studies in political science: A Bayesian regression discontinuity analysis that shows

priors improve the precision and credibility of causal estimates, and a Bayesian meta-analysis that uses priors to highlight the consequences of modeling assumptions.

3.1 Overview of Key Concepts

3.1.1 Causal models

As an area of scientific development, *causal inference* refers to the formal modeling of causal effects, the assumptions required to identify causal effects, and research designs that make these assumptions plausible. Scientific disciplines, especially social sciences, have long been interested in substantiating causal claims using data, but the rigorous definition of the full causal model and identifying assumptions are what distinguish the current causal inference movement from other informal approaches. This section reviews causal inference by breaking it into a three-part hierarchy: causal models, causal identification, and statistical estimation.

The first level of the causal inference hierarchy is the *causal model*. The causal model is an omniscient view of a causal system that defines its mathematical first principals. The dominant modeling approach to causal inference in political science is rooted in a model of *potential outcomes* (Rubin 1974, 2005). This "Rubin model" formalizes the concept of a causal effect by first defining a space of potential outcomes. The outcome variable Y for unit i is a function of a treatment variable A. "Treatment" refers only to a causal factor of interest, regardless of whether the treatment is randomly assigned.[1] Considering a binary treatment assignment where $A = 1$ represents treatment and $A = 0$ represents control, unit i's outcome under treatment is represented as $Y_i(A = 1)$ or $Y_i(1)$, and the outcome under control would be $Y_i(A = 0)$ or $Y_i(0)$. The benefit of expressing Y in terms of hypothetical values of A allows the causal model to describe, with formal exactitude, the entire space of

[1] Some causal inference literatures refer to treatments as "exposures," which may feel more broadly applicable to settings beyond experiments. For this project, I make no distinction between treatments and exposures.

possible outcomes that result from treatment assignment as well as causal effects of treatment. The treatment effect for an individual unit, denoted τ_i, is the difference in potential outcomes when changing the treatment A_i.

$$\tau_i = Y_i(A_i = 1) - Y_i(A_i = 0) \tag{3.1}$$

This formulation generalizes to multi-valued treatments as well. If τ_i equals any value other than 0, then A_i has a causal effect on Y_i. Defining the causal model in terms of unit-level effects provides an exact, minimal definition of a causal effect: A affects Y if the treatment has a nonzero effect *for any unit*. A causal model may describe more complex features of a causal system, such as whether a unit complies with their treatment assignment, whether the unit's potential outcome depends on other variables, and so on.

Although the causal model perfectly describes the structure of a causal system, the model is only a hypothetical device. Because a unit can receive only one treatment, the researcher can observe only one outcome per unit. This renders the causal effect τ_i unidentifiable from data. This is the core philosophical problem in causal inference, and it means that no causal effects can ever be observed with data. Causal effects can only be *inferred* by layering on additional assumptions (Holland 1986).

Causal identification assumptions are the second level of the causal inference hierarchy. Identification assumptions specify the conditions under which observed data reveal what the data would look like if units received counterfactual treatments (Keele 2015). The implications of identification assumptions are typically posed in terms of *expectations* about potential outcomes that average over units, $\mathbb{E}\left[Y_i\left(A_i\right)\right]$, instead of unit-level potential outcomes. This is because it requires fewer assumptions to identify aggregate causal effects than to identify individual potential outcomes. Aggregate level causal effects, defined in terms of expectations over potential outcomes, are typically known as causal estimands. Example estimands include average treatment effects, conditional average treatment effects, local average treatment effects,

and so on.

The final layer of the causal inference hierarchy is *statistical estimation*. Identification assumptions describe minimally sufficient conditions for *nonparametric* identification of causal estimands (Keele 2015). Causal estimands are infinite-data expectations in perfectly defined covariate strata. Real data are often less convenient, with noisily estimated averages and continuous covariates whose strata often must be modeled in some way to make causal estimation feasible. There is no guarantee that linear regression models, or any parametric models, will correctly model the data and recover causal effects, so methodologists developing causal methods often seek methods that minimize additional statistical assumptions.

This hierarchy is helpful for organizing this chapter because it helps clarifies why researchers use certain research designs or statistical approaches to overcome particular problems with their data. Statistical assumptions can undermine identification assumptions (Blackwell and Olson 2020; Goodman-Bacon 2018; Hahn et al. 2018), which is why causal inference scholars tend to promote estimation strategies that rely on as few additional assumptions as possible (Keele 2015). One way to avoid these assumptions is to use research designs that eliminate confounding "by design" rather than through statistical adjustment, such as randomized experiments, instrumental variables, regression discontinuity, and difference-in-differences (for instance, Angrist and Pischke 2008). Research projects without those designs must invoke "selection on observables"— the statistical approach that assumes that confounders are controlled—although many methodological advancements in matching, semi-parametric models, and machine learning allow researchers to relax functional form assumptions in their statistical models (Hill 2011; Ratkovic and Tingley 2017a; Samii, Paler, and Daly 2016; Sekhon 2009). Causal inference is not synonymous with the new "agnostic statistics" movement (e.g. Aronow and Miller 2019), but it is animated by a similar motivation to identify statistical methods that rely on as few fragile assumptions as possible.

The three-part hierarchy is also useful because it clarifies where my contributions around

Bayesian causal estimation will be focused. As I discuss below, the "easiest way in" for Bayesian methods is through statistical estimation (level 3), since some flexible estimation methods are convenient to implement using Bayesian technologies (Imbens and Rubin 1997; Ornstein and Duck-Mayr 2020). I push this further by arguing that Bayesian analysis changes the interpretation of the causal model (level 1) by specifying probability distributions over the space of potential outcomes. This probability distribution allows the researcher to say which causal effects and counterfactual data are more plausible than others, which is a desirable property of statistical inference that is not available through conventional inference methods. The Bayesian approach also has the power to extend the meaning of identification assumptions (level 2) by construing them also as probabilistic rather than fixed features of a causal analysis (Oganisian and Roy 2020).

3.1.2 Bayesian inference

Bayesian inference is a contentious and misunderstood topic in empirical political science, so it is important to establish some foundations and intuitions before melding it with causal modeling. This section introduces Bayesian methods by skipping past the common descriptions that are often unhelpful and confusing—subjective probability, prior "beliefs," the posterior is proportional to the prior times the likelihood—and instead describes an "inside view" of Bayesian analysis on its own terms (McElreath 2017b).

Bayesian analysis uses conditional probability to conduct statistical inference. It begins with a joint probability for all variables in a model. In most cases these variables are denoted as data $\mathbf{y}$ and parameters $\boldsymbol{\pi}$, but in Bayesian analysis, the distinction between data and

parameters has only to do with which variables are observed or unobserved.[2]

$$p\left(\mathbf{y}, \boldsymbol{\pi}\right) = p\left(\mathbf{y} \cap \boldsymbol{\pi}\right) \tag{3.2}$$

The joint probability model represents the multitude of ways that the variables could be configured in the world. Conditioning on observed variables rules out many configurations of the unobserved variables, leaving behind only the unobserved variables that are consistent with observed data.

$$p\left(\boldsymbol{\pi} \mid \mathbf{y}\right) = \frac{p\left(\mathbf{y} \cap \boldsymbol{\pi}\right)}{p\left(\mathbf{y}\right)} \tag{3.3}$$

From this perspective, Bayesian analysis is "just counting" (McElreath 2017a)—counting the number of model configurations that are remain after conditioning on known information.

Bayes' Theorem is an expression for this conditioning process based on a particular factorization of the joint model,

$$p\left(\boldsymbol{\pi} \mid \mathbf{y}\right) = \frac{p\left(\mathbf{y} \mid \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi}\right)}{p\left(\mathbf{y}\right)} \tag{3.4}$$

which reveals how researchers commonly interface with Bayesian analysis: specifying a model for data conditional on parameters, $p\left(\mathbf{y} \mid \boldsymbol{\pi}\right)$, and a model for the marginal distribution of parameters, $p\left(\boldsymbol{\pi}\right)$. These models are often called the "likelihood" and "prior distribution," respectively.

The controversy surrounding Bayesian analysis arises from different perspectives about which constructs we choose to describe using probabilities. Researchers routinely model

[2]The semantic distinction between "data" and "parameter" is often sloppier in practice than many researchers would like to think. Many statistical analyses use aggregate estimates of lower-level processes as if they were known, such as per-capita income or the percentage of women who vote for the Democratic presidential candidate. These quantities are not knowable from finite data, and instead behave like random variables in that their values could differ under repeated sampling, so it might make sense to view their "true values" as parameters. From a Bayesian point of view, these are meaningless semantics, since both data and parameters are merely random variables modeled with probability distributions. The Bayesian view has a similar spirit to the Blackwell, Honaker, and King (2017) view of measurement uncertainty, where "measurement error" falls on a spectrum between fully observed data and missing data.

data given parameters, but many feel that modeling the marginal distribution of parameters is unscientific. This is because the marginal parameter distribution often represents "prior information" about which parameter values are plausible without observing the data. The inside view demystifies priors by acknowledging that a prior and a likelihood are fundamentally the same thing: using a probability distribution to quantify uncertainty about the value of a yet-unseen variable (McElreath 2017b). Likelihoods, in turn, are priors for data: assumptions that relate observed data to unobserved variables (Lemm 1996).[3] Using likelihoods to learn from data presents a similar epistemic problem as the fundamental problem of causal inference: assumptions are required for learning to be possible.

Bayesian updating, from the inside view, means considering a multitude of possible model configurations and pruning the configurations based on their consistency with the observed data. The prior model, $p(\mathbf{y} \mid \boldsymbol{\pi})p(\boldsymbol{\pi})$ describes an overly broad set of possible assumptions about the world. These assumptions include a distribution of possible data given parameters, $p(\mathbf{y} \mid \boldsymbol{\pi})$, and a distribution of possible parameters, $p(\boldsymbol{\pi})$. Bayesian updating decides which configurations of the world are more plausible based on how likely it would be to observe our data under those configurations. The plausibility of a parameter value—its posterior probability—is greater if the observed data are more likely to occur under that parameter value versus another value. In turn, the posterior distribution downweights parameter values that are implausible or inconsistent with the data (McElreath 2020, chap. 2). This is an important distinction from non-Bayesian statistical inference as conventionally performed in political science, which has no comparable notion of "plausible parameters given the data." As it connects to causal inference, this means that discussing "plausible causal effects" is not possible without a probability distribution over causal effects. The mission in the remainder of this chapter is to establish a framework for causal inference in terms of plausible effects

[3] We see this with meta-analysis as well. Results of a specific study generalize to a population by invoking assumptions about a specific study's relationship to the population. These assumptions can be encoded as priors, which I demonstrate at the end of this chapter.

and plausible counterfactuals.

3.2 Probabilistic Potential Outcomes Model

Having reviewed the basics of causal models and Bayesian inference, we now turn to a framework for Bayesian causal modeling. The distinguishing feature of a Bayesian causal model is that the elemental units of the model, the potential outcomes, are given probability distributions. This probability distribution reflects available causal information that exists outside the current dataset. Bayesian inference proceeds by updating our information about causal effects and counterfactual potential outcomes in light of the observed data. This section introduces this modeling framework at a high level, provides a probabilistic interpretation and notation for potential outcomes modeling and describes how the Bayesian framework affects the "hierarchy of causal inference."

As with other causal models, we begin at the unit level. Unit i receives a treatment $A_i = a$, with potential outcomes $Y_i(A_i = a)$. Suppose a binary treatment case where A_i can take values 0 or 1, so the unit-level causal effect is $\tau_i - Y_i(1) - Y_i(0)$. Although τ_i is unidentified, it is possible to estimate population-level causal quantities by invoking identification assumptions. For instance, the conditional average treatment effect at $X_i = x$, $\bar{\tau}(X = x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$, can be estimated from observed data assuming no hidden treatments, no interference, conditional ignorability, and positivity (Rubin 2005). Suppressing the unit index i,

$$\begin{aligned} \bar{\tau}(X = x) &= \mathbb{E}[Y(A = 1) - Y(A = 0) \mid X = x] \\ &= \mathbb{E}[Y(A = 1) \mid X = x] - \mathbb{E}[Y(A = 0) \mid X = x] \\ &= \mathbb{E}[Y \mid A = 1, X = x] - \mathbb{E}[Y \mid A = 0, X = x] \end{aligned} \tag{3.5}$$

where the third line is obtained by the identification assumptions. The identification assumptions connect *causal estimands* and what I will call *observable estimands*. Causal estimands

are the true causal quantities, but they are unobservable because they are stated as contrasts of potential outcomes. Observable estimands are the observable analogs of causal estimands and are equivalent to causal estimands only if identification assumptions hold. Other literature refers to observable estimands as "nonparametric estimators" (Keele 2015), but I steer clear of this language because the distinction between observable estimands and estimators is important for understanding the contributions of the Bayesian causal approach.

The transition to a Bayesian probabilistic model begins with an acknowledgment that no estimate of the observable estimand, $\mathbb{E}\left[Y \mid A = a, X = x\right]$, will be exact. The assumptions identify causal effects only in an infinite data regime where the observable estimand is known exactly. Inference about causal effects from finite samples, however, requires further statistical assumptions that link the observable estimand to an estimator or model. Let $f(A_i, X_i, \boldsymbol{\pi}) + \varepsilon_i$ be a model for Y_i consisting of a function $f(\cdot)$ of treatment A_i, covariates X_i, and parameters $\boldsymbol{\pi}$, and an error term ε_i where $\mathbb{E}\left[\varepsilon_i\right] = 0$. This setup is similar to any modeling assumption that appears in observational causal inference to link an estimator to the observable estimand, including parametric models for covariate adjustment, propensity models, matching, and more (Acharya, Blackwell, and Sen 2016; Sekhon 2009). We use the statistical model to estimate the CATE, $\hat{\tilde{\tau}}(X = x)$, by differencing these model predictions over the treatment.

$$\begin{aligned} \tilde{\tau}(X = x) &= \mathbb{E}\left[Y \mid A = 1, X = x\right] - \mathbb{E}\left[Y \mid A = 0, X = x\right] \\ &= \mathbb{E}\left[f(A_i = 1, X_i = x, \boldsymbol{\pi}) - f(A_i = 0, X_i = x, \boldsymbol{\pi})\right] \end{aligned} \tag{3.6}$$

The Bayesian approach, inspired largely by Rubin (1978), constructs $f()$ as a joint model for data and parameters: $p\left(Y, \boldsymbol{\pi}\right) = p\left(Y \mid f(A, U, \boldsymbol{\pi})\right) p\left(\boldsymbol{\pi}\right)$. The data are distributed conditional on the model prediction $f()$, which is a function of parameters $\boldsymbol{\pi}$. The parameters also have a prior distribution $p\left(\boldsymbol{\pi}\right)$, or a distribution marginal of the data. These models for data and parameters are added statistical assumptions on top of causal identification assumptions. The data model is similar to any estimation approach that uses a probability model for errors

(e.g. any MLE method or OLS with Normal errors). The parameter model has no analog in OLS or unpenalized MLE, but this added statistical assumption will be leveraged as a major benefit as we explore Bayesian causal estimation below.

The joint generative model is sufficient to characterize the probability distribution for the conditional average treatment effect as defined in Equation (3.6),

$$p(\tilde{\tau}(X = x)) = \int p\left[f(A = 1, X = x, \boldsymbol{\pi}) - f(A = 0, X = x, \boldsymbol{\pi}), -\boldsymbol{\pi}\right] \mathrm{d}\boldsymbol{\pi} \tag{3.7}$$

which is the probability distribution of model contrasts for $A = 1$ versus $A = 0$. Integrating over $\boldsymbol{\pi}$ in Equation (3.7) marginalizes the distribution with respect to the uncertain parameters. Because the marginalized parameters are distributed according to the prior $p(\boldsymbol{\pi})$, the expression in (3.7) represents a prior distribution for the CATE. This is an inherent feature of the Bayesian approach: probability distributions of causal quantities even before data are observed. Conditioning on the observed data returns the posterior distribution for the CATE,

$$p(\tilde{\tau}(X = x) \mid Y) = \int p\left[f(A = 1, X = x, \boldsymbol{\pi}) - f(A = 0, X = x, \boldsymbol{\pi}), \boldsymbol{\pi} \mid Y\right] \mathrm{d}\boldsymbol{\pi} \tag{3.8}$$

which marginalizes over the parameters after conditioning on the data.

Causal models, at their core, are models for counterfactual data. Because Bayesian models are a *generative* model for parameters and data, they contain all machinery required to directly quantify counterfactual potential outcomes using probability distributions. Bayesian causal models facilitate probabilistic causal inference at the unit level by generating posterior distributions for counterfactual observations. To see this in action, we start by acknowledging that we can use any joint model to generate a predictive distribution for data Y from fixed model parameters (McElreath 2017a). Denote these generated observations as $\tilde{Y}$ to distinguish them from the data observed Y. If we average this predictive distribution $p\left(\tilde{Y} \mid \boldsymbol{\pi}\right)$ over the prior distribution of parameters, we obtain a "prior predictive distribution"—the distribution

of data we would expect under the prior (Gelman et al. 2013).

$$p\left(\tilde{Y} \mid A = a, X = x\right) = \int p\left(\tilde{Y} \mid A = a, X = x, \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi}\right) \, \mathrm{d}\boldsymbol{\pi} \tag{3.9}$$

If we condition on the observed data before generating new observations, this is called a "posterior predictive distribution"—the distribution of data that we expect from the posterior parameters.

$$p\left(\tilde{Y} \mid Y, A = a, X = x\right) = \int p\left(\tilde{Y} \mid A = a, X = x, \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi} \mid Y\right) \, \mathrm{d}\boldsymbol{\pi} \tag{3.10}$$

These predictive distributions are the basis for out-of-sample inference in any Bayesian generative model.[4] Invoking the causal identification assumptions, we generate a predictive distribution for counterfactual data as well by setting the treatment A to some other value $A = a'$. Denote these counterfactual predictions $\tilde{Y}_i'$, which I subscript i to show that this model implies a probability distribution for individual data points as well as aggregate treatment effects. The posterior predictive distribution for counterfactual data is

$$p\left(\tilde{Y}_i' \mid Y, A_i = a', X_i = x\right) = \int p\left(\tilde{Y}_i' \mid A_i = a', X_i = x, \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi} \mid Y\right) \, \mathrm{d}\boldsymbol{\pi}, \tag{3.11}$$

which is sustained by causal identification assumptions as well as a distributional assumption for data given parameters.

Bayesian causal models can be so summarized: if a causal model defines a space of potential outcomes, then a Bayesian causal model gives potential outcomes a probabilistic representation. Probability densities over potential outcomes are defined in the prior and in the posterior, and they can be defined all the way to the unit level if the generative model

[4] Simulations of this sort are possible under any likelihood-based model that posits a generative probability distribution for the data, but Bayesian predictive distributions marginalize over the parameter distribution instead of conditioning on fixed parameters. This makes Bayesian predictive distributions a more complete accounting of statistical and epistemic sources of uncertainty.

contains a probability distribution for unit data.[5] In short, the Bayesian view of causal inference is a missing data for unobserved means or unobserved counterfactuals—a view that is at least as old as Rubin (1978).[6] Bayesian methods for causal inference have appeared in political science only sporadically in the decades since (e.g. Horiuchi, Imai, and Taniguchi 2007; Green et al. 2016; Ornstein and Duck-Mayr 2020).

3.2.1 Why Bayesian causal modeling?

A Bayesian view of causal inference is possible, but why it is valuable? This section describes several benefits that are related to this project, although other projects could certainly find other benefits. In short: Bayesian methods facilitate direct inference about plausible causal effects without notions of repeated sampling, which makes it valuable to observational data often used in political science. Probability distributions provides a convenient interface for incorporating uncertainty in multi-stage estimation routines, data with measurement error, and flexible models with many correlated parameters, all of which appear in subsequent chapters.

The Bayesian causal approach is sensible for causal inference because it facilitates direct, probabilistic inference about treatment effects given the data: which effect sizes are more likely or less likely than others. While p-values and confidence intervals are often misused to make probabilistic statements about parameters, the posterior distribution and posterior intervals actually enable the researcher to state the probability of substantive treatment

[5] Some modeling approaches can estimate average causal effects with group-level statistics only, eliding the unit-level model altogether. This can weaken the model's dependence on parametric assumptions for units, falling back onto more dependable parametric assumptions for the statistics, e.g. the Central Limit Theorem for group means. A model of this type will naturally stop short of defining probability distributions for counterfactual units, but it does define probability distributions for counterfactual means. In some cases, such as binary outcome data, means in each group are sufficient statistics for the raw data, so the unit level model is implied by the group-level model.

[6] In more general modeling contexts beyond causal inference,Jackman (2000) makes a similar argument that all estimates, inferences, and goodness-of-fit statistics can be unified as functions of missing data, with Bayesian posterior sampling as a natural way to describe our information about these functions.

effects, negligible effects (Rainey 2014), and more. Positivistic statements about plausible causal effects are a natural way to discuss the results of causal research: "the world probably works in this way, given the evidence."[7] This language requires probability distributions over parameters, which are the distinguishing feature of Bayesian methods. Non-Bayesian methods, meanwhile, formally conduct inference about the plausibility of *data* given fixed parameters; inferences about parameters is indirect and requires an additional layer of decision theory. Non-Bayesian inference can be awkward as a result—for instance, using a p-value to claim that data are inconsistent with a null hypothesis that the researcher never thought was credible to begin with (Gill 1999). Restated more formally, there non-Bayesian researcher routinely conducts inference by estimating $p\left(\mathbf{y} \mid \text{Null Hypothesis}\right)$, when they are usually more interested in $p\left(\text{Alternative Hypothesis} \mid \mathbf{y}\right)$.

Probabilistic inference for parameters is especially valuable when the observed data represent the entire population, which is common for observational causal inference in political science. Historical data have no possibility to be resampled from the broader population, so estimators cannot inherit their statistical properties their sampling distributions (Western and Jackman 1994).[8] The foundations of uncertainty in Bayesian inference, meanwhile, are probability distributions that represent imperfect pre-data information about the generative processes underlying the variables in a model. Whether this imperfect information corresponds to sampling randomness or other epistemic uncertainty can be subsumed in the Bayesian framework (Rubin 1978).

It is common for advocates of Bayesian inference to celebrate the fact that the posterior

[7] Rubin writes, in the context of causal inference, that "a posterior distribution with clearly stated prior distributions is the most natural way to summarize evidence for a scientific question" (Rubin 2005, 327).

[8] Causal researchers have been exploring a "design-based" uncertainty framework, where randomness in treatment assignment is a source of uncertainty, as a non-Bayesian alternative to population resampling (Abadie et al. 2020; Keele, McConnaughy, and White 2012). This approach is uncommon except among researchers on the cutting edge of "agnostic" statistical practices (Aronow and Miller 2019). Bayesian statisticians remain interested in the frequency properties of their methods such interval coverage (Rubin 1984), which partially motivates an interest in "objective Bayesian inference" (Berger 2006; Fienberg 2006).

distribution quantifies uncertainty in all random variables simultaneously, but this is especially useful for causal methods that entail multiple estimation steps. Multi-stage procedures require estimates from one stage to serve as inputs in other stages, introducing additional measurement error into the estimates from later results. These multi-stage procedures are common in causal inference and include instrumental variables, propensity score weighting, synthetic control, and other structural models (Acharya, Blackwell, and Sen 2016; Angrist and Pischke 2008; Blackwell and Glynn 2018; Imai et al. 2011; Xu 2017). Bayesian methods combine all estimation stages into one joint model, so a Bayesian treatment effect estimate will naturally reflect uncertainty in all model stages by marginalizing the posterior distribution over all "design stage" parameters (Liao 2019; McCandless, Gustafson, and Austin 2009; Zigler and Dominici 2014). Joint modeling is similar to "uncertainty propagation" methods that use numerical approaches to simulate early-stage uncertainty in later-stage models. Unlike uncertainty propagation, however, a fully Bayesian model effectively treats early-stage estimates as priors for later-stage estimates, so all uncertain parameters are updated using full information from all stages of the model (Liao and Zigler 2020; Zigler 2016; Zigler et al. 2013).

The combined modeling approach is important for this project because the key independent variable, district-party public ideology, is an uncertain estimate from a measurement model. Estimates for the causal effect of district-party ideology therefore contain two sources of error: statistical uncertainty about the causal effect itself, and measurement error in the underlying data. Building a combined model to estimate ideal points and causal effects simultaneously would be logistically overwhelming, but the full model can be approximated by drawing ideal points from a prior in the causal analyses. This is a method that I implement in Chapter 4.

One final justification for Bayesian causal modeling is that prior information is everywhere. This is a longer discussion that I untangle in Section 3.3, but to preview, priors matters

for the way researchers think about their modeling decisions, and they affect the inferences that researchers draw from data, even if they wish to avoid explicit Bayesian thinking about their analyses.

3.2.2 Bayesian modeling and the hierarchy of causal inference

This section interprets the Bayesian causal inference framework in light of the "hierarchy of causal inference" described in Section 3.1.1. The hierarchy helps us account for the ways that Bayesian methods have already been invoked for causal inference in political science and in other fields, and it helps us understand how the Bayesian statistical paradigm reinterprets causal inference more broadly. To review, the hierarchy consisted of three parts:

1. The causal model: definition of potential outcome space, causal estimands expressed in terms of potential outcomes.
2. Identification assumptions: linkage from causal estimands expressed as potential outcomes to observable estimands expressed using observed data.
3. Estimation: Methods for estimating observable estimands with finite data.

We began our discussion of the Bayesian causal model above by considering a plug-in estimator for an observable estimand that came from a Bayesian statistical model. Bayes was invoked as "mere estimation," so we began our understanding of Bayesian causal modeling at level 3 of the hierarchy. As only an estimation method, a Bayesian estimator (such as a posterior expectation value) doesn't obviously change the meaning of the observable estimand or the causal estimand. We can evaluate the Bayesian model for its bias and variance like any other estimator.

The realm of "mere estimation" is where many Bayesian causal approaches appear in political science and other fields. The estimation benefits of Bayes tend to fall into three categories: priors provide practical stabilization or regularization, posterior distributions

are convenient quantifications of uncertainty, or MCMC provides a tractable way to fit a complex model. "Mere estimation" regards Bayesian inference as practically valuable but theoretically unnecessary, since researchers might prefer non-Bayesian means to the same ends. Examples include the use of Bayesian Additive Regression Trees (or BART, Chipman, George, and McCulloch 2010; Hill 2011) for heterogeneous treatment effects (Green and Kern 2012), Gaussian processes for smooth functions in regression discontinuity (Ornstein and Duck-Mayr 2020) and augmented LASSO estimators (Park and Casella 2008; Tibshirani 1996) for sparse regression methods (Ratkovic and Tingley 2017a, 2017b).[9] These methods use priors to regularize richly parameterized functions and MCMC to estimate models, but the theoretical implications of Bayesian causal estimation are not a major focus.

What does it mean for Bayesian estimation to have theoretical implications for causal inference? This brings our focus to level one of the causal inference hierarchy: the model of potential outcomes. Any estimation method that invokes Bayesian tools requires a prior for model parameters, which imply prior densities on causal estimands. If the joint model contains a unit-level data model as well, which is the case for most regression approaches, then unit potential outcomes also have prior probability densities: some potential outcomes are more plausible than others, even before seeing data. This is a decisive theoretical departure from a non-Bayesian approach to causal modeling, where potential outcomes and causal effects are merely defined. The benefit of this departure is the ability to specify posterior distributions for unobserved potential outcomes directly, which a few recent methodology papers in political science have invoked for missing data due to noncompliance (Horiuchi, Imai, and Taniguchi 2007), synthetic control estimation (Carlson 2020) and regression settings (Ratkovic and Tingley 2017a). But these papers do not highlight the fact that these methods also imply *priors* for counterfactuals. As a result, skeptical applied researchers

[9] A recent, notable example from economics is Meager (2019), who uses Bayesian random effects meta-analysis to aggregate evidence from micro-credit experiments. See Rubin (1981) for an introduction to Bayesian meta-analysis of experiments.

have little guidance for understanding what it means to have priors on counterfactual data, theoretically or practically. I discuss priors in more detail in Section 3.3.

Priors do not have to be an inconvenience. There are many scenarios where priors can relax assumptions, building robustness checks directly into a statistical model. This is how Bayesian inference affects layer two of the hierarchy: identification assumptions. By their nature, identification assumptions can never be validated by consulting the data, so most causal inference research projects simply condition the analysis on the identification assumptions holding. A Bayesian model can relax these assumptions by instead posing them as priors that reflect the researcher's reasonable expectations about the remaining biases in a research design (Oganisian and Roy 2020). This generalizes the notion of "sensitivity tests" for measuring the robustness of causal inferences to violated identification assumptions (e.g. Imai et al. 2011; Acharya, Blackwell, and Sen 2016) by *marginalizing over* these sensitivity parameters instead of conditioning on fixed, stipulated values. One recent political science example of this approach is Leavitt (2020), who frames the parallel trends assumption in a difference-in-differences design as a prior over unobserved trends. This introduces an additional layer of "epistemic uncertainty" into Bayesian causal inference that is ordinarily assumed to be zero. For a more general discussion of identification assumptions as priors, see Oganisian and Roy (2020).

3.3 Understanding Priors in Causal Inference

The distinguishing feature of Bayesian analysis that attracts most of its controversy is the prior distribution over model parameters. At the same time, priors also deliver most of the benefits of Bayesian modeling. This section unravels several common confusions about priors as they relate to modeling in general and especially for causal inference. What do priors do, and how can they be used responsibly? I have two goals with this section. The first is to undermine the

view that flat priors are sensible default choices. Flat priors are not always uninformative, and uninformative priors are not always flat, depending on the parameter space being considered. The second goal is to provide guidance for specifying priors that supersedes

3.3.1 Information, belief, and data falsification

Bayesian analysis is often characterized as overly subjective. If priors are a way for researchers to insert their "beliefs" into a statistical analysis, what is the point of data? Some have argued that Bayesian analysis with informative priors is analytically equivalent to "data falsification" because priors and data influence the posterior distribution through the same mechanism: adding information to the log posterior distribution (García-Pérez 2019).

This hesitation can be eased with two lines of thought. First, it is helpful to think of priors as *information*, not "belief." A prior is any assumption that brings probabilistic information into a model. This is not unique to Bayesian models, since all likelihood functions represent probabilistic assumptions about data as well. This project regards priors as "beliefs" in a pragmatic sense only. They are "belief functions" in the sense that they represent the support for a parameter value *within a model*, but the researcher is "morally certain" that the model is wrong, so their degree of belief in a prior is actually zero (Gelman and Shalizi 2013, 19–20). Priors, like other model assumptions, represent reasonable approximations that impose structure on the information obtained from data. Researchers often care about other pragmatic consequences of priors such as the frequency properties of their estimates, noninformative priors for optimal learning from data (Berger 2006; Fienberg 2006; Rubin 1984), and model-building workflow practices even if they "fall outside the scope of Bayesian theory" (Betancourt 2018; Gabry et al. 2019; Gelman and Shalizi 2013).

3.3.2 Flatness is a relative, not an absolute, property of priors

Researchers commonly encounter Bayesian methods to solve an inconvenient estimation problem but would like to avoid the difficulty of specifying priors. It is common for these researchers to err toward "flat" or "diffuse" priors that assign equal or nearly equal probability to all possible values. This feels "least biased" because the Bayesian model most-closely represents an unpenalized maximum likelihood model, with which the researcher is more familiar. One common Bayesian argument against flat priors is that they understate the researcher's actual prior information—an argument that is simultaneously obvious yet uncompelling. Instead, I argue that flat priors often lead researchers to misunderstand what their models actually say. If a parameter has a flat prior, functions of that parameter are not guaranteed to have a flat prior. Furthermore, flat priors are only flat with respect to a particular parameterization of a model. If the parameterization changes, a flat prior will not represent the same prior information. In general, the researcher must understand how data functionally depend on parameters in order to understand the consequences of priors that attempt to be non-informative.

To understand the consequences of prior choices, it is essential to understand the *implied prior*. Suppose we have a parameter π and a function of that parameter, $h(\pi)$. If π has a prior density, then $h(\pi)$ has an implied prior density, which is affected by the density of π and the function $h(\cdot)$. Consider a simple example where π is distributed Normal $(0, 1)$, and $h(\pi) = 3 + 2\pi$. The implied prior for $h(\pi)$ is Normal $(3, 2)$, which is shown in Figure 3.1. Importantly, note that for a given value of π, the density of π is almost certainly not equal to the density of $h(\pi)$. This shows that functions of parameters have prior density, but the density of the function will almost certainly differ from the density of the original parameters.

Implied priors help us understand the unintended consequences of flat priors by highlighting circumstances where flat priors, believed to be reasonable and conservative, create

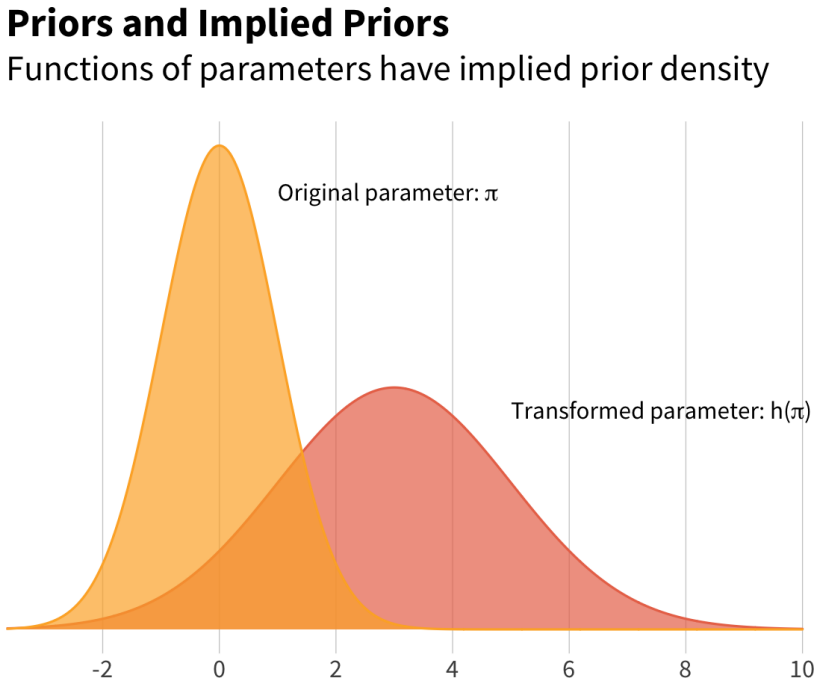


Figure 3.1: If a parameter has a density, a function of the parameter also has a density that is almost always unequal to the original density.

problematic data (Seaman III, Seaman Jr, and Stamey 2012). Consider a binomial random variable that counts X successes out of N independent trials with success probability α. We represent prior ignorance about α using a flat Beta $(1, 1)$ density. Now consider the identical model but reparameterized as a logit model, which is a for estimating α with covariates. The logit model introduces the parameter η, the logit-scale equivalent of α.

$$\begin{aligned} X &\sim \text{Binom}(N, \alpha) \\ \text{logit}(\alpha) &= \eta \end{aligned} \tag{3.12}$$

How do we put a prior on η that represents diffuse information about X? We follow a default instinct and give η a "vague" prior with a wide variance, Normal$(0, 10)$. If we take both of these models and generate 100,000 prior simulations for $X \sim \text{Binom}(N, \alpha)$, depicted as histograms in Figure 3.2, the implied priors for X do not resemble one another at all. The first panel shows the implied prior for X when α has a flat Beta prior, resulting in a distribution

for X that is also flat. The middle panel shows the implied prior for X when η has a wide Normal prior, resulting in a prior that concentrates X at very small and very large values. This is because the Normal $(0, 10)$ prior places most probability density on η values that represent unreasonably small or large α values. Only a thin range of logit-scale values map to probabilities that we routinely encounter in political science: logit values between -3 and 3 correspond to probabilities between 0.047 and 0.953. In order to obtain a flat prior for α using a logit model, we would actually use the prior $\eta \sim \text{Logistic}(0, 1)$, shown in the third panel of Figure 3.2.[10]

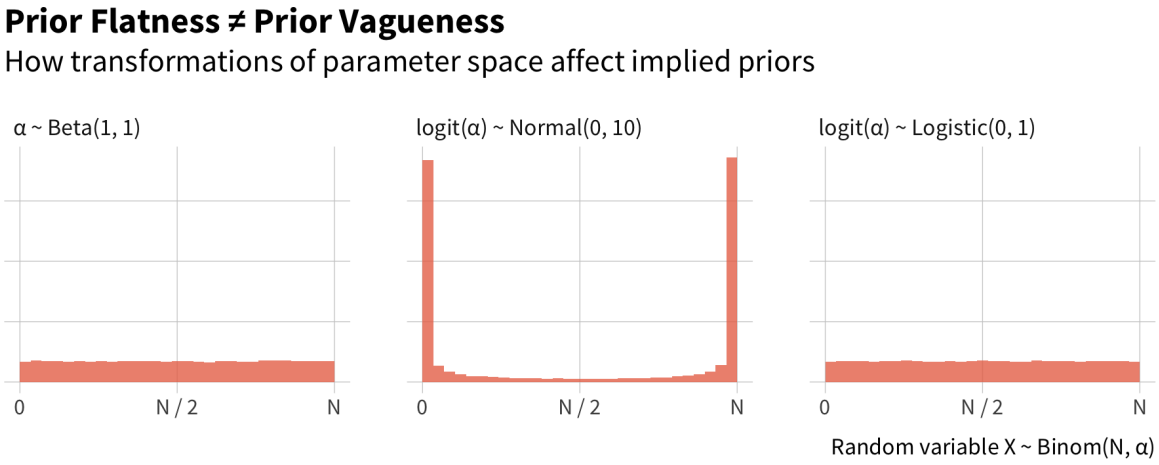


Figure 3.2: How parameterization affects priors. Transforming a model likelihood requires a commensurate transformation in the prior in order to produce the same model.

What general lessons can we draw from these exercises? It is a mistake to assume that the *shape* of a prior represents its informativeness. The relationship between shape and informativeness is contingent on the functions that map priors over parameters to implied priors over data.[11]

[10] The standard Logistic prior creates a flat density on the probability scale because the logit model uses the cumulative Logistic distribution function to convert logit values to probabilities. This same intuition holds for a probit model: a Normal $(0, 1)$ prior on the probit scale represents a flat prior on the probability scale.

[11] In other words, flat priors are not "invariant to reparameterization" of the likelihood. Understanding invariant priors is an animating motive for so-called "objective Bayesian" methods. Objective Bayes is the domain of "Jeffreys priors" and "reference priors." More in Section 3.3.4.

These mismatches between prior shape and prior information can be exaggerated by implied priors created from nonlinear functions, which compress and stretch probability mass through non-constant transformations. In other words, the model matters. This is a general principal of Bayesian model-building that essential for understanding Bayesian causal inference: "the prior can often only be understood in the context of the likelihood" (Gelman, Simpson, and Betancourt 2017). We should be prepared to encounter models where flat priors for parameters to yield data with highly informative prior distributions.[12] We should also be prepared to encounter models where non-informative priors over data are achieved using non-flat priors for parameters. As it relates to causal inference, this means that the model that "lets the data speak" the most may not be the model with the flattest priors.

3.3.3 Priors and model parameterization

Researchers have prior information about the *world*, but they must specify prior distributions in a model. The parameter space of the model may not obviously represent the natural scale of prior information, making it challenging to specify priors. We have also seen that transforming a parameter space can change the shape of a prior. The contingency of priors on parameter spaces is initially inconvenient, but researchers can it to their advantage when building a model for causal effects. If a prior is challenging to specify in one parameter space, the researcher can (and should) reparameterize the problem in order to specify priors in a more convenient parameter space. This section briefly discusses three points: what is parameterization, how does parameterization affect prior distributions, and how can the researcher use parameterization to their advantage.

All Bayesian models feature a model of data, $p\left(\mathbf{y} \mid \boldsymbol{\pi}\right)$. Like all models, the data model could be written in multiple equivalent ways according to different parameterizations. For ex-

[12] This has affected Bayesian causal inference in political science already: for instance, Horiuchi, Imai, and Taniguchi (2007) model treatment propensity with a probit model where coefficients are given Normal priors with variance of 100.

ample, the Binomial likelihood model above could be parameterized in terms of the probability parameter α or the log-odds parameter η. Because these are equivalent parameterizations, then for any $\eta = \text{logit}(\alpha)$, the likelihood of a data point under both parameterizations will be equal. These reparameterizations are ubiquitous in statistics and computing, and researchers often leverage them to expedite analytic or computational tasks.[13]

Parameterization is important for Bayesian statistics because it affects the parameter space. Parameters that are difficult to understand in one parameterization may be easier to understand in an equivalent parameterization. These parameterizations are not only opportunities for researchers to understand their models better, but they can reveal the consequences of certain prior choices, helping the researcher select priors that better represent the desired about of prior information.

Consider a randomized experiment with a binary outcome measure y_i and a binary treatment assignment z_i. Suppose that the causal estimand of interest is a difference in means, $\bar{y}_{z=1} - \bar{y}_{z=0}$, which is commonly estimated using a linear probability model. We can parameterize the model in two equivalent ways. First is a conventional regression setup,

$$y_i = \alpha + \beta z_i + \varepsilon_i \tag{3.13}$$

where α is the control group mean, $\alpha + \beta$ is the treatment group mean, β represents the difference in means, and ε_i is an error term for unit i. I refer to this parameterization as the "treatment-effect" parameterization, because it contains the effect parameter β. This parameterization is common even for estimation in experimental settings because it can be estimated using standard regression software. This parameterization is unappealing for Bayesian causal inference because it presents a challenge for prior specification. If a researcher simply places a flat prior on the treatment effect, the implied prior for the treatment group

[13] A fun example: OLS regression typically do not calculate $\hat{\beta} = X^\top X^{-1} X^\top Y$ directly. Instead, they calculate a reparameterization of the same problem (typically the "QR parameterization") that returns equivalent results but is easier to implement in the computer.

mean will differ from the prior for the control group mean. If a researcher wanted to give equivalent priors to the mean in each group, they must construct a non-flat prior for the treatment effect that accomplishes that goal

The researcher can instead use parameterization to their advantage, setting up a model that estimates the mean for each experimental group directly. Letting the mean in group z be μ_z, this parameterization would be

$$y_i = \mu_{z[i]} + \varepsilon_i \tag{3.14}$$

where the difference in means is calculated as $\mu_1 - \mu_0$. I refer to this parameterization as the "difference-in-means" parameterization. The difference-in-means parameterizations is equivalent in the likelihood to the treatment-effect parameterization, but it is much simpler to place equivalent priors on each group mean when the model is parameterized directly in terms of the means.

This example is also enlightening because it highlights an instance where flat priors have unanticipated consequences. Suppose that we use the difference-in-means parameterization and specify flat priors on each group mean: $\mu_z \sim \text{Uniform}(0, 1)$. The left panel of Figure 3.3 shows a histogram of prior simulations for the group mean, which is simply flat on the $(0, 1)$ interval. If we specify flat priors on both group means, what is the implied prior for the difference in means? The right panel of Figure 3.3 shows that the difference in means will actually have a triangular prior distribution with a mode at 0. The prior is still non-informative—after all, it results from flat priors about each group—but just because the prior is non-informative does not mean that it will always be flat.

This example is important because it highlights how a researcher's default tendencies—the treatment-effect parameterization and an impulse toward flat priors—can be incompatible when we actually examine the consequences of these modeling choices. A flat prior on the treatment effect creates non-equivalent priors for the group means, but equivalent (flat)

priors on the group means create a non-flat prior on the treatment effect. When a researcher confronts a modeling scenario, it is not enough to simply assume that a flat prior will return an optimal "data-driven" result, because the actual informativeness of a flat prior depends on the parameterization of the model and the functions being calculated with the model parameters.

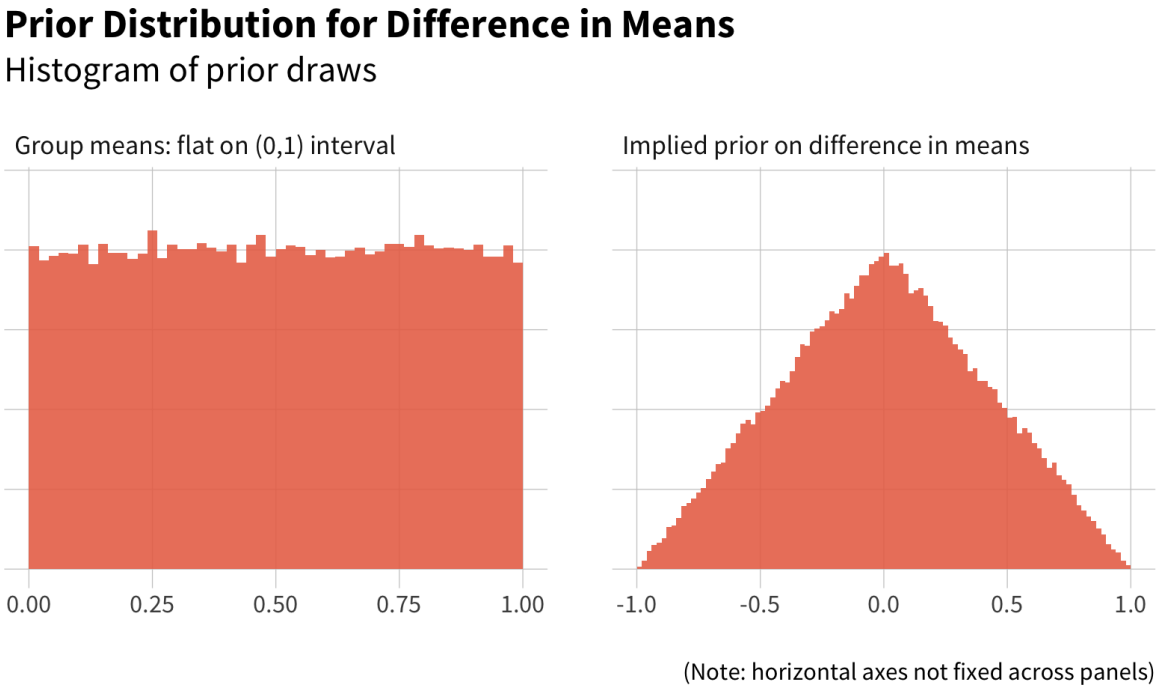


Figure 3.3: Model parameterization affects prior distributions for quantities of interest that are functions of multiple parameters or transformed parameters. The left panel shows that the difference between two means does not have a flat prior if the two means are given flat priors. Note that the x-axes are not fixed across panels.

3.3.4 Principled and pragmatic approaches to prior specification

How do we construct sensible priors if prior flatness is not always sensible and sensible priors are not always flat? This section offers productive guidance for specifying priors and discusses the appropriateness of different prior strategies for causal inference applications. I emphasize

the use of "weakly informative" priors and discuss some heuristic rules that can guide prior choices in many scenarios.

A Bayesian approach to causal inference does not mean using magical priors to somehow de-confound a treatment variable. Causal identification is a matter of research design, and simply asserting a prior belief that the treatment assignment is ignorable would be a misunderstanding of the role of priors. Priors provide structure for a model to learn from data, so data are still of paramount importance.

It is helpful to imagine different approaches to prior specification as lying on a spectrum from least informative to most informative. The least informative priors might be regarded as "nuisance" priors that are specified for no other purpose than to express uncertainty with a posterior distribution (Gelman, Simpson, and Betancourt 2017). We have already discussed how prior flatness is a misleading heuristic for choosing a non-informative prior. Statisticians in the "objective Bayes" tradition have developed several rule-based strategies for specifying non-informative priors without researcher intervention (Kass and Wasserman 1996), most notably Jeffreys priors and reference priors. The rules that determine these priors have complex information-theoretic justifications— Jeffreys priors (Jeffreys 1946, 1998) are defined in relation to the Fisher information matrix with the goal of extracting the most unobstructed information from the data as possible; reference priors (Bernardo 1979) maximize the KL divergence of the posterior distribution relative to the prior, i.e. the prior that maximizes the "amount learned" from data. An important property of this approach to prior distributions is that they are invariant to the model parameterization—if model 1 and model 2 are equivalent reparameterizations of one another, then the objective prior for model 1 yields the same posterior distribution as the objective prior for model 2. Objective priors are principled and general representations of prior ignorance, which make them superior to flat priors as representations of prior ignorance. Objective priors may still be undesirable because, like flat priors, they may still misrepresent the amount of information that the researcher has about

model parameters.

On the other end of the spectrum are informative priors. These priors are more often used to represent substantive beliefs or specific information about model parameters. Fully informative priors concentrate probability mass in narrower regions of probability space, excluding other regions that might be possible. These priors are commonly used for regularizing estimates and stabilizing weakly- or non-identified quantities, which are more common in measurement or predictive modeling than in causal modeling. Fully informative priors may be undesirable for causal inference because the bias–variance trade-off may be too great, a situation that Hahn et al. (2018) describe as "regularization-induced confounding." Regularization-induced confounding occurs when confounding effects, such as regression adjustments, are under-adjusted due to regularization, which can severely bias the treatment effect estimate even if all confounders are observed.[14]

In between non-informative and fully-informative priors is a region where Bayesian methods to causal inference will be both sensible and efficacious. Bayesian researchers refer to this neighborhood as "weakly informative priors" (**???**; Gelman, Simpson, and Betancourt 2017) Weakly informative priors provide some regularization of parameter estimates but still allow the model to be surprised by data. Weak information is more commonly thought of as "downweighting" unreasonable parameter values rather than "upweighting" the researcher's subjective beliefs. Weak information can take many forms, but I highlight four sources of weak information that will always be available to the researcher: structural information about data and parameters, the likelihood model itself, the number of predictor variables, and the tail behavior of a log prior density.

[14] In high-dimensional causal inference problems, regularization is necessary to estimate sparse signals and prevent overfitting (Samii, Paler, and Daly 2016). In these settings, regularization-induced confounding can be ameliorated by modeling the treatment assignment separately and estimating the treatment effect by residualizing the observed treatment. This routine, sometimes called "Neyman orthogonalization," facilitates unbiased causal inference in the presence of strong, high-dimensional confounding (Chernozhukov et al. 2018; Hahn et al. 2018; Hahn, Murray, and Carvalho 2020; Ratkovic 2019).

One source of weak information is "structural" information about the mathematical properties of model constructs (Gelman, Simpson, and Betancourt 2017). For example, probabilities are bounded in the $[0, 1]$ interval, correlations in $[-1, 1]$, and variances in $[0, \infty)$. These constraints sound trivial, but they are often consequential. Linear probability models (LPMS), for example, are often preferred over over logistic models in experimental settings because of their similarity to a difference-in-means T-test, but LPM estimates can sometimes escape the $(0, 1)$ interval. Even if a point estimate doesn't escape its structural bounds, structural priors can improve the posterior precision of an estimate by removing invalid parameter values from the prior. I present such an example below in Section 3.4.1.

A related source of weak information is the likelihood model itself. If we know the scale of the outcome data (and we ordinarily do), the likelihood provides prior information by defining the data's functional dependence on parameters. That was an important feature of the prior for the IRT model in Chapter 2: knowing that the probit model maps quantile values to probabilities using the Normal CDF provides a lot of information about which quantile values are plausible to obtain in the model. This principle generalizes to other models with other link functions, including linear models with an identity link.

A third source of weak information is the number of predictors in a model. Bayesian statisticians generally give regression coefficients tighter priors as the number of coefficients increases (Simpson et al. 2017). To understand the intuition for this decision, recall that the mathematical structure of a regression is a weighted sum of the predictors. As the number of predictors increases, the weight on each predictor ought to decrease on average. If a model includes additional predictors without adjusting the prior, the regression function's prior distribution will grow in variance because each coefficient adds another random variable to the regression function.

One final tactic for specifying weak priors is to have foreknowledge of the tail behavior of certain families of prior distributions. All priors will regularize estimates, but some priors

regularize more aggressively than others even if they look similar. A prior's tail behavior is especially important for circumstances when the data contradicts the prior. Prior distributions with flat tails will not regularize estimates as strongly, allowing the data to overcome the prior more easily. Other priors have thinner tails that decay more rapidly, which regularizes estimates much more strongly in the tails. The tail behavior of a selection of prior distributions is highlighted in Figure 3.4. The figure compares the density and log density of Normal, T (5 degrees of freedom), and Cauchy distributions, along with a Laplace distribution that serves as a conceptual stand-in for a sparsity-inducing prior. The log scale is helpful visualizing the impact of a prior's "penalty" on an estimate, where lower values of log density indicate a greater prior penalty on that parameter value. The log scale representation of the prior highlights how prior densities that look similar to one another—such as the Normal, T5, and Cauchy—can differ dramatically in their practical performance. Normal distributions have a quadratic log density, which is a Bayesian analog to quadratic or "L2" ridge penalty,[15] which is a much stronger penalty compared to the gentler T5 and Cauchy priors. The Laplace distribution regularizes more strongly than the other priors near zero, indicated by the fact that the Laplace log density does not gently approach its maximum. Despite its aggressive behavior near zero, the Laplace prior does not regularize as strongly as the Normal prior in its tails, which is why the Laplace prior is often regarded as a prior for sparse regression.[16] Comparing prior *log* densities in this way is generally helpful for deciding between prior families based on their regularizing behaviors.

If a researcher is ever in doubt about the consequences of their prior choices, a principled Bayesian workflow contains several important model-checking tools (Betancourt 2018; Gabry

[15] Indeed, the maximum *a posteriori* estimate from a Normal prior is equivalent to the maximum likelihood estimate using an L2 norm penalty Bishop (2006).

[16] Like the analogy between the Normal prior and L2 regularization, the MAP estimate under a Laplace prior is equivalent to the maximum likelihood estimate using L1 (absolutely value norm) regularization (Bishop 2006; Park and Casella 2008). More work on sparsity-promoting priors finds that horseshoe priors have more flexibility to control sparsity near zero and non-regularized signal detection farther from zero (Carvalho, Polson, and Scott 2010; Piironen and Vehtari 2017a, 2017b).

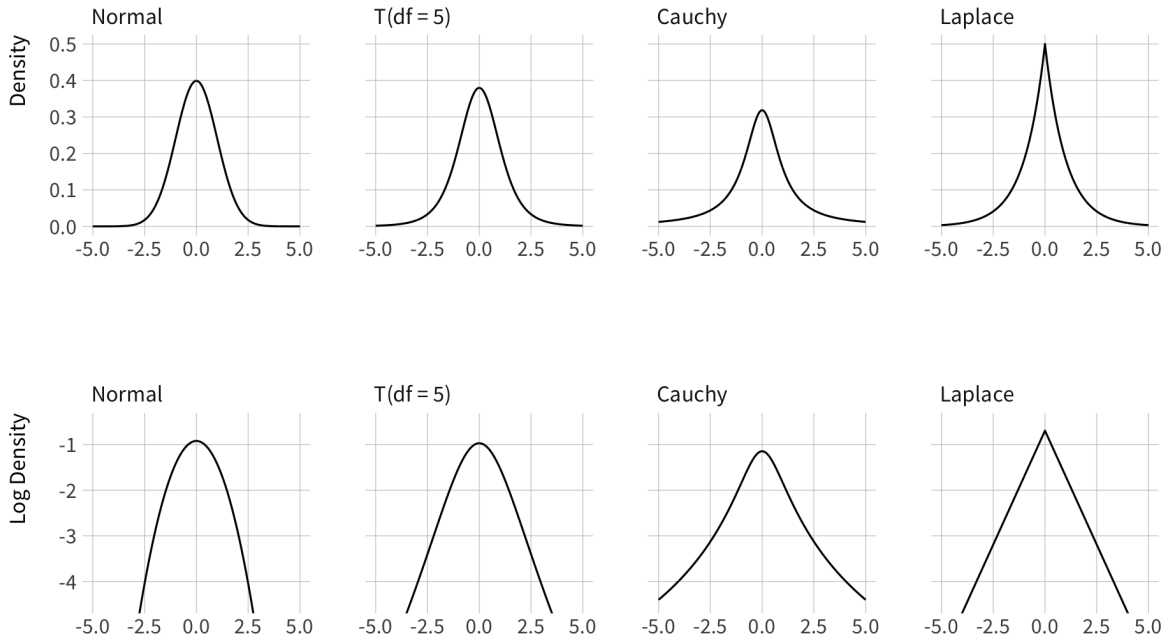


Figure 3.4: Density and log density for common mean-zero distributions. Log densities highlight the differences between families of prior distributions by emphasizing the tail properties.

et al. 2019). Most important are prior predictive simulation and model-checking with fake data. Prior predictive simulation (sometimes called prior predictive "checking" or prior "pushforward" simulation) consists of drawing a sample of parameters from the joint prior and using those parameters to simulate data. The result is a prior predictive distribution for the data, $\int p\left(\mathbf{y} \mid \boldsymbol{p}\right) p\left(\boldsymbol{\pi}\right) \, \mathrm{d}\boldsymbol{\pi}$. The distribution of simulated data should be only weakly informative—the draws are concentrated near the region of possible outcome values, but the distribution should be more broader than the marginal distribution of the observed data. Different features of a model can be stress-tested by estimating the model using simulated data. Fitting fake data is helpful for exposing and correcting undesirable features of a model while avoiding unnecessary "looks" at the data, which violate both Bayesian coherence and

frequentist p-values[17]

3.4 Bayesian Opportunities

Causal inference, like any statistical modeling scenarios, presents a number of problems that can be addressed with a Bayesian framework. Some of these opportunities are showcased in later chapters. Chapter 4 contains a multi-stage causal model and measurement error in the treatment of interest, two "problems" with uncertainty quantification that are handled naturally by the joint posterior distribution. Chapter 5 models heterogeneous treatment effects with continuous interactions, highlighting how informative priors can be used to regularize a flexible, highly-parameterized regression function.

There are many more scenarios where Bayesian tools are useful for causal modeling. I explore two such scenarios by replicating the analyses from two recent political science publications: an observational regression discontinuity study by Hall (2015), and a meta-analysis of field experiments by Green et al. (2016). The Hall (2015) study shows how a research finding how the precision and credibility of a causal study can be strengthened by incorporating weak prior information. With the Green et al. (2016) study, I translate key modeling assumptions into prior distributions and showcase how the conclusions from the study strongly depend on the choice of prior.

3.4.1 Application: weakly informed regression discontinuity

This section presents a Bayesian reanalysis of Hall's (2015) regression discontinuity study of congressional elections. Portions of the original analysis contain a pathological result where confidence intervals for key parameters of interest contain values that could not possibly occur. I overcome the pathology using weakly informative priors that contain structural

[17] Although see Devezer et al. (2020) on the subject of using the data twice.

information about the dependent variable only, excluding impossible parameter values while remaining uninformative over the possible values. This minor prior intervention successfully guides the posterior distribution away from impossible regions of parameter space, resulting in a posterior distribution that is consistent with the data as well as external structural information about the problem. This intervention does not undermine the main takeaway from the original study, but the Bayesian estimates for the effect of interest are notably smaller and more precisely estimated.

With similar aims to this project, Hall (2015) examines primary elections and their impact on ideological representation in Congress. The study asks if extremist candidates for Congress are more likely or less likely to win the general election contest than candidates who are relatively moderate by comparison. To identify the effect of candidate ideology in the general election, Hall (2015) leverages the vote margin in the primary election as a forcing variable in a regression discontinuity design (RDD). In a primary contest between an extremist and a moderate, the extremist advances to the general election if their vote share in the primary is greater than the moderate's, i.e. the extremist's *margin* (difference) over the moderate is greater than 0. If the extremist's primary margin is any less than 0, the moderate advances to the general election instead.

Hall (2015) applies the RD design by assuming that the effect of candidate ideology on win probability in the general election is locally identified at the threshold where the extremist's primary margin crosses 0. Hall estimates RD models using a few different specifications, but I replicate his simplest design, which is a linear probability model (LPM) of the following form.[18] The outcome y_{dpt} is a binary indicator that the general election candidate running in

[18] Data were obtained from Hall's replication materials, available on his website. http://www.andrewbenjaminhall.com/, last accessed July 02, 2020.

district d for party p in election year t wins the general election.

$$\begin{aligned} y_{dpt} = \beta_0 + \beta_1(\text{Extremist Wins Primary})_{dpt} + \beta_2(\text{Extremist Primary Margin})_{dpt} \\ + \beta_3(\text{Extremist Wins Primary} \times \text{Extremist Primary Margin})_{dpt} + \varepsilon_{dpt} \end{aligned} \tag{3.15}$$

where *Extremist Primary Margin* is running variable, the extremist's margin over the moderate candidate with the highest vote in the primary, and *Extremist Wins Primary* is the treatment variable, equaling 1 if the extremist's margin exceeds 0, plus error term ε_{dpt}. When the extremist margin exceeds 0, the general election candidate in case dpt is the extremist, otherwise the general election candidate is the moderate. The coefficient β_1 represents the intercept shift associated with the extremist primary win: the effect of candidate extremism at the discontinuity. I replicate this LPM using OLS as well as a Bayesian model with a slight reparameterization: instead of using a dummy variable and an interaction term, I subscript coefficients by w, which indexes the treatment status (*Extremist Wins Primary*),

$$\begin{aligned} y_{dpt} &= \alpha_{w[dpt]} + \beta_{w[dpt]}(\text{Extremist Primary Margin})_{dpt} + \varepsilon_{dpt} \\ \varepsilon_{dpt} &\sim \text{Normal}(0, \sigma) \end{aligned} \tag{3.16}$$

where α_w is an intercept for treatment status w, and β_w is the slope for treatment w. This parameterization implies two lines, one line for $w = 0$ and another line for $w = 1$. The treatment effect at the discontinuity is the difference between the intercepts, $\alpha_1 - \alpha_0$. This parameterization is helpful for extending the model below.

I plot the OLS win probability estimates near the discontinuity in Figure 3.5. At the discontinuity, we estimate that extremism decreases a candidate's win probability by 0.53 percentage points, which is the same effect found by Hall (2015). Visualizing the RD predictions reveal that the confidence set for model predictions at the threshold contain many values that cannot possibly occur. For moderate candidates, the confidence interval for the win probability at the discontinuity includes values as high as 1.24, which far exceeds the maximum possible value of 1.0.

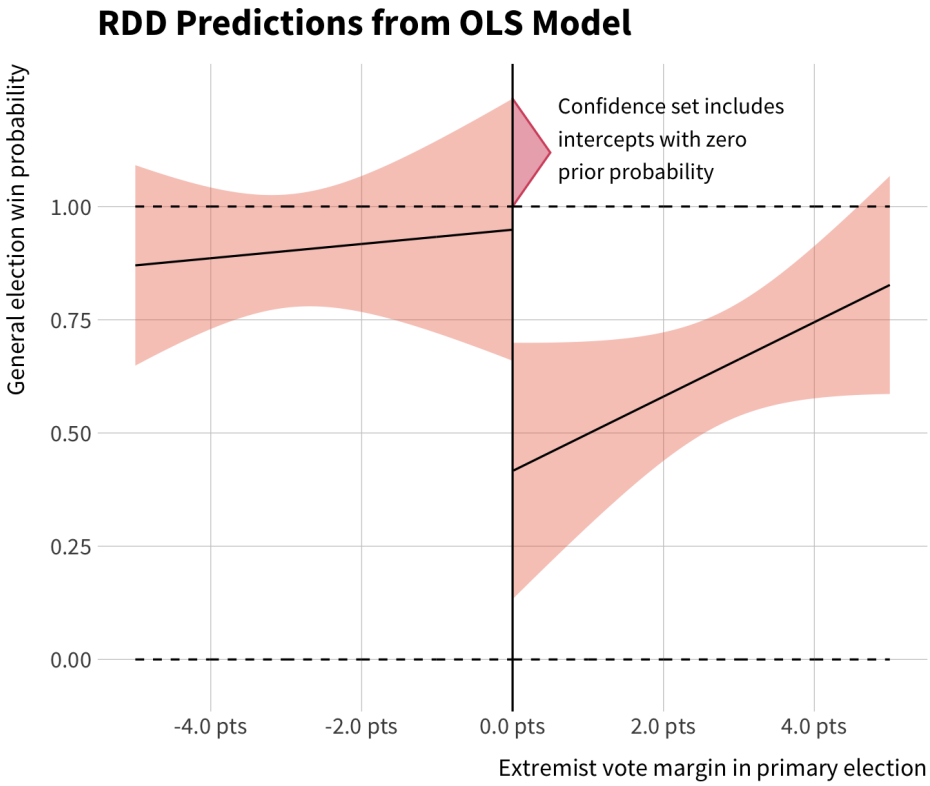


Figure 3.5: OLS estimates of the predicted probability that a candidate wins the general election. Local average treatment effect is estimated at the threshold. Treatment effect is defined by parameter estimates whose confidence sets contain values with no possibility of being true.

This pathology is possible in any LPM with finite data, but there are a few pragmatic reasons why we might not worry about it. First, for fully saturated model specifications, predicated probabilities from a model LPM are unbiased estimates of the true probabilities, and thus are an unbiased estimate of the treatment effect of interest. For frequentist inference, constructing a 95 percent confidence interval on this unbiased estimate might be enough to suit the researcher's needs. In this particular case, however, these reasons may not satisfy our goals because the model is not fully saturated. Instead, the design employs a local linear regression, so the extrapolation of the regression function to the threshold is model-dependent (Calonico, Cattaneo, and Titiunik 2014). It makes sense, then, to build a model that constrains

those extrapolations only to regions of parameter space that are mathematically possible for the problem at hand. Most importantly, because this intercept estimate is essential for calculating the treatment effect, the degree to which it is corrupted presents a significant problem for the inferences we can draw from the analysis.

To visualize just how much posterior probability this model places in impossible regions of parameter space, Figure 3.6 shows a histogram of posterior samples for the treatment effect from the Bayesian version of this model using (improper) flat priors on all parameters. Because flat priors do nothing to concentrate prior probability density away from pathological regions of parameter space, a large proportion of posterior samples contain intercept estimates that do not and cannot represent win probabilities. Of the 8,000 posterior samples considered by this model fit, 36% of the non-extremist intercepts are "impossible" to obtain because they are greater than 1 or less than 0. A small number of MCMC samples for the extremist intercepts take impossible values as well. As a result, just 64% of MCMC samples for the treatment effect is composed of parameters that are mathematically possible. Even invoking the practical benefits of the LPM, such a high level of corruption in the most important quantity of this analysis is cause to rethink the approach.

The Bayesian approach incorporates structural prior information about the intercepts estimated at the discontinuity. In particular, we specify a prior that these constants can only take values in the interval $[0, 1]$, which represents the substantive belief that the win probability at the threshold must not be less than 0% or greater than 100%. We remain agnostic as to which values within that interval are more plausible in the prior, which results in a uniform prior for the intercepts.

$$\alpha_{w=0}, \alpha_{w=1} \sim \text{Uniform}\,(0, 1) \tag{3.17}$$

The structural information in this prior is indisputable. We know with certainty that no probability can be less than 0 or greater than 1. Accordingly, this prior does nothing except

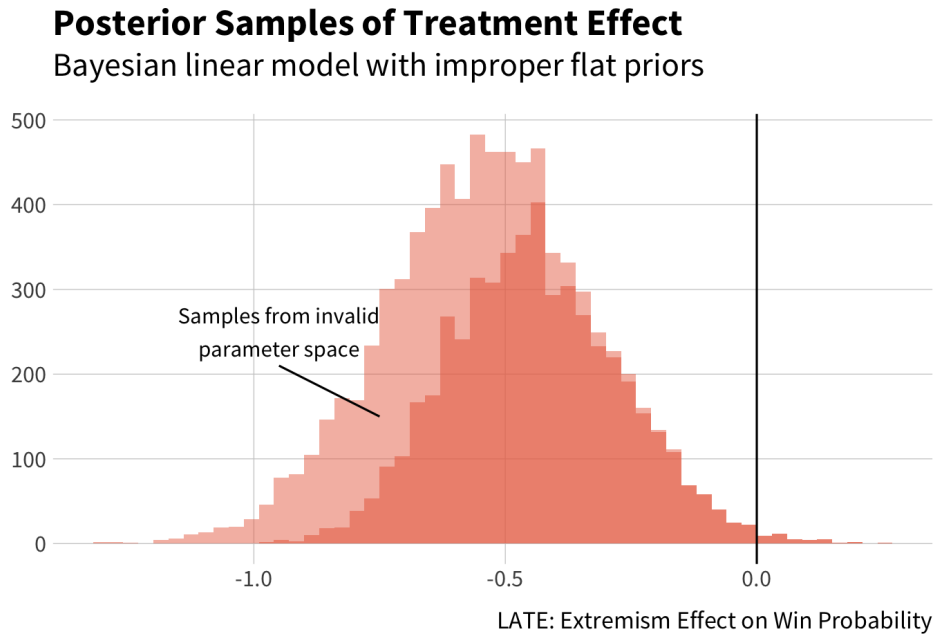


Figure 3.6: Histogram of posterior samples for the treatment effect. Using flat priors for each win probability intercept, a substantial share of the treatment distribution consists of parameters that cannot possibly occur.

exclude treatment effects that cannot logically be possible. Because we give flat priors to the individual intercepts rather than the treatment effect itself, the implied prior for the treatment effect inherits the triangular shape introduced above in Figure 3.3, which is vague being non-flat.

We complete the model by specifying distributions for the outcome data and the remaining parameters.

$$
\begin{aligned}
y_{dpt} &\sim \text{Normal}\left(\alpha_{w[dpt]} + \beta_{w[dpt]}(\text{Extremist Primary Margin})_{dpt}, \sigma\right) \\
\beta_{w} &\sim \text{Normal}\left(0, 10\right) \\
\sigma &\sim \text{Uniform}(0, 10)
\end{aligned}
\tag{3.18}
$$

The Normal model for the outcome data in the first line is equivalent to the Normal error term defined in (3.16). The priors for the β_w slopes and residual standard deviation σ are very diffuse given the scale of the outcome data, $\{0, 1\}$, and the running variable that only takes

values in the interval $[-5, 5]$, a bandwidth of ±5 percentage points around the threshold.

I also consider a logit model, pursuant to an argument that the problem with Hall's original analysis is simply that it employs an LPM. This setup considers the binary election result as a Bernoulli variable with a probability parameter specified by a logit model,

$$y_{dpt} \sim \text{Bernoulli}\left(\phi_{dpt}\right) \tag{3.19}$$

$$\text{logit}\left(\phi_{dpt}\right) = \zeta_{w[dpt]} + \omega_{w[dpt]}(\text{Extremist Primary Margin})_{dpt} \tag{3.20}$$

with logit-scale intercepts ζ_w and slopes ω_w.

Although this logit specification constrains all win probability estimates to fall in the appropriate region, specifying priors for logit models is more challenging because regression parameters are defined on the log-odds scale instead of the probability scale. Fortunately for the case of regression discontinuity, the treatment effect is defined at the threshold where the running variable is 0, so our prior for the treatment effect can be constructed in a region of parameter space where the running variable and its coefficients have dropped from the equation.

$$\begin{aligned} \text{logit}\left(\phi_{dpt}\right) &= \zeta_{w[dpt]}, \text{ at Extremist Primary Margin}_{dpt} = 0 \\ \text{which implies } \phi_{dpt} &= \text{logit}^{-1}\left(\zeta_{w[dpt]}\right) \end{aligned} \tag{3.21}$$

To construct a prior for the logit-scale intercepts that implies a flat win probability at the intercept, I use a standard Logistic prior on the intercepts,

$$\zeta_w \sim \text{Logistic}\left(0, 1\right) \tag{3.22}$$

which is the same transformation described in the discussion of Figure 3.2 above.[19] This prior resembles the same prior information as the structural prior in the Bayesian LPM above: I rule out impossible values while remaining agnostic about all valid win probabilities.

[19] It would also be possible to specify a flat prior on the probability and convert the parameter to the logit-scale, which would imply the same Logistic prior.

What effect do these minor prior interventions have? Figure 3.7 plots the results from these three Bayesian models: the problematic original model with improper flat priors, the Bayesian LPM with structural priors to constrain the intercepts, and the logit model that recreates the structural prior using the Logistic prior. The left panel shows a histogram of posterior MCMC samples for the extremist and non-extremist win probabilities at the threshold. The LPM at the top of the panel includes no parameter constraints whatsoever. As a result, we see the pathological behavior where the posterior distribution places positive density on win probabilities that we know with certainty to be impossible. The histograms in the second and third rows show the LPM and logit models with structural priors. Both models concentrate prior density on possible win probabilities only, resulting in posterior distributions that reflect prior information better than the unconstrained model. The posterior distributions are asymmetric and place a lot of posterior density at high win probabilities, but this should not alarm us. The asymmetry in the distribution reflects the signals obtained from the data, rationalized against weak information encoded in the prior. In other words, the asymmetry accurately captures our prior information about these win probabilities.

The right panel of Figure 3.7 shows how these parameter constrains ultimately manifest in our LATE estimates by plotting posterior means and 90 percent compatibility intervals for each model. As with most Bayesian models, our priors have the effect of shrinking important effects toward 0 while reducing their variances. In this particular case, the posterior mean for the local average treatment effect shrinks from -0.53 with flat/unconstrained priors to -0.44 using the LPM with constrained intercepts: a 17% reduction in the magnitude of the effect. The LATE from the Bayesian logit is -0.42, which is a 21% reduction in magnitude. This shrinkage comes from the fact that some of the largest treatment effects in our original posterior distribution were composed of impossible draws (evident in Figure 3.6 above). The standard deviation of the posterior samples is reduced for the models with structural prior constraints, so these prior interventions are also improving the precision of our estimates.

This is because a fair amount of posterior uncertainty in the unconstrained model was owed to impossible parameter values as well.

Results of Bayesian Regression Discontinuity
How weakly informative priors affect inferences

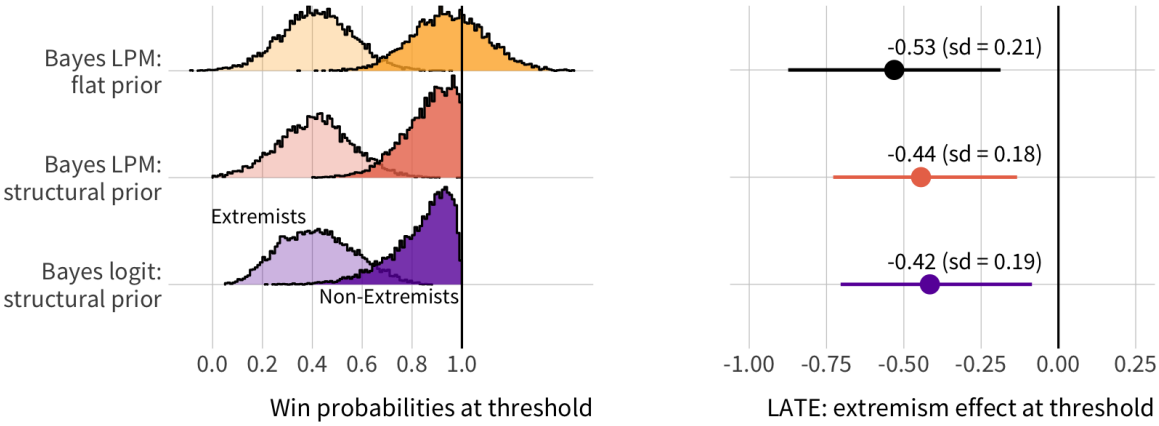


Figure 3.7: Comparison of posterior samples from Bayesian models with improper flat priors and weakly informative priors. Weakly informative models have flat priors over valid win probabilities at the discontinuity. Priors reduce effect magnitudes and variances by ruling out impossible win probabilities.

It bears emphasizing that the prior interventions in this case study were no more controversial than declaring what is already known: probabilities lie between 0 and 1. Since many causal research designs estimate treatment effects on binary variables, and many causal research designs are limited to small numbers of relevant real-world observations or budget-limited experimental samples, simple interventions like this have the potential to substantially improve the precision of research findings in contexts where researchers would otherwise leave out relevant prior information.

3.4.2 Application: modeling assumptions in meta-analysis

Green et al. (2016) present experimental results of the effects of yard signs on vote shares. The original study contains four separate field experiments with small samples and uncertain treatment effects. The research team conducted a "fixed-effects" meta-analysis to synthesize the results from all studies, which is equivalent to a prior distribution that the variance of treatment effects across studies is exactly zero. I relax the fixed-effects assumption using a Bayesian random effects model with different priors for the cross-study variance. My results show that the findings of original Green et al. (2016) are not robust across different priors. This exercise highlights how common modeling assumptions can be encoded as priors, how priors can be used to *relax* these assumptions rather than rigidify them, and how these priors are consequential enough to change our interpretations of experimental findings.

The Green et al. (2016) implemented their field experiment in coordination with campaigns and interest groups active during primary campaign cycles for county commissioner and mayor and general election cycles for congress and governor. These four races had varying salience levels, varying sign placements (yard versus road), and varying sign designs that differently emphasized the candidates' party, ideological affiliation, and messages about candidates' viability in the race. Different precincts within each electoral district were assigned to be treated with yard sign placements. Campaigns sometimes designated certain precincts as "must-treat" and "untreatable," so the original statistical analysis was a weighted least squares regression with inverse propensity weights to adjust for treatment assignment probability. Vote share data were measured using aggregate precinct-level returns, which was the lowest level of analysis for randomization and for outcome measurement. The team estimated all treatment effects with and without covariates, with covariates used to improve the precision of estimates and correct any residual covariate imbalance across treatment assignments. I take all propensity weighting and covariate adjustments conducted by the orig-

inal team as appropriate, reestimating the meta-analysis using covariate-adjusted treatment effect estimates (intent-to-treat) as presented in the published paper.

Figure 3.8 plots the original treatment effect estimates from each of the four studies, and the population treatment effect estimate from the original fixed-effects meta-analysis conducted by the authors. I use the standard errors reported in the paper to calculate confidence intervals as the point estimate ± two standard errors. The original estimates are noisy, owed to the low number of precinct observations in each study. The preponderance of evidence suggests, however, that the population treatment effect is more likely to be positive than negative, since the positive point estimates are more precisely measured than the negative point estimates. The pooled estimate is consistent with, indicating that yard signs increase a candidate's vote share by an estimated 1.7 points on average, with a standard error of 0.7 percentage points.

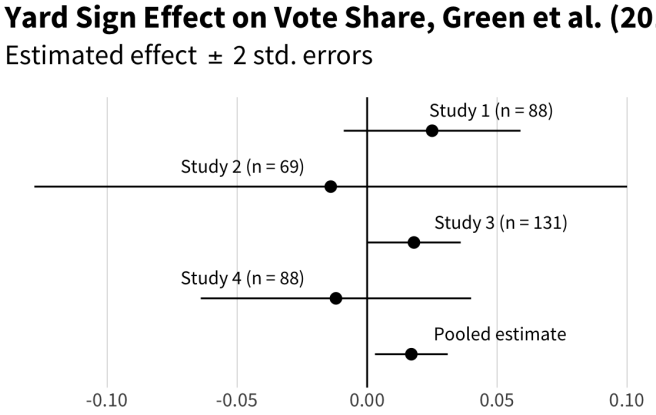


Figure 3.8: Original estimates from Green et al. (2016) field experiments. Pooled estimate from fixed-effects meta-analysis.

The meta-analysis finding suggests that yard signs do have an overall beneficial effect for candidates, even if the estimates from one-off studies yield uncertain results. The fixed-effects meta-analysis model, however, represents a special case where the "true" treatment effect in each experiment is assumed to be exactly the same. In other words, the model assumptions allow no heterogeneity whatsoever across study contexts to systematically modify

the effectiveness of yard signs. This assumption is unlikely to be true, since the settings cover both primary and general election campaigns, offices as high-profile as governor and as low-profile as county commissioner, electorates in different states, and more. The assumption of no study-level heterogeneity is relaxed in "random-effects" meta-analysis. The random-effects model corresponds to a view that "true" treatment effects will be slightly different from one another, depending on the context of a study. These contextual effects are captured by an additional variance parameter that describes the heterogeneity in treatment effects across studies as a separate source of variance aside from sampling error.

I now present a Bayesian random effects model that generalizes both the fixed- and random-effects, showing how the different model assumptions are represented by different prior distributions. This Bayesian approach is similar to the model proposed by Rubin (1981) to analyze parallel experiments in different schools (for a recent extension, see Meager 2019). Notationally, we have four experiments each indexed j with treatment effect estimates $\hat{y}_j$ and standard errors s_j. By the central limit theorem, each study estimate is modeled as a draw from a Normal distribution with a mean of τ_j, the true effect in study j, and standard error s_j.

$$\hat{y}_j \sim \text{Normal}\left(\tau_j, s_j\right) \tag{3.23}$$

The next step generalizes the random- and fixed-effects approaches to meta-analysis. The true effect for study j is modeled as a draw from the population effect, the average effect across all possible studies. The choice of the prior density depends on the assumptions that link a study to the population of possible studies. A common choice is a Normal distribution, which represents a generative assumption that individual study effects differ from the global effect by the addition of a large number of independent fluctuations—different studies have candidates, different campaign strategies, different voter populations, and so on, and the effects of each of these forces are additive shocks to a study treatment effect. This prior draws

the true study effect from a Normal distribution with mean μ and scale σ,

$$\tau_j \sim \text{Normal}\left(\mu, \sigma\right) \tag{3.24}$$

where μ represents the population mean and σ represents the amount of cross-study variation in treatment effects. The random-effects model represents a prior that σ takes some value greater than 0, leaving open the possibility that there is heterogeneity in treatment effects across studies. The fixed-effects model, meanwhile, restricts the value of σ to be exactly 0—*perfect prior knowledge* that all τ_j are unquestionably equal to μ, with no admissible possibility that the true effects contain study-level heterogeneity. This is a highly restrictive assumption to which the results of this meta-analysis are highly sensitive.

I complete the model by specifying priors for μ and σ. For μ, I follow Green et al. (2016) and estimate separate models that represent "agnostic," "skeptical," and "optimistic" prior beliefs about the population treatment effect, represented as Normal distributions with different mean and scale values.

$$\text{Agnostic:} \quad \mu \sim \text{Normal}(0, 0.05) \tag{3.25}$$

$$\text{Skeptical:} \quad \mu \sim \text{Normal}(0, 0.01) \tag{3.26}$$

$$\text{Optimistic:} \quad \mu \sim \text{Normal}(0.05, 0.05) \tag{3.27}$$

For σ, I estimate models with a range of priors that represent different prior expectations about cross-study heterogeneity. I use an Exponential prior with a rate parameter λ,

$$\sigma \sim \text{Exponential}(\lambda) \tag{3.28}$$

and set λ to different values in each model. The Exponential prior has an always-decreasing density, meaning that larger σ values (more heterogeneity) are always less likely in the prior. The expected value of the Exponential prior is $\sigma = \lambda^{-1}$, so we can directly interpret λ^{-1} as an expression of the expected prior heterogeneity across studies. The least informative prior

for σ sets λ equal to 20, or an expected cross-study variance of 5 percentage points. This is the same value as the prior scale for the agnostic and optimistic priors, representing a weakly informative assumption that the heterogeneity in the data is equal parts sampling error and study-level variation. I also estimate models with λ values equal to 100 and 200, which correspond to expected study heterogeneities of 0.5 and 1.0 percentage points, respectively. I also estimate a fixed-effects model where σ is forced to be zero, which is equivalent to an Exponential $(\lambda = \infty)$ prior.

I plot the population treatment effect from each model in Figure 3.9. The figure contains estimates using the optimistic, agnostic, and skeptical priors for the population treatment effect itself. The plot uses different point shapes and colors to display the estimates from different priors for the cross-study heterogeneity. Across all of the μ priors, the population treatment effect is largest and most precisely estimated with using the fixed-effects prior $(1/\lambda = 0)$. Indeed, the only σ prior that rejects the null hypothesis for the skeptical, agnostic, and optimistic prior is the fixed-effects prior of $\sigma = 0$. Any prior that allows the possibility of cross-study variance attenuates the treatment effect estimated toward zero and increases the posterior variance of the treatment effect. Many of these priors yield population estimates that fail to reject a null hypothesis.

It is important to note that the priors used in the random-effects models are not being "unfair to the data." There are many possible priors that could have been used that would have created more posterior uncertainty in the population treatment effects than the priors currently used. For instance, research commonly use Cauchy or Student's T priors to model "robust" random effects that regularize less aggressively. Those priors would have pooled study effects less than the Normal prior, resulting in more uncertainty about the population treatment effect. It is also important to note that the Exponential priors on the cross-study heterogeneity are not injecting "too much variance" into the model. A naïve approach to random-effects meta-analysis would have been to use a *flat* prior on the study variance, or

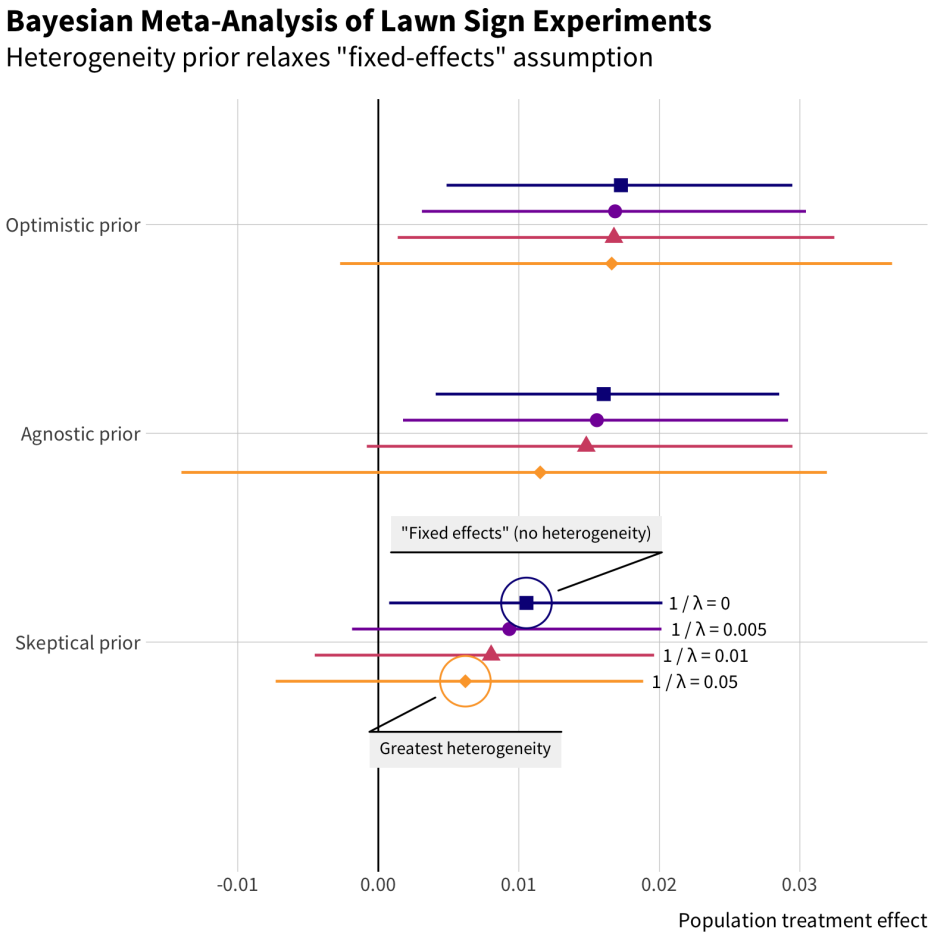


Figure 3.9: Population estimates from Bayesian meta-analysis models. Exponential priors with lower rate parameters represent the relaxation of the "fixed-effects" assumption.

Exponential(λ = 0), which would have resulted in much greater posterior uncertainty in the population treatment effect.

This exercise is useful for demonstrating how priors are consequential for the inferences drawn from causal studies. Sensitivity to priors is often regarded as a downside of the Bayesian approach, but this example is instructive because it highlights how priors don't disappear just because we wish to avoid the work of specifying them. The fixed-effects approach, which was employed as the "default" choice, did not avoid the issue of priors whatsoever. Instead, the fixed-effects model secretly posits a highly specific prior by assuming no cross-study

heterogeneity. The Bayesian model exposes this prior, allowing the researcher to scrutinize and improve their model.

3.5 Other Frontiers of Bayesian Causal Inference

There are many roads that this chapter does not take in its discussion of Bayesian causal modeling. In particular, this chapter is light on epistemology. While the Bayesian approach to learning and uncertainty is a valuable metaphor for scientific inference, especially as it relates to direct probabilistic inference about data that cannot be repeatedly sampled, it should not be taken too far. The hyper-Bayesian visions of entire scientific fields operating formally as gigantic Bayesian models is ridiculous, especially considering that all models are wrong and that priors almost never represent actual beliefs (Gelman and Shalizi 2013; Navarro 2019). Instead, I have tried to present Bayesian causal modeling as a collection of *pragmatic* tools to address *practical* problems with causal inference, not as theoretical tools for philosophical problems. As much as advocates of Bayesian inference like to discuss the "coherence" of the framework, there are many legitimate criticisms of its philosophical posture and operating characteristics (Mayo 2013, 2018) including from practitioners of the approach (Gelman and Yao 2020).

There are other approaches to Bayesian causal modeling in other fields that were beyond the scope of this chapter. Baldi and Shahbaba (2019) discuss "Bayesian Causality" with a broader vision where entire hypotheses are subjected to Bayesian updating from hypothesis tests. More practically, Hinne, Gerven, and Ambrogioni (2019) discuss a Bayesian approach to causal inference as a model-comparison problem. This framework considers two alternative models for data—a continuous model that assumes no causal effects, and a discontinuous model that separately models treated and control outcomes—and compares the evidence for each model using Bayes factors. This framework is interesting because it can quantify

evidence in favor of both competing models instead of only evidence *against* a null model (Wagenmakers 2007), but it is a strong methodological departure from traditional statistical practices in political science. Finally, Lattimore and Rohde (2019) argue that the *do*-calculus machinery behind Pearl (2009) can be entirely subsumed into Bayesian inference using probabilistic graphical models, which is interesting but not of immediate concern to political science.

— 4 —

How District-Party Ideology Affects Primary Candidate Positioning: A Bayesian De-Mediation Model

Do primary elections effectively transmit citizens' policy preferences into government? For this to be true, we should expect that the policy ideology with a partisan constituency to affect the ideological positioning of candidates who run for that party's nomination. This chapter explores the effect of district-party public ideology on the positioning of primary candidates running in that district.

It is important to distinguish the influence of the district-party public from the influence of the district overall. Does a candidate like Senator Susan Collins have a reputation as a moderate Republican because of a close balance between the number of Republican and Democratic voters in Maine? Or is the Republican constituency in Maine relatively moderate compared to Republican constituencies in states that elect more conservative Republicans? Although past research has been interested in the threat of primary challenges as a cause of ideological divergence between partisan legislators (for example Boatright 2013; Hill 2015; Hirano et al. 2010; McGhee et al. 2014), many of these studies lacked the capability to observe the preferences within local partisan groups as a concept distinct from aggregate partisanship

or aggregate voting in the entire district. This chapter uses my new measures of district-party ideology to investigate this question in ways that previous research projects could not.

The effect of district-party ideology on candidate positioning is a challenging causal inference problem. We cannot directly compare the "explanatory power" of district-party ideology and district-level voting by measuring whether one is more strongly correlated with candidate ideal points than the other, nor can we simply control for aggregate voting to recover the "partial effect" of district-party ideology. This is because aggregate policy ideology and aggregate voting are causally related: if a district contains a voter base with more conservative policy preferences, these policy preferences should influence aggregate voting behavior in the district as well as the positioning of candidates who try to respond to those policy preferences. Simply controlling for district voting in a regression will likely introduce collider bias by conditioning on a post-treatment variable (Greenland, Pearl, and Robins 1999; Montgomery, Nyhan, and Torres 2018).

This chapter advances this literature's use of moder causal inference methods by estimating the effect of district-party ideology on primary candidate positions using sequential *g*-estimation, a structural modeling approach that measures the direct effect of district-party ideology while fixing the mediating effect of district-level voting. Substantively, I translate the primary candidate's strategic positioning dilemma into the language of causal graphs, highlighting how aggregate districting voting mediates a relationship between district-party ideology and candidate positioning. Methodologically, I take the sequential *g* method as it appears in political science (Acharya, Blackwell, and Sen 2016) and embed it in a Bayesian framework. The Bayesian framework estimates all components of the structural model simultaneously, quantifying uncertainty in all model parameters in a single posterior distribution. This includes measurement uncertainty in ideal point estimates from the IRT model in Chapter 2, which is included as a prior distribution over ideal points. The key payoff for the Bayesian structure, therefore, is a unified framework for conducting inference about

treatment effects by marginalizing over other sources of uncertainty, including imprecise data and design parameters.

I find that primary candidates position themselves to fit district-party policy preferences: Republican candidates run more conservative campaigns in districts where the Republican constituency is more conservative, and Democrats run more progressive campaigns to please more progressive partisan constituencies. This finding holds even when controlling for aggregate district voting using sequential g-estimation. I also find, unlike other studies of primary representation, that primary candidates' responsiveness to district-party ideology is greater in closed primaries and weaker in open primaries.

4.1 Candidate Positioning and Voters' Policy Preferences

How do constituent preferences affect candidate positioning? This project explores the implications of what Brady, Han, and Pope (2007) call the "strategic positioning dilemma" (SPD), striking a balance between moderate position-taking to appease the general election constituency and ideological positioning taking to appease the partisan primary election constituency. Existing research contains plenty of studies that support the general theoretical intuition of the SPD theory, although I review some conflicts and ambiguities in detail in Chapter 1. To briefly review, general election candidates are rewarded at the ballot box for taking more moderate campaign stances (Canes-Wrone, Brady, and Cogan 2002; Hall 2015) and aligning themselves with local public opinion on specific issues (Canes-Wrone, Minozzi, and Reveley 2011; Fenno 1978; though see Fowler and Hall 2016). Nevertheless, no candidate makes it to the general election without first winning a primary nomination, where many scholars theorize that candidates benefit by taking more ideological positions that represent conventional views within the party. This could be a within-party Downsian incentive: the median primary voter is a ideological partisan with off-median policy preferences, so candi-

dates take more extreme positions to appeal to partisan constituency preferences (Aldrich 1983; Burden 2001). This is consistent with evidence from safe congressional districts, where candidates experience less general election threat and can more freely position their campaigns to target the primary electorate (Ansolabehere, Snyder, and Stewart 2001; Burden 2004). Pressures for primary candidates to take non-median stances may come through mechanisms unrelated to bottom-up voter pressures, instead reflecting candidates' need to organize committed staff and volunteers for their campaigns (Aldrich 2011; Layman et al. 2010; McClosky, Hoffmann, and O'Hara 1960), seek campaign funds from policy-seeking contributors (Barber 2016; Barber, Canes-Wrone, and Thrower 2016; La Raja and Schaffner 2015), or garner support from policy-demanding groups that control access to connections and resources to support candidates (Cohen et al. 2009; Koger, Masket, and Noel 2009; Masket 2009).

As I explained in more detail in Chapter 1, the explicit evidence about the SPD from primary elections themselves is surprisingly weak. This is mainly because most studies do not explicitly measure the ideological preferences of primary voters, instead using aggregate measures of voting that are not able to identify policy preferences (Kernell 2009). Furthermore, most aggregate-level measures of policy ideology do not differentiate between partisan constituencies in the same district (e.g. Tausanovitch and Warshaw 2013), which is important for understanding how candidates respond to their specific primary constituency. Without data that more closely resembles the theoretical story in question, studies that claim to demonstrate that candidates "handle the [strategic positioning] dilemma by positioning themselves closer to the primary electorate" rarely show anything of the sort (Brady, Han, and Pope 2007, 79). Brady, Han, and Pope (2007) show that among incumbent members of Congress, more liberal Democrats and more conservative Republicans attract fewer primary challenges and perform better in primary elections, but neither of these findings say anything specific lessons about how candidates are positioned relative to primary constituencies.

Hirano et al. (2010) get closer to this project's contribution by creating a statewide measure of primary electorate ideology from exit poll questions on ideological self-placement. They find that incumbent NOMINATE scores are more strongly related to general electorate ideological self-placement than primary electorate self-placement, a similar pattern as shown in Chapter 1, Figure 1.2. They further find no evidence that incumbent members of Congress do not have more extreme NOMINATE scores under greater threats of primary competition or higher primary turnout. This finding conflicts partially with Clinton (2006), who uses a large opinion survey of partisan voters in every congressional district to show that Republican members of Congress have NOMINATE scores that are more strongly related to the median Republican in their district, while Democrats' NOMINATE scores are more strongly related to the overall district median constituent. The IRT measures of district-party ideology that I create in Chapter 2 is a more direct measure of policy preferences than ideological self-placement scores or additive issue scores, which respectively tap a large amount of "symbolic" or identity-focused conceptions of political ideology or do not accord for differential measurement error across policy issues (Ellis and Stimson 2012; Treier and Hillygus 2009). My IRT ideal point scores, to my knowledge, are the first ideology scores for district-party groups to be applied to the study of congressional primaries.[1]

Research on primary competition and candidate positioning is also held back by the availability of primary candidate data, which until recently was not available. Past studies of primary competition and polarization typically use ideology scores from legislative roll call votes, which include only incumbent members of congress or state legislators (Brady, Han, and Pope 2007; Hirano et al. 2010; McGhee et al. 2014), or they use surveys of general election candidates, which may include non-incumbent candidates for the general election but no candidates who ran in the primary and lost (Ansolabehere, Snyder, and Stewart 2001;

[1]For IRT estimates of partisan constituencies and "ideological nationalization" at the U.S. state level, see Caughey, Dunham, and Warshaw (2018).

Burden 2004). Recent ideal point methods that use political financial contributions data have the capacity to scale a much broader universe of political actors, including candidates, political parties, PACS and interest groups, and donors (Bonica 2013, 2014; Hall and Snyder 2015). Few notable studies have yet used these contribution-based scores to study primary candidates (Ahler, Citrin, and Lenz 2016; Porter and Treul 2020; Rogowski and Langella 2015; Thomsen 2014, 2020).

Even with better measures of district-party ideology and candidate positions, there are some reasons to suspect that district-party ideology does not have a straightforward effect on candidate extremism. Thomsen (2014) finds that politicians are less likely to run for Congress (versus a state office) if their moderate stances make them a weaker "fit" for their party, measured as a greater distance between a candidate's CF score and the average CF score in their state-party. This process could weaken the relationship between district-party ideology and candidate positioning because moderate candidates select themselves out of the campaign, even if they might appeal to district partisan voters. Porter and Treul (2020) find that the number of primary candidates who lack prior elected experience is increasing over time. Interestingly, they find no pattern between candidate experience and candidate extremism. This non-relationship may indicate that the positioning of inexperienced candidates may be less related to district-party ideology, either because the candidates are not adept at perceiving local ideology, they do not have the resources to conduct or acquire surveys of their constituencies, or emphasize non-ideological appeals in their campaigns. Incumbents, in turn, may be more responsive to local ideology as a result of their political skills and survivorship bias—ill-fitting candidates don't become incumbents in the first place—although incumbents systematically misperceive public opinion as well (Broockman and Skovron 2018).

4.1.1 Primary rules

More recent studies of primary candidates tend not to focus on the overall relationship between voters' policy ideology and candidate positions. Instead, there is an enduring interest in the way primary rules affect candidates' positioning incentives by altering the composition of the primary electorate. Primary "rules" refer to regulations in state election law that control which voters can participate in primary elections and which inclusion criteria parties can define within those legal confines. A state primary is "closed" if only registered members of a political party are allowed to vote in the party's primary election. On the other side of the spectrum, an "open" primary allows any registered voter to cast a vote in any party's primary. There are many more states whose primary rules fall between these two extremes, such as "semi-closed" systems that allow party-unaffiliated voters to choose which party primary they want to vote for, even if registered partisans must vote in the primary of their party registration (McGhee et al. 2014). Political scientists and political observers speculate that these rules shape candidates' positioning incentives by changing the composition of primary voters. Closed primary elections, so the argument goes, are limited to registered party identifiers in a district, so candidates running in closed primaries must appeal to a more ideologically homogeneous primary electorate in order to win the nomination. More open systems, especially "blanket" systems where candidates from all parties run on the same primary ballot to advance to the general election, encourage primary candidates to take more moderate stances to attract a more ideologically diverse primary coalition. The general hypothesis, therefore, is that more restrictive primary rules exacerbate ideological polarization in congress and in state legislatures, and furthermore that polarization might be combated by moving primary elections to more open rules.

This "primary rules hypothesis" receives essentially zero support across several studies of U.S. elections. Hill (2015) studies the relationship between primary rules and the ideological

makeup of the primary electorate, asking if primary voters in states with more open primary rules are in fact more moderate on average than primary voters under closed primary rules. Estimating an IRT ideal point model on individual CCES respondents in each congressional district and validated voter turnout data, Hill finds that primary voters are more ideologically consistent than general election voters in the same party, but primary voters are no more extreme in closed primary states than in open primary states. Even if primary electorates are not affected by state primary rules, candidates may still suspect that primary rules matter and position themselves accordingly. Rogowski and Langella (2015) study the relationship between primary rules and candidate positioning as measured with CF scores. They find no systematic evidence that either congressional candidates or state legislative candidates are more extreme in closed primary states or more moderate in open primary states. McGhee et al. (2014) also study state legislators but using ideology scores that bridge the NOMINATE ideal point space to state legislative voting (Shor and McCarty 2011b), again finding no convincing evidence that primary systems matter for polarization in state legislatures. Within-state studies of changes to primary rules over time find mixed and highly qualified results, focused primarily on California's change to a "top two" primary system.[2] Bullock and Clinton (2011) find that the shift to the top-two primary promoted the election of more moderate candidates in California in competitive districts, measured using the two-party presidential vote, but no effects in more lopsided districts. Looking at the mechanisms underlying this, however, Hill (2015), who finds no effect on the ideological composition of primary voters after California's reforms, and survey experiments by Ahler, Citrin, and Lenz (2016) broadly show that voters are unable to identify which candidates are more ideological and which are more moderate.

I leverage my new data on district-party ideology to revisit the primary rules hypothesis in Section 4.3.1 below. Unlike most studies, I find that more candidates are less responsive to

[2] In a top-two system, candidates from all parties compete in a single primary for two spots on the general election ballot, which are awarded to the top two plurality winners in the primary.

district-party ideology in states with more open primary rules, consistent with the primary rules hypothesis.

4.1.2 Ideology within the two major parties

The SPD claims that candidates should be differentially responsive to primary and general electorates, but other theories on the ideological nature of U.S. parties may be relevant as well. An increasingly prominent theoretical perspective in U.S. political research holds that the parties are not asymmetrical but instead exhibit many "asymmetries" that help explain recent political conflict. In particular, the Republican Party is understood as an "ideological" party committed to a smaller welfare state, less regulation of business, and conservative cultural values, while the Democratic Party is a "group-based" party whose priorities reflect the mixture of social groups that compose the party's core constituency (Grossman and Hopkins 2016). The mixture of group interests within the Democratic Party leads to internal conflicts about which policies to prioritize, while the Republican Party is more concerned with who is a "real Republican" or a "real conservative" (Freeman 1986). The ideological consensus in Republican political thought provides constituents with many different values-based rationales for supporting conservative policies, while Democrats remain more conflicted about how to rationalize their desires for activist government policies against an individualistic American value system that downplays the significance of group identities (Feldman and Zaller 1992; Free and Cantril 1967; Lelkes and Sniderman 2016).

The ideological foundations of U.S. political parties could be relevant for the way candidates conceive of their "responsiveness" to constituents. Because the Republican Party has an ideological underpinning, and elite political actors will be more aware of partisan ideology than individual constituents will be, Republican candidates may not exhibit much ideological responsiveness even if their constituents' *policy attitudes* contain real variation. In other words, the ideological identity of the Republican Party could be a stronger organiz-

ing principle for Republican candidate positioning than the heterogeneous views of local constituencies. Democrats, meanwhile, may appear more responsive to district-party ideology because local opinion variation reflects the social group profile of the constituency, which is the organizing feature of Democratic Party representation. The intuition of these "asymmetric party" predictions diverge from (Clinton 2006), who finds that Republicans are *more* responsive to within-party opinion variation than Democrats but doesn't provide much theoretical exploration of why this should be the case.

4.1.3 Exploratory analysis

The analysis begins by examining the topline correlation between district-party ideology and candidate positions. For the dependent variable, I use the dynamic CF score included in the DIME congressional candidate database for 2012, 2014, and 2016 candidates (Bonica 2019b). For district-party ideology, I use the mean from the MCMC samples of the IRT model in Chapter 2, which are estimated from polling data over the 2010s districting cycle. Using only the mean understates the amount of uncertainty in the ensuing analysis, which is later corrected in the full analysis. For now, these initial investigations serve only to give us an impression of the raw data.

Figure 4.1 shows the topline relationship between primary candidate CF scores and the ideal point mean for the district-party they ran to represent. Each point represents a primary candidate for Congress in either the Democratic or Republican Party primary in years 2012, 2014, and 2016 as they appear in the DIME congressional database. This totals 1,975 Democratic candidates and 2,197 Republican candidates over three election cycles. In addition to each candidate, I plot least-squares regression lines calculated separately for each party. Confidence intervals reflect standard errors that are clustered at the district-party level to capture correlated error among candidates who run in the same primary race.

The figure shows a weak but decisively positive relationship between ideal point means

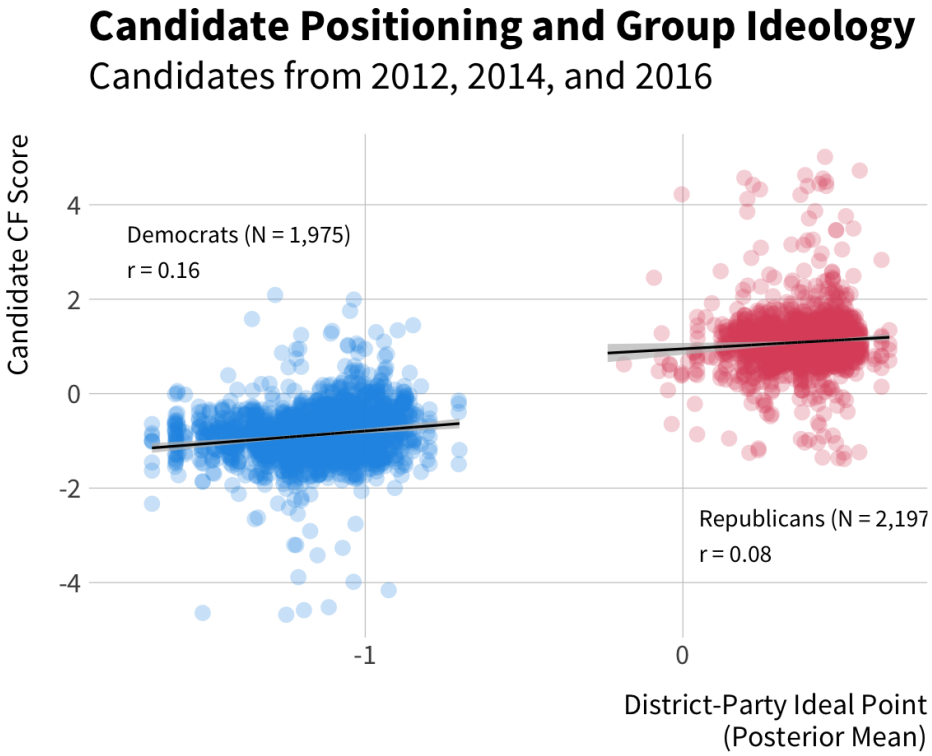


Figure 4.1: Topline relationship between district-party ideology and candidate positions. The horizontal axis plots the posterior mean for a group ideology, and the vertical axis is the dynamic CF score for primary candidates included in the DIME congressional candidate database.

and CF scores; district-parties that are more conservative see more conservative primary candidates. The relationship is stronger for Democrats than for Republicans in slope (0.53 versus 0.38) and in correlation (0.16 versus 0.08). Because one-unit increases are large on the ideal-point scale, it is helpful to standardize these coefficients in terms of standard deviations in the raw data. Conveniently for a bivariate regression, the standardized coefficient is equivalent to the correlation coefficient. Increasing district-party ideology by one within-party standard deviation is associated with a CF score increase of 0.16 standard deviations among Democrats ($p < .01$) and 0.08 standard deviations among Republicans ($p = 0.01$). A small number of outlier CF scores exist for each party, but these outliers are few compared to the approximately 2,000 observations in each party. The regression lines track the center of

each party's ideal point distribution, so these outliers do not appear to be large influences on the topline relationship.

Figure 4.2 plots the same relationship with candidates divided into incumbents, challengers of incumbents, and candidates running for a district with no incumbent running for reelection. It is immediately noticeable that CF scores have lower variance among incumbent candidates than among non-incumbent incumbents, and the correlations between CF scores and ideal point means are markedly higher: 0.39 among Democrats and 0.37. The overall higher correlation suggests incumbents could be more capable at positioning themselves for their primary electorates or that prior elections are effective at screening out ill-fitting candidates to represent the district. The higher correlation among Democrats also disappears among incumbent candidates, suggesting that no party is obviously more responsive to primary constituencies than the other, contrary to Clinton (2006)'s finding that Republican incumbents were uniquely sensitive to ideological variation within their partisan bases. CF scores are higher variance among challengers and open-seat candidates, and their relationship to district-party ideology is weaker but still positive. This may be because challengers and open seat candidates must seek other ways to appeal to candidates aside from ideological fit. This would be consistent with Boatright (2013)'s key finding that primary challenges focused purely on ideological fit to the district are relatively rare. A more mechanical explanation could be that CF scores are higher variance for challengers and open seat candidates because they struggle to raise from the same concentrated network of donors as incumbents do, attenuating the relationship between CF scores and district-party ideologies. Even still, the data generally suggest that candidates of all statuses have some awareness, on average, of how to position themselves as more conservative or progressive to corner their local partisan constituencies. Almost all relationships are statistical significant at a 1% level except among open seat candidates, whose sample sizes are also smaller.

Another notable finding among incumbents is the appearance of a much smaller "inter-

cept shift" between the two parties. While other studies typically find a large gap between Republicans and Democrats who represent otherwise equivalent districts [McCarty, Poole, and Rosenthal (2009); among others], the predicted CF scores for Democratic and Republican incumbents appear to diverge less dramatically if each regression line were extrapolated to meet at moderate values of district-party ideology. This interpretation comes with several caveats, naturally. First, there is no way to know if a linear extrapolation is an appropriate method for comparing parties in "otherwise equivalent districts," since there are no districts whose partisan constituencies are similar enough to make that extrapolation without strict functional form assumptions. Second, a cursory regression analysis of CF scores on district-party ideology and party still finds mean difference between the two parties, even though it is smaller among incumbents. Nevertheless, the data broadly reinforce the theoretical notion responsiveness to partisan constituencies partially explains at least some of the ideological distance between Republican and Democratic candidates running in the same district. Future research on inter-party divergence could incorporate district-party ideology scores and address this issue more directly.

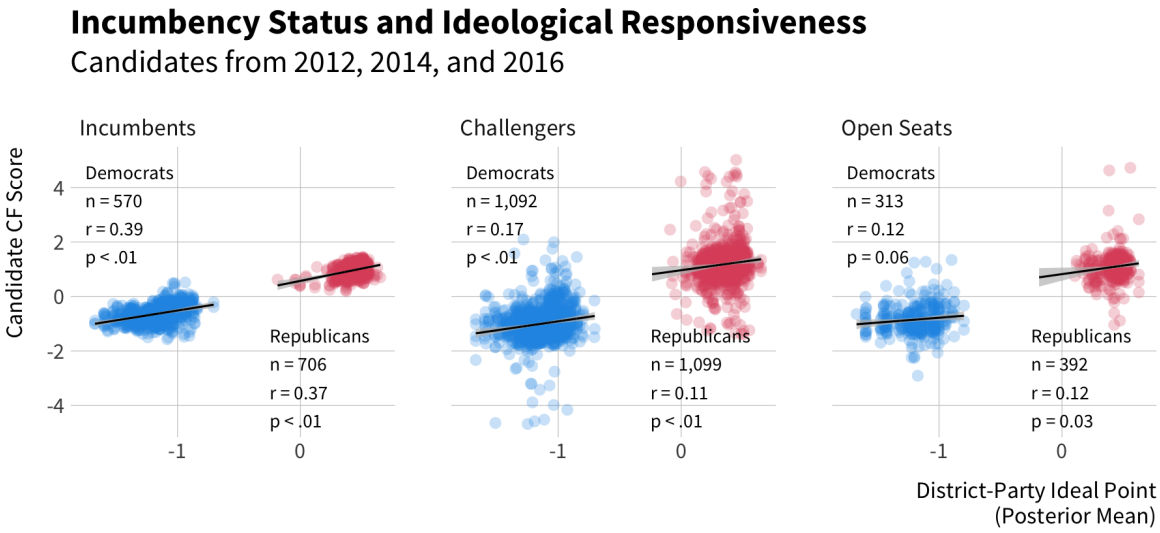


Figure 4.2: District-party ideology and candidate positioning across candidate incumbency status.

Figure 4.3 shows how the relationship between district-party ideology and CF scores varies by election year, splitting the 2012, 2014, and 2016 campaign cycles into three panels. There is no definitive trend toward increasing polarization or increasing responsiveness to partisan ideology in more recent years, which which is inconsistent with theoretical speculation that candidates have become more sensitive to primary electorates over the past few elections. Instead, the same weak but positive relationship appears in nearly all subsets of data with no overwhelming explanation for differences over time or between parties. The flattest relationship actually appears most recently for Republicans in 2016, among whom the relationship is nearly flat. We should hesitate to interpret too much from one estimate, but future researchers could investigate whether financial contributions by Republican donors had a different ideological character in 2016, if perhaps Donald Trump's unusual campaign platform altered which donors wanted to support which candidates, or if moderate donors directed their money away from Republican candidates in anticipation of a Democratic national victory.

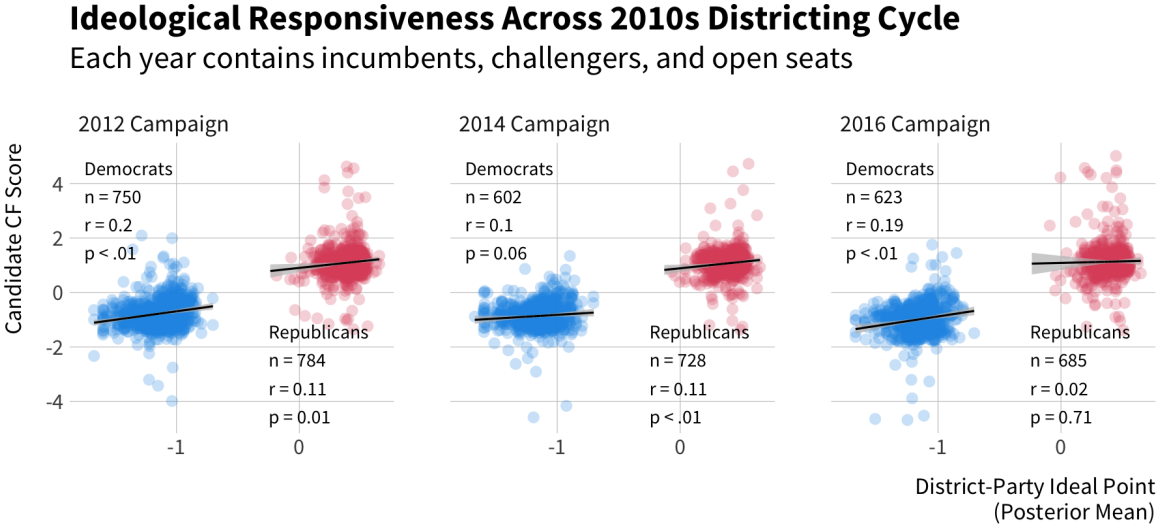


Figure 4.3: District-party ideology and candidate positioning across election cycles within the 2010s districting cycle.

Finally, Figures 4.4 and 4.5 plot the topline relationship between district-party ideology

and CF scores for candidates running in states with different levels of primary openness. Data on primary systems come from Boatright, Moscardelli, and Vickrey (2017) for 2012 and 2014, and I coded 2016 by consulting the National Conference of State Legislators, Ballotpedia, and OpenPrimaries.org.[3] I code primary rules using a three-category scheme. "Closed" primaries allow only pre-registered partisans to participate in a primary election. "Semi-closed" primaries are closed to party registrants, but they allow independent or non-partisan voters to choose which primary in which they wish to participate. Lastly, "open" primaries allow any eligible voter to participate in any party's primary. I code nonpartisan blanket and top-two primaries as "open." I choose a coarser three-level scheme over the five-level scheme in McGhee et al. (2014) because it is unlikely that voters process the fine legal differences that lead the authors to classify (for instance) "semi-closed" states and "semi-open" states differently. The three-part scheme is also more specific than the two-part open/closed scheme used by Hill (2015), since it is difficult to group independent and non-partisan participation in semi-closed states as either closed or open.

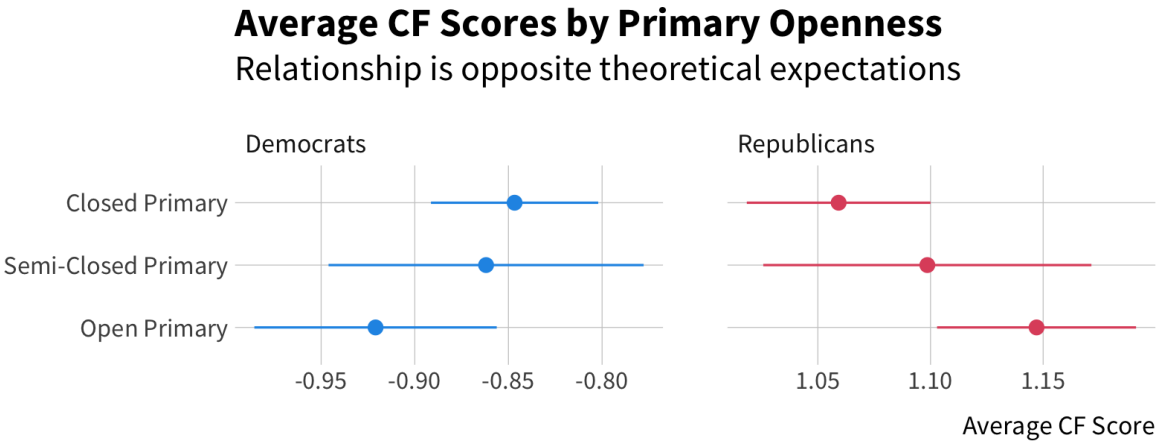


Figure 4.4: Average CF scores candidate positioning in states with closed, semi-closed, and open primary rules.

Conventional wisdom about primary openness implies that closed primaries should

[3] Accessed May 27, 2020.

produce more extreme candidates, and open primaries should produce more moderate candidates. Figure 4.5, which plots the average CF score for Republicans and Democrats under each primary system, contradicts this hypothesis.[4] Democrats are in fact more progressive and Republicans more conservative in states with increasingly open primary participation rules, which is the opposite direction of the commonly hypothesized pattern. Figure 4.5 goes on to plot the linear relationship between CF scores and district-party ideology in each state to see if candidates in closed primaries are more responsive to district-party ideology. The relationships do not support the primary rules hypoethesis either. While Republican candidates are indeed least sensitive to local partisan ideology in states with the model open rules—indeed the point estimate of the relationship is negative—they are *most* sensitive in semi-closed rather than closed states. The evidence from Democrats is even less consistent with the primary rules hypothesis. Democrats in semi-closed systems are also most responsive to district-party ideology, and they are less sensitive to district-party ideology in closed primary systems than they are in open systems.

The descriptive results from Figures 4.1 through 4.5 inform a few modeling choices for the causal analysis to follow. Because incumbency status appears to modify the relationship between citizen and candidate ideology, some of the analysis below estimates the effect of district-party ideology using subsets of data on incumbents, challengers, and open seat candidates. By comparison, estimates exhibit no clear time variation, so I choose to pool election cycles into one model, using fixed effects where appropriate to the design, rather than estimating entirely separate models for different cycles. And although the descriptive relationships were broadly similar for both parties—contradicting an "asymmetric parties" prediction that Republicans would be less responsive to district party ideology as well as the Clinton (2006) finding that Republicans are *more* responsive— the variables that could

[4] Estimates are calculated from a linear regression on indicator variables for each primary system type with group-clustered standard errors.

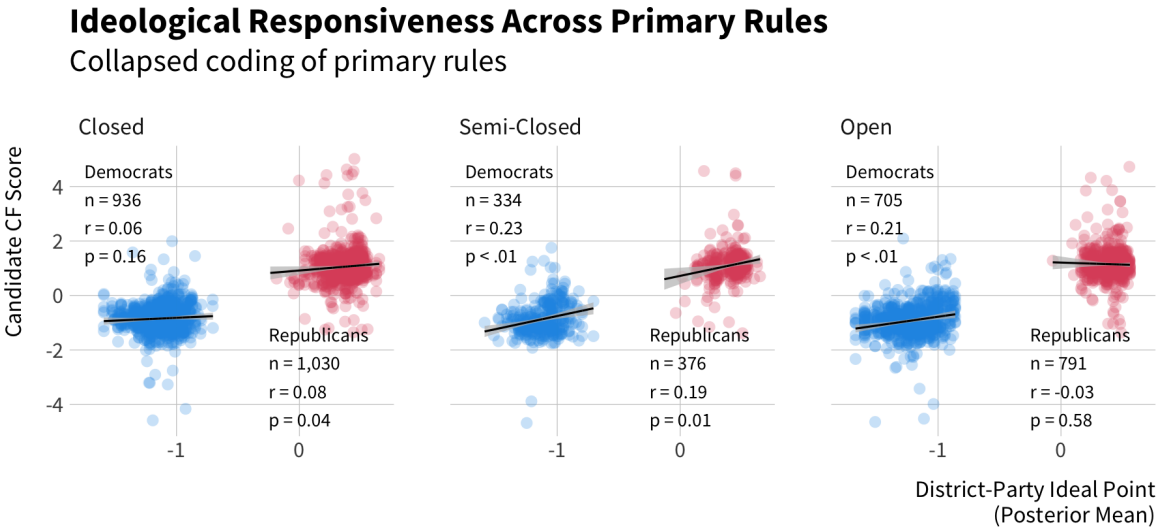


Figure 4.5: District-party ideology and candidate positioning in states with closed, semi-closed, and open primary rules.

confound the relationship between district-party ideology and candidate ideology are likely to differ dramatically across parties. In order to increase the credibility of the selection-on-observables regression design, I therefore estimate models for Democrats and Republicans separately.

4.2 The Causal Effect of District-Party Ideology

The descriptive picture in Figures 4.1 through 4.5 are suggestive about the relationship between district-party ideology and candidate ideology, but the correlational analysis is insufficient to identify the causal effect of the partisan constituency. As with many regression-based analyses using observational data, we are concerned about variables that confound the relationship between district-party ideology and candidate positioning—district characteristics that affect the degree of conservatism among both voters and candidates simultaneously. Even more troublesome than confounding is the causal structure linking theoretical components of the strategic positioning dilemma: how the ideological locations of the partisan constituency *and*

the total district constituency affect one another, and how they jointly influence a candidate's positioning calculus.

The strategic positioning dilemma (SPD) describes the campaign incentives that arise when a candidate chooses their campaign's ideological location for a multi-stage campaign season. Let CF_i represent the campaign position for candidate i in left-right ideological space. The theory states that candidates optimize their chances of winning by taking a position between the ideological location of their partisan constituency, denoted $\bar{\theta}_g$ for the district-party group g in which the candidate runs, and the ideological median among the district constituency, denoted V_d for the district d where group g resides.[5] Figure 4.6 plots a hypothetical candidate's location between the partisan and district constituencies. Whether the candidate positions themselves closer to the partisan constituency or to the district constituency is a function of the candidate's perceived degree of electoral threat from each constituency. In an electorally safe district, the candidate of the advantaged party is reasonably assured to win the general election, so they may take a campaign position closer to the partisan constituency in order to neutralize the threat from other partisan candidates in the primary election. In a competitive district, the candidate faces greater general election competition, so they position themselves closer to the district constituency to avoid alienating moderate voters who could decide the election (Aldrich 1983; Burden 2001). As it relates to causal inference, the theoretical setup implies a causal effect of the primary constituency on candidate positioning (a causal path $\bar{\theta}_g \rightarrow \text{CF}_i$) and a causal effect of the district constituency on candidate positioning ($V_d \rightarrow \text{CF}_i$).

This analysis proposes an even more specific causal structure, the details of which are crucial for the research design and statistical approach. I invoke a causal model where the party constituency location affects both the district constituency location and the candidate

[5] We distinguish individual candidates i from district-party groups g and districts d, because every district contains two major party groups, and every group can contain multiple primary candidates in a given election cycle.

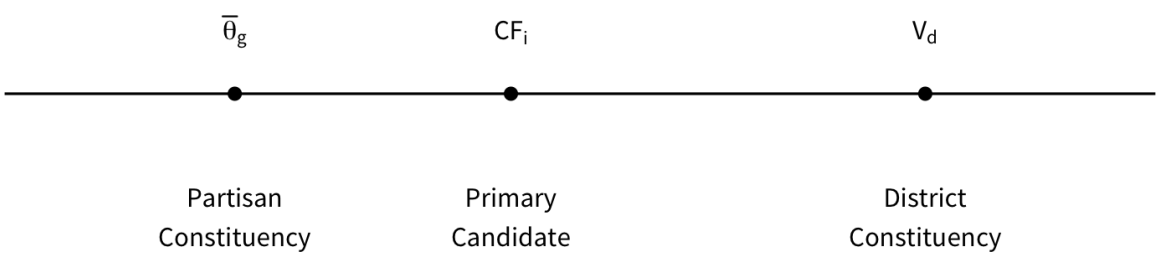


Figure 4.6: Spatial representation of key variables affecting the strategic positioning dilemma.

location, but the district location only affects the candidate location. The diagram in the left half of Figure 4.7 captures this structure. A causal structure where the party location affects the district location, but not the other way around, makes sense by appealing to the spatial theory underpinnings of the SPD. The party location and the district location are aggregations of many individual constituents who each have ideological ideal points. The district constituency contains all of the party constituency as well as any other constituent who has a different party affiliation or no party affiliation. This means that if we could causally intervene on the party constituency location by, for example, shifting it to the left, this is entails a leftward shift among the individual partisans in that constituency. Because these partisan individuals are also members of the district constituency, this *ceteris paribus* intervention on the party location also directly affects the district location, justifying the causal path $\bar{\theta}_g \rightarrow V_d$. From a spatial theory point of view, the only way to intervene on a party location with no downstream effect on the district location would be to introduce an offsetting shift in the ideal points of constituents outside that party or an offsetting change in the partisan composition of constituents in district overall, neither of which is entailed by a causal intervention on the party location. Furthermore, intervening on the location of the *district* has no necessary effect on the party location. This is because the district location is affected by factors other than the partisan constituency, namely the ideal points of constituents outside that partisan constituency or the relative sizes of partisan constituencies within the district.[6]

[6]If we generalize the SPD to a panel data framework, we can stipulate a mechanism by which past district

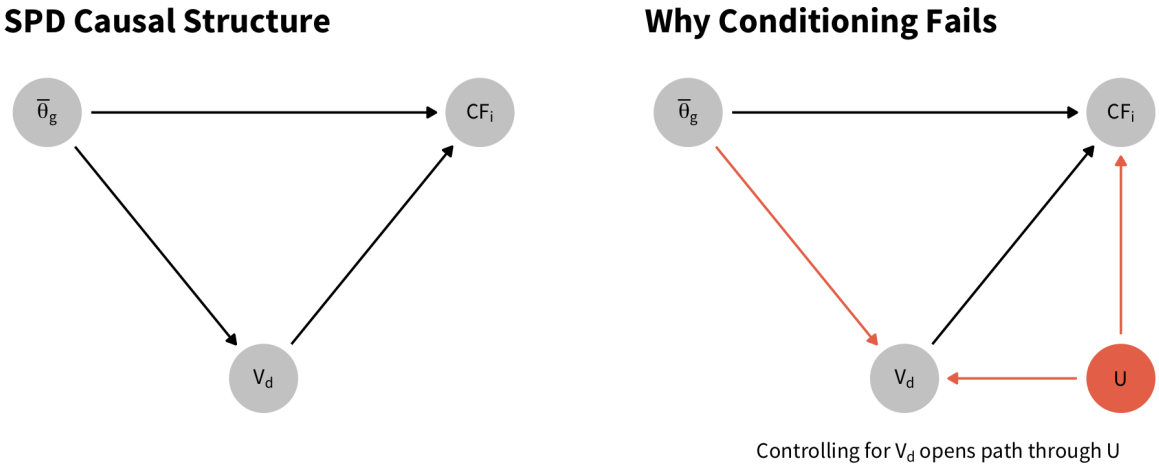


Figure 4.7: Left: A causal diagram for the strategic positioning dilemma. The district location V_d is a collider between district-party ideology $\bar{\theta}_g$ and the candidate position CF_i. Right: Conditioning on post-treatment variables can bias causal estimates by creating artificial associations through unobserved confounders.

If our primary interest is the treatment effect of district-party ideology on candidate positioning, but the district location also affects candidate positioning, it is a natural impulse to want to condition on some measure of district preferences. For instance, the previous two-party presidential vote share is an available signal of district-level preferences that candidates can consult when they position themselves, so researchers may control for presidential voting in a district in order to isolate the effect of district-party ideology on candidate positioning. This is similar to the motivation of Clinton (2006), who compared incumbent responsiveness to district median preferences vs party median preferences by including measures of each variable in a regression. Because the district location is a post-treatment variable, or a "collider" (Pearl 2009) on the causal path from the party location to candidate location ($\bar{\theta}_g \rightarrow V_d \rightarrow \text{CF}_i$), identifying the effect of district-party ideology separately from the overall district effect is

voting may affect future district-party ideology through a thermostatic opinion mechanism: heavy Democratic voting in one election leads to a Democratic presidency, which causes issue opinions across the country to become more conservative. The data in this project are unable to capture this mechanism because district-party ideology is measured for a district-party group over a districting cycle, but this is something that researchers could explore if they approach primary representation through a dynamic causal inference framework (Blackwell and Glynn 2018; Imai and Kim 2019).

more complicated than controlling for the presidential vote in a regression. This is because conditioning on a collider variable can bias causal effects by opening unblocked causal paths from treatment to outcome through unobserved confounders. This problem is diagrammed in the left half of Figure 4.7, which shows how conditioning on the presidential vote, a measure of V_d, opens a new pathway $\bar{\theta}_g \rightarrow V_d \rightarrow U \rightarrow \text{CF}_i$, biasing the estimated causal effect of $\bar{\theta}_g$. At the same time, doing nothing to adjust for aggregate district preferences may identify the total effect of the district-party constituency under certain assumptions, but it would not detect whether presidential voting absorbs all of the effect of the partisan constituency. There would be no way to contrast the unique impact of the partisan constituency from its downstream effect on overall district voting.

This analysis resolves this problem by using sequential *g*-estimation to identify a quantity called the average controlled direct effect (ACDE) of district-party ideology on candidate positioning (Acharya, Blackwell, and Sen 2016; Vansteelandt 2009). In terms of potential outcomes, let $\text{CF}_i\left(\bar{\theta}_g, V_d(\bar{\theta}_g)\right)$ be candidate *i*'s CF score as a function of district-party ideology and the past presidential vote, which represents district-level preferences V_d and is itself a function of district-party ideology. The controlled direct effect (CDE) for a unit i is the difference in CF scores for different district-party ideology treatments, holding the presidential vote fixed at some value v.

$$\text{CDE}_i\left(\theta, \theta', m\right) = \text{CF}_i\left(\bar{\theta}_g = \theta, V_d = v\right) - \text{CF}_i\left(\bar{\theta}_g = \theta', V_d = v\right) \tag{4.1}$$

The ACDE is the average of the CDEs if all units were fixed at the same mediator value v.

Sequential *g*-estimation is a structural modeling routine that estimates ACDEs by subtracting intermediary causal effects without creating collider bias (Acharya, Blackwell, and Sen 2016; Vansteelandt 2009). The routine requires adjusting for two sets of confounders: pre-treatment confounders X_d that affect both district-party ideology and CF scores, and intermediate confounders Z_d that affect both the presidential vote and CF scores. Inter-

mediate confounders are allowed to be affected by both pre-treatment confounders and district-party ideology. The first graph in Figure 4.8 diagrams the stipulated causal structure among the district-party ideology treatment, the presidential vote mediator, and both sets of confounders. In the first stage of the model, the researcher estimates the effect of the mediator on the outcome variable, conditional on all confounders.

$$\begin{aligned}\mathbb{E}\left[\mathrm{CF}_i\left(\theta, v\right) - \mathrm{CF}_i\left(\theta, v'\right) \mid \bar{\theta}_g = \theta, X_d = x, Z_d = z\right] = \\ \mathbb{E}\left[\mathrm{CF}_i \mid v, \theta, x, z\right] - \mathbb{E}\left[\mathrm{CF}_i \mid v', \theta, x, z\right]\end{aligned} \tag{4.2}$$

The left side of Equation (4.2) represents the conditional effect of the mediator in terms of potential outcomes, and the right side is the quantity that can be estimated from observed data assuming that the mediator is conditionally ignorable given all confounders and has positive assignment probability [Acharya, Blackwell, and Sen (2016);]. This specification blocks all back-door paths from V_d to CF_i, which can be seen in the first panel of Figure 4.8.

The next step of sequential g-estimation is to subtract the effect of the mediator from the outcome, also known as "demediation" or "blip-down." Demediation removes all variation in the outcome variable that is attributable to the causal effect of mediator. This stage is algebraically equivalent to subtracting a *demediation function* from the observed outcome. The mediation function in terms of potential outcomes is as follows:

$$\delta_d\left(\theta, v, \tilde{v}, x\right) = \mathbb{E}\left[\mathrm{CF}_i\left(\theta, v\right) - \mathrm{CF}_i\left(\theta, \tilde{v} = 0.5\right) \mid X_d = x\right] \tag{4.3}$$

which represents the expected effect on CF scores by setting the presidential vote to its observed value versus some fixed reference value $\tilde{v}$ for all units, conditional on X_d (Acharya, Blackwell, and Sen 2016).[7] In this analysis, I fix the two-party presidential vote to 0.5, an even split between Republicans and Democrats in the district. We subtract the demediation

[7] The model specification below functionally assumes to interactions between the presidential vote and intermediate confounders Z_d, so the specification of the demediation function in Equation (4.4) does not depend on Z_d, although this assumption is not strictly necessary for nonparametric identification of the ACDE (Robins 1997).

function from the original outcome to obtain the demediated CF score, $b\left(\text{CF}\right)_i$, which is equivalent to the potential CF score if all units had a presidential vote of 0.5.

$$\begin{aligned} b\left(\text{CF}\right)_i &= \text{CF}_i - \delta_d\left(\bar{\theta}_g, V_d, \bar{v} = 0.5, x\right) \\ \mathbb{E}\left[b\left(\text{CF}\right)_i\right] &= \mathbb{E}\left[\text{CF}_i\left(\theta, V_d = \bar{v}\right)\right] \end{aligned} \tag{4.4}$$

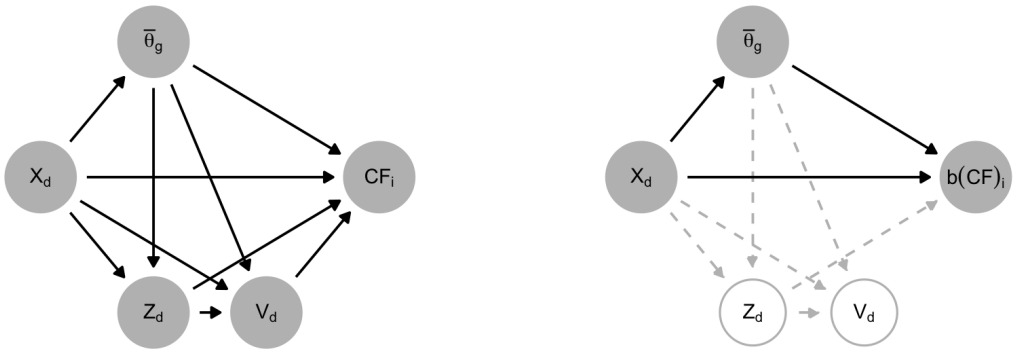
Finally, the researcher estimates the effect of the treatment on the demediated outcome, which is equivalent to the controlled direct effect on the original outcome.

$$\begin{aligned} \mathbb{E}\left[\text{CF}_i\left(\theta, \bar{v}\right) - \text{CF}_i\left(\theta', \bar{v}\right) \mid x\right] &= \mathbb{E}\left[b\left(\text{CF}\right)_i\left(\theta\right) - b\left(\text{CF}\right)_i\left(\theta'\right) \mid x\right] \\ &= \mathbb{E}\left[b\left(\text{CF}\right)_i \mid \theta, x\right] - \mathbb{E}\left[b\left(\text{CF}\right)_i \mid \theta', x\right] \end{aligned} \tag{4.5}$$

The first statement in Equation (4.5) relates the controlled direct effect in terms of CF scores to the total effect on demediated CF scores. The second statement defines how the ACDE can be estimated with observable data under the assumption that assignment to district-party ideology ignorable and with positive probability given pre-treatment covariates (Acharya, Blackwell, and Sen 2016). The top-right graph in Figure 4.8 shows how the demediation step recovers the controlled direct effect of district-party ideology. Demediating the outcome removes all variation in CF scores caused by the presidential vote, so it deletes the path $V_d \rightarrow \text{CF}_i$ in the diagram. In turn, there is no need to condition on the presidential vote V_d in Equation (4.5) to identify the ACDE, even though the diagram shows that district-party ideology has an effect on the presidential vote. Furthermore, the graph contains remains a causal path from the intermediate confounders Z_d to the CF score, the stage-two model does not condition on these confounders because these pathways are a part of the district-party ideology's ACDE on CF scores. It is worth noting here that if we estimate the stage-two model using the original CF score rather than the demediated CF score, this would estimate the total effect of district-party ideology. This quantity can be valuable because valuable because the difference between the total effect and the controlled direct effect shows how much of the total effect flows is carried by a mediating mechanism.

Stage 1
Identifies mediator effect

Stage 2
Identifies controlled direct effect
of treatment using demediated outcome



Violations of Sequential Ignorability
In stage 1 (U1) and stage 2 (U2)

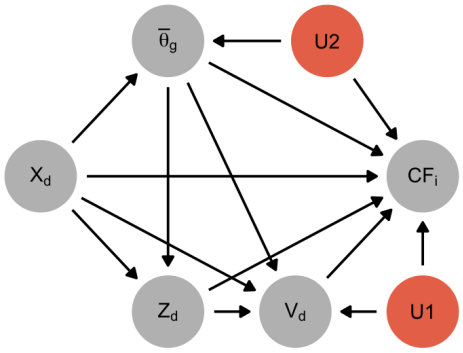


Figure 4.8: Causal graphs describing the modeling problem and sequential g estimation. The stage 1 graph identifies the effect of the past district voting (V_d) on candidate positioning (CF_i). The stage 2 treatment-outcome model subtracts the district vote effect from candidate positions and identifies the effect of district-party ideology $\bar{\theta}_g$ on the demediated CF score $b(\text{CF})_i$, which is equivalent to the controlled direct effect on the raw CF score. The final graph shows where unadjusted confounders violate the nonparametric causal identification assumptions in stage 1 ($U1$) and in stage 2 ($U2$).

The bottom graph in Figure 4.8 shows where unmeasured confounding can violate the two ignorability assumptions required to estimate the ACDE. The stage-one model identifies the causal effect of the past presidential vote, so if an unmeasured variable (represented in the figure by $U1$) affects both district voting and CF scores, the mediator's effect is not identified. Similarly, the stage-two model does not identify the effect of district party ideology on the demediated CF score if they share an unmeasured cause $U2$. Unmeasured variables in other locations of the graph certainly exist, but they do not violate the sequential ignorability assumptions unless they can be represented by an unblocked back-door path through $U1$ or $U2$.

With few exceptions, it is not typical for research on primary politics to incorporate explicit causal inference methodologies (Doherty, Dowling, and Miller 2019; Fowler and Hall 2016; Hall 2015; Hall and Thompson 2018). Although many other studies use a selection-on-observables research design to study candidate positioning in primary elections, a major contribution of this study's research design is the use of structural causal modeling to define, identify, and estimate a specific causal quantity.

4.2.1 Sequential g implementation

The average controlled direct effect of district-party ideology on primary candidate CF scores is *nonparametrically* identifying under SUTVA, sequential ignorability, and positivity. I estimate the sequential g model using a linear model specification laid out by Acharya, Blackwell, and Sen (2016). As with any linear model, this specification imposes additional functional form and distributional assumptions on the data, which are helpful for confounder adjustment in the absence of other design-based variation from instruments or discontinuities, but could lead to incorrect inferences if assumptions are false. One goal for future work in this literature would be to combine the sequential g structural model with semi- and non-parameteric estimation methods, an nascent area of work that exists largely outside of

political science (Athey, Imbens, and Wager 2016; Chernozhukov et al. 2017; Hahn, Murray, and Carvalho 2020; Hill 2011; Ratkovic 2019; Samii, Paler, and Daly 2016; Wager and Athey 2018)

This section lays out the linear specification for sequential g-estimation. I define the model using notation that is general enough to apply to any subset of data where the model is estimated. As mentioned above, I investigate subsets of data by incumbency and primary rules, and I estimate each model separately for Republicans and Democrats. The data contain measures of candidates i in district-party groups g in districts g. Because all models are estimated separately for each party, groups and districts perfectly overlap. Nonetheless, I use g for variables that can vary across party within a district and d for variables that are fixed for both parties in a district. Because the estimation contains in two stages of regression modeling, I sometimes subscript some parameters with 1 or 2 to indicate which equation they belong to.

The first stage is a mediator–outcome model that estimates how the Republican two-party presidential vote in the past election V_d affects the CF score of candidate i within group g. We set up the sequential g method to control for the previous Republican presidential vote share in district d, denoted $pvote_g$. This is done with the following multilevel regression model.

$$\begin{aligned} \text{CF}_i &= a_0 + \mu \text{V}_{d[i]} + \eta \bar{\theta}_{g[i]} + \mathbf{x}_{d[i]}^{\top} \beta + \mathbf{z}_{d[i]}^{\top} \gamma + \alpha_{d[i]} + \varepsilon_i \\ \alpha_d &\sim \text{Normal}\left(0, \sigma_\alpha\right) \\ \varepsilon_i &\sim \text{Normal}\left(0, \sigma_\varepsilon\right) \end{aligned} \tag{4.6}$$

The group ideal point $\bar{\theta}_g$ is included in the regression as a control, so its coefficient η is estimated as a nuisance parameter, as are the coefficients β for district-level pre-treatment confounders $\mathbf{x}_d$, the coefficients γ for district-level intermediate confounders $\mathbf{z}_d$, and the constant a_0. Because the mediator is measured at the district level, I include a district error term α_d in addition to the candidate-level error term ε_i. This multilevel model accounts

correlated error among candidates in the same district-party group, similar to clustering standard errors when a treatment is dosed at the cluster level.

We then use the estimates from the stage-1 model to demediate the CF score variable. Because the first stage is a linear model, the demediation function has a straightforward parametric definition,

$$\delta_d = \mu \left(\mathrm{V}_d - \bar{v} \right) \tag{4.7}$$

where $\bar{v}$ is the reference value for the mediator where all units are fixed, which I set equal to 0.5 to represent a 50-50 split in the previous two-party presidential vote.[8] The demediation function is subscripted d because it varies across districts according to the observed value of the previous Republican vote share, which also entails that the demediation function is equivalent for all candidates in the same district-party group, regardless of their original CF score. We calculate the demediated outcome $b(\mathrm{CF})_i$ by subtracting each district's demediation function from its observed outcome value,

$$b \left(\mathrm{CF} \right)_i = \mathrm{CF}_i - \delta_d \tag{4.8}$$

The demediated outcome is then used in the second stage estimation of the controlled direct effect. The second stage is different from the first stage in two ways. First, because the demediated outcome fixes the value of the mediator, there is no more variation across observations that can be attributed to the mediator, so V_d is omitted from the equation. Second, intermediate confounders are omitted because they can be affected by district-party ideology, and as such should be left unadjusted in order to identify the ACDE. This new

[8] The demediation function can be more complex if the mediator's effect on the outcome is modeled with a more complex model containing interactions or nonlinearities.

equation is,

$$
\begin{aligned}
b\left(\mathrm{CF}\right)_i &= a_1 + \tau\bar{\theta}_g + \mathbf{x}_{g[i]}^{\top}\omega + v_{d[i]} + u_i \\
v_d &\sim \mathrm{Normal}\left(0, \sigma_v\right) \\
u_i &\sim \mathrm{Normal}\left(0, \sigma_u\right)
\end{aligned}
\tag{4.9}
$$

where a_1 is a constant, τ is the coefficient for district-party ideology, ω are coefficients for district-level pre-treatment confounders $\mathbf{x}_d$, v_d is a district error term, and u_i is a candidate error term. In a linear model specification, τ measures the average total effect of a one-unit increase in district-party ideology on demediated CF scores, which is equivalent to the ACDE on the original CF scores. More generally, the ACDE of setting district-party ideology from $\bar{\theta}'$ to $\bar{\theta}$ is as follows.[9]

$$
ACDE(\tau, \theta, \theta') = \tau\left(\theta - \theta'\right) \tag{4.10}
$$

4.2.2 Bayesian modeling and interpretation

Another key methodological innovation in this chapter is embedding sequential *g*-estimation in a Bayesian framework. Chapter 3 describes a number of special advantages for Bayesian causal modeling, stemming from the fact that Bayesian inference allows the researcher to conduct posterior inference about causal effects using probabilities: the treatment effect is *probably* greater than x, where "probably" is defined in relation to an empirical cumulative probability distribution function of posterior samples.

The most important feature of Bayesian causal inference for this chapter is the fact that one posterior distribution quantifies uncertainty in all parameters from the multi-stage sequential *g* method. This is valuable because uncertainty in stages 1 and 2 are directly related to one another by way of the demediation function. Whereas non-Bayesian analysis requires either

[9] Again, the exact formula for the ACDE depends on the model specification. The linear model with no treatment interactions produces a simple linear ACDE definition, but a more complex model would entail a more complex formula.

ad-hoc variance corrections for the multistage model or a bootstrapping approach (Acharya, Blackwell, and Sen 2016), the posterior distribution from the Bayesian model captures all variances and covariances among model parameters by its very nature. Inferences about the ACDE can then be expressed by marginalizing the posterior distribution with respect to these auxiliary model parameters, isolating the remaining dimension of the posterior corresponding to the ACDE. Letting $\boldsymbol{\pi}$ represent all auxiliary model parameters, and using the definition of the ACDE in Equation (4.10), the posterior distribution for the ACDE is given by

$$p\left(\tau\left(\theta - \theta'\right) \mid \mathbf{CF}\right) = \int p\left(\tau\left(\theta - \theta'\right), \boldsymbol{\pi} \mid \mathbf{CF}\right) \, \mathrm{d}\boldsymbol{\pi}, \tag{4.11}$$

which is a the distribution of ACDE values that condition on the observed data and marginalize over the associated auxiliary parameter values. In practice, we use posterior samples to approximate this posterior distribution by picking any two values θ and θ', extracting posterior samples for τ, and calculating the ACDE for each sample iteration. Posterior expectations for the ACDE can be calculated, up to Monte Carlo error, by averaging the ACDE values from a draw of posterior samples. Uncertainty intervals for the ACDE can be estimated by noting which posterior samples bound the inner 90 or 95% of posterior samples.

The joint posterior distribution is also essential for incorporating uncertainty in district-party ideal points themselves. The IRT model in Chapter 2 does not estimate district-party ideal points exactly. Instead, ideal points are estimated only up to a posterior distribution, with uncertainty that reflects both prior ignorance and a finite sample of polling data. To estimate the effect of district-party ideology on candidate positioning, the causal analysis incorporates measurement error in ideal points "the Bayesian way," where posterior uncertainty from one analysis becomes prior uncertainty in later analyses. One way to implement this "full uncertainty" would be to build one joint model containing both the IRT measurement model and causal inferential models, although this would be a burdensome feat of computer

programming and computationally expensive to run each model. Instead, I approximate the joint model by constructing a prior for $\bar{\theta}_g$ in the causal model by approximating the posterior distribution from the IRT model. This prior is constructed by defining the group ideal point $\bar{\theta}_g$ as an element of Θ, the vector of all group ideal points, which gets a multivariate Normal prior.

$$\Theta \sim \text{MultiNormal}\left(\hat{\Theta}, \hat{\Sigma}_{\Theta}\right) \tag{4.12}$$

The hyperparameters in the prior are estimated from MCMC samples for the ideal points. The mean vector $\hat{\Theta}$ contains MCMC means for each ideal point, and the matrix $\hat{\Sigma}_{\Theta}$ contains variances for each ideal point on the diagonal and covariances between any two ideal points on the off-diagonals. Because the IRT model partially pools ideal points using a hierarchical Normal regression, the multivariate Normal prior is a reasonable stand-in for representing prior ideal point uncertainty in causal analyses.

Including ideal point uncertainty as a prior distribution effectively adds a measurement error model overtop to the sequential *g* analysis. Although Acharya, Blackwell, and Sen (2016) describe a variance estimator and a bootstrap method for dealing with multi-stage modeling uncertainty, neither of these methods is naturally suited to a measurement error context where an additional layer of ideal point uncertainty is represented in posterior samples. Past research has explored the use of inverse-variance weighting to downweight observations with greater measurement uncertainty (e.g. Adolph et al. 2003), which accomplishes a similar goal as the prior distribution but throws out all prior information about the covariance between ideal points. More recently, researchers have employed numerical methods for "uncertainty propagation," where the researcher estimates an uncertain quantity, simulates values from the quantity's posterior distribution, and pushes those simulated values through a downstream analysis. Quantities of interest are then averaged across the posterior simulations (see Kastellec et al. 2015, 791; Caughey and Warshaw 2018, 6, 2019, 360). This is similar in spirit

to "multiple overimputation" (Blackwell, Honaker, and King 2017), which extends multiple imputation to a measurement error setting by iteratively replacing mismeasured observations with draws from their posterior distribution. This project regards these propagation methods as insufficiently Bayesian because they "cut" the flow of information between models. Posterior cutting allows model 1 to inform model 2, but model 2 can never inform model 1 (Plummer 2015). This can be an undesirable modeling property if the causal model can inform the measurement model (Treier and Jackman 2008). For instance, if ideal points are related to candidate positioning, and ideal points are measured with error, then observing candidate positions can update our information about ideal points. By specifying the prior directly and updating the parameters, there is no need for any additional imputation steps or post-estimation model averaging (Gelman and Hill 2006, 542).

The Bayesian sequential *g* approach, of course, requires priors for other model parameters as well. As I describe below in Section 4.2.4, most variables in the model are binary indicators or are standardized to be mean 0 and variance 1. Furthermore, most of the outcome data in each model fall mostly in the $[-2, 2]$, so we can place standard Normal priors on the covariates without being overly informative about the covariate effects and introducing regularization bias (Hahn et al. 2018). Constants are given wider priors to account for the fact that not all predictors are exactly centered at means.

$$\begin{aligned} a_0, a_1 &\sim \text{Normal}\,(0, 5) \\ \eta, \mu, \beta, \gamma &\sim \text{Normal}\,(0, 1) \\ \tau, \omega &\sim \text{Normal}\,(0, 1) \end{aligned} \tag{4.13}$$

Both modeling stages contain Normal district-level error terms with estimated variances that facilitate partial pooling. These variances are themselves given half-Cauchy priors with weakly informative scale parameter values that are about half the range of the raw outcome

data within each party.

$$\begin{aligned}
\alpha_d &\sim \text{Normal}\,(0, \sigma_\alpha) \\
\sigma_\alpha &\sim \text{Half-Cauchy}\,(0, 1) \\
\nu_d &\sim \text{Normal}\,(0, \sigma_\nu) \\
\sigma_\nu &\sim \text{Half-Cauchy}\,(0, 1)
\end{aligned} \tag{4.14}$$

And finally, each stage of the model has a Normal error term for candidates within districts. The variances for these errors are given half-Cauchy priors with wider scale values than the districts errors, since residual variation between any two candidates is likely larger than the variation between average candidates in any two districts.

$$\begin{aligned}
\varepsilon_i &\sim \text{Normal}\,(0, \sigma_\varepsilon) \\
\sigma_\varepsilon &\sim \text{Half-Cauchy}\,(0, 2) \\
u_i &\sim \text{Normal}\,(0, \sigma_u) \\
\sigma_u &\sim \text{Half-Cauchy}\,(0, 2)
\end{aligned} \tag{4.15}$$

4.2.3 Causal inference with multilevel data

The research question in this chapter presents us with multilevel data: how is the ideological positioning of primary candidates affected by the policy ideology of partisans in their district, when there are potentially multiple candidates per district? In this scenario, the outcome is a variable specific to an individual candidate i, but the treatment is fixed for an entire district-partisan group g. This introduces a few issues for statistical assumptions and causal assumptions.

On the statistical front, multilevel models bias coefficient estimates when the aggregate errors are not exchangeable. Mechanically, this is similar to "omitted variable bias" in a single-level regression. Although this concern is well founded for many multilevel models, for these

models we can be less concerned. Because all predictors in these regressions are measured at the district level, the district error term is analogous to an error term that we would obtain by averaging every candidate's CF score within a district-party group and running a single-level regression on those averages. Both of these model specifications require an exchangeable errors assumption at the district level. The only difference for the models in this analysis is the additional candidate-level errors, but this too is a non-issue. Averaging candidate data within each district would invoke a similar assumption about the exchangeability of candidates given the district, otherwise it would be inappropriate to average data within a district.

Even though the multilevel model has similar assumptions as a regression on averages, it has certain benefits that are convenient for these data. Because the number of candidates in a district isn't fixed across all districts, we would expect heteroskedasticity in a regression-on-averages model, since some districts would have higher variances due to fewer candidates. In the extreme case, if a district contained only one unopposed primary candidate, a naïve estimator would be unable to distinguish district-level variance from candidate-level variance. The multilevel model addresses this by estimating the distributions of district errors and candidate errors simultaneously, enhancing the model's ability to recognize when larger district errors are caused by signal versus noise. Errors from smaller districts borrow more information from the overall distribution of districts, downweighting the contributions of smaller districts by shrinking their error terms toward a mean of zero. This has a similar intuition as a weighted least squares regression on the district-averaged data, where groups with more observations are more informative about global parameters and receive greater weight. This is yet another example where priors stabilize pathological model behavior, underscoring the flexibility afforded by Bayesian model-building for confronting the idiosyncrasies of a dataset with tactics that are both intuitive and feasible.

The multilevel data structure also raises causal inference issues that are worth clarifying. As with many causal models where treatments are assigned to clusters of observations, it

makes sense to consider SUTVA as violated within a cluster: there is no way for one candidate in a district to be treated by a different district-party ideology than other candidates.[10] The positioning of one candidate may also affect the positioning of another, which could violate the "no interference" component of SUTVA. Under this violation, the treatment effect at the individual level is not identified. If SUTVA holds *between* groups, however, it is possible to identify a treatment effect by considering average effects across groups (Hill n.d.). In potential outcomes notation, even if we can define potential outcomes at the individual level ($CF_i(\bar{\theta}_{g[i]})$), the lowest level where we could credibly *identify* treatment effects would be the group level, where the potential outcome for a group is the average outcome within the group ($\overline{CF}_g\left(\bar{\theta}_g\right)$). This is consistent with the multilevel model setup that we have so far, where the ACDE is a function of aggregate data and aggregate parameters only (see Equation (4.10)).

There are a few additional considerations for causal inference with hierarchical data that, although I do not pursue these threads in this project, could be relevant for future work with similar data. A correlation between treatment effects and group size may arise if a crowded primary field causes larger treatment effects because stiffer competition leads candidates to be more responsive to district-party ideology. On the other hand, more crowded fields would lead to smaller treatment effects if candidates take heterogeneous ideological positions to differentiate themselves. If treatment effects are correlated with group size, then the average causal effect for a candidate is not equivalent to average difference among groups. Instead, the average effect for candidates must be a size-weighted average of group effects (Hill n.d.). I do not pursue this possibility in this project because these dynamics are not identifiable with data on primary candidates only, since incumbents may take ideological positions to deter challengers even if no challengers actually emerge. As such, the observed number of candidates in a district may not capture the true degree of primary threat (Hirano et al. 2010;

[10] Candidates may vary in their ability to perceive district-party ideology, but that might also be described as an issue of treatment compliance or treatment effect heterogeneity.

Maisel and Stone 1997; Stone and Maisel 2003).

One additional consideration for group-level effects is the possibility that group size affects treatment assignment. This may be true if the long-run dynamics of primary competition within a district-party have feedback effects on local ideology, for instance if partisan constituents become more ideologically aware by experiencing stronger intra-party competition in their district, or less ideological after a long period of representation by a single incumbent with little primary competition. There is evidence that primaries contain more ideological campaign content in certain periods of heightened partisan mobilization (Boatright 2013), which could increase voters' ideological awareness as well. Whether voters are responding to primary competition in the district *per se* or to a national state of partisan agitation is an interesting but thoroughly challenging question for future researchers to explore, were they to extend the data and methods in this project to a greater number of election cycles and dynamic causal modeling approaches (e.g. Blackwell and Glynn 2018; Imai and Kim 2019).

4.2.4 Data

For the CF score outcome measure, I specifically use the dynamic CF score provided by Bonica (2019b), which is re-estimated for each two-year FEC cycle. The measure of district-level partisanship, the mediator in the sequential g-estimation routine, is the two-party Republican presidential vote share from the previous election cycle, which is provided and matched to candidacies by Bonica (2019b). District-party ideology $\bar{\theta}_g$, the treatment variable, is measured from the ideal point model in Chapter 2.

Pre-treatment confounders $\mathbf{x}_d$ are included to identify the effect of the presidential vote in stage 1 and the effect of district-party ideology in stage 2. These covariates were organized and matched to primary candidacies by the Primary Timing Project by Boatright, Moscardelli, and Vickrey (2017) District demographic variables (sourced originally from the American Community Survey) include district median income, population density, and land area,

as well the percent of a district population that is White, Latino, college educated, below the federal poverty line, unemployed, employed in the service industry, employed in blue-collar jobs, aged 28–24, and aged 65+. The Boatright, Moscardelli, and Vickrey (2017) data also provide Mayhew (1986)'s five-level "traditional party organization" and binary "persistent factionalism" classifications for state level (Mayhew 1986). The stage-1 model also contains intermediate controls $\mathbf{z}_d$ for identifying the effect of the past presidential vote on CF scores. These variables include the district-party ideal point for the *other* party group in the same district, because both parties should partially affect aggregate district voting, and year fixed effects to capture average shifts in CF scores that are correlated with average shifts in presidential voting.[11]

All controls except for the fixed effects and binary indicators from Mayhew (1986) are centered at their means and scaled by their standard deviations. This makes it easier to specify priors for coefficients, since standardized coefficients are unlikely to exceed a 1 except when predictors are highly correlated. District-party ideology $\bar{\theta}_g$, the treatment variable, is measured on its original scale anchored by the item parameters in the Chapter 2 model. The presidential vote variable is centered at its reference value for demediation, 0.5, and then divided by 10 so that a one-unit change in the model represents a 10 point change in vote share. CF scores are measured on their original scale, which spans roughly as wide as -5 to 5 across both parties, both the vast majority of Democrats occupy values in $[-2, 0]$ and Republicans in $[0, 2]$.

4.3 Findings

I estimate models in several subsets of data. First, I estimate separate models for all Democratic and all Republican candidates in the sample. I then estimate models that divide each

[11] Fixed effects are not included in pre-treatment controls because district-party ideology is fixed across the redistricting cycle, so they do not improve causal identification in stage 2.

Table 4.1: Sample sizes in all estimated models.

Full Sample	By Incumbency Status	By Primary Rules
Democrats = 1,970	Incumbents (D = 568, R = 704)	Closed (D = 933, R = 1,026)
Republicans = 2,192	Challengers (D = 1,089, R = 1,096)	Semi-Closed (D = 332, R = 375)
	Open Seats (D = 313, R = 392)	Open (D = 705, R = 791)

partisan subsample into incumbent candidates, challengers of incumbents, and candidates running for an open congressional seat. Finally, I estimate separate models for candidates running in closed primaries, semi-closed primaries, and open primaries. Table 4.1 shows the sample sizes in each model subset.

For the sake of computation time, I estimate these models by approximating the posterior distribution using the mean-field variational Bayes routine available in Stan (Kucukelbir et al. 2015). Variational Bayesian inference (VB) finds and optimizes a simpler distribution that is similar to the true posterior distribution in terms of Kullback–Leibler divergence (Grimmer 2011). Variational estimators are asymptotically consistent and can be used to estimate any Stan model. But because they require approximations, samples from the approximate distribution tend to underestimate the variance in the true posterior distribution (Wang and Titterington 2012). The benefit, however, is that models that would take hours to estimate using MCMC can be estimated using VB in roughly one minute, which is extremely valuable for building and estimating the 14 models included in this chapter.

Before turning to the results, recall that a distinguishing feature of the Bayesian approach is the use of priors to represent measurement error in district-party ideal points. This is an intuitive solution to the problem of uncertainty propagation because uncertainty in causal effects naturally reflects uncertainty in the data by marginalizing over the ideal points. Another interesting consequence is that the posterior distribution for the ideal points could be different from the prior, depending on the information that the inferential the model

can provide about the values of the ideal point parameters. I plot the prior and posterior distributions for ideal points alongside one another in Figure 4.9, using points to represent prior and posterior means and bars to represent uncertainty intervals. Posterior estimates come from the models estimated on the full samples of Democrats and Republicans. The data fall along the 45-degree line, indicating that prior and posterior ideal points are similar to one another, which is a sensible result that increases our confidence in the computational accuracy of the model. The similarity between prior and posterior ideal points is also convenient for understanding the regression results, because coefficients can be interpreted in the original scale of the data without any need to post-process results into a familiar scale.[12]

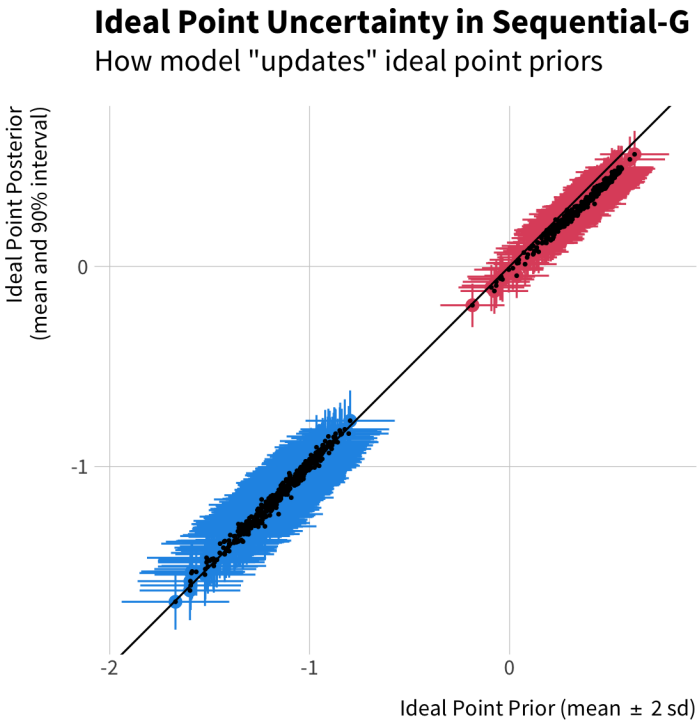


Figure 4.9: Marginal prior and posterior distributions for ideal points in sequential *g* model.

[12] In a regression context like this, the ideal points and their regression coefficient are not mutually identifiable because each could be arbitrarily scaled in offsetting ways. Estimation with MCMC may suffer under this non-identifiability by "wandering" through the weakly identified regions of parameter space, which can be corrected by post-processing parameter samples by applying scale constraints to each MCMC interaction. The variational algorithm is deterministic and does not wander in the same way, which eliminates the need to post-process estimates to recover sensible answers.

We now turn to the sequential *g* results. Figure 4.10 plots VB posterior means and 90% intervals for the full sample of Democratic candidates. Stage 1 parameters are plotted as squares, and stage 2 parameters are plotted as circles. The top-left panel contains the key parameters most relevant to causal inference, including the coefficients for the mediator and treatment variables from both stages of the estimation. The bottom-left panel plots the standard deviations of the district errors and residual errors. The right-side panel contains regression coefficients for all control variables.

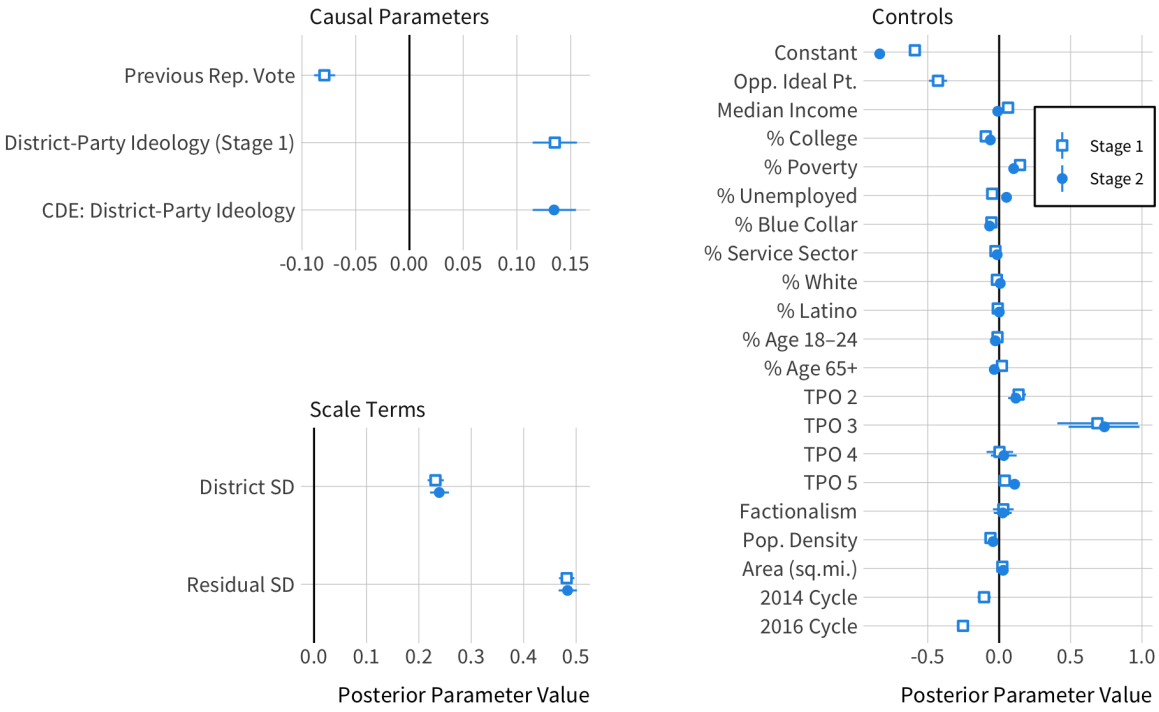


Figure 4.10: Sequential *g* results for Democratic candidates. Points and intervals are variational point estimates and 90 percent intervals from the approximate posterior distribution.

Under sequential ignorabilitiy and no intermediate interactions assumptions, stage 1 estimates the effect of the past presidential vote share on CF scores. Interstingly, this coefficient

is negative, indicating that Democratic primary candidates running in districts that vote more heavily Republican are actually more *progressive* than Democratic candidates who represent more Democratic-leaning districts. This is is an unexpected relationship given the general findings in the literature that candidates take more moderate stances in more moderate districts, which should manifest as a positive coefficient. It is important to keep in mind that this sample of data contains incumbent and non-incumbent candidates, as well as primary winners and losers. Because most primary elections do not occur under highly competitive circumstances, the typical primary campaign may not resemble the predictions from "Downsian" formal models that assume perfect candidate competition (Burden 2004). The mixture of incumbent and non-incumbent candidates may weight the sample more heavily toward candidates who corner more extreme or idiosyncratic candidate positions in their attempts to attract attention away from incumbents, who are likely to position themselves more in line with overall district preferences (Ansolabehere, Snyder, and Stewart 2001). The models that stratify on incumbency in Section 4.3.1 are consistent with these possibilities. One other possibility is that the introduction of my new measure, district-party ideology, is responsible. If candidates are indeed sensitive to *partisan* preferences in their district, and partisan preferences are positively correlated with district voting on average, then the strong relationship between candidate positions and the district vote was at least partially confounded in all past studies. The results in Figure 4.10 are consistent with that possibility, since the coefficient on district-party ideology is large and positive even in stage 1, when it does not have a causal interpretation but is instead included as a control to identify the presidential vote effect.

The controlled direct effect of district-party ideology is positive, with a posterior mean indicating that one-unit increase in district-party ideology causing a 0.13-unit increase in CFscores. Using standardized coefficients, the posterior mean suggests that an increase of one standard deviation in district-party ideology causes 0.04-standard deviation increase in

CF scores, which is a substantively small effect.

The control coefficients do not have causal interpretation, and even interpreting them as partial correlations is problematic because of collider bias. Nonetheless readers might find some of the coefficient estimates intriguing or intuitive given the makeup of the Democratic Party coalition and the polarized environment of the 2010s. The opposing party ideal point mean has a clear negative coefficient, indicating that more conservative Republicans coincide on average with more progressive Democrats. Progressivism among Democrats is greater in districts with greater density, less land area, and greater numbers of service sector employees. The year fixed effects suggest candidates are more liberal in more recent years, which fits a pattern of polarization over time. Some interesting or counter-intuitive findings are that Democrats are more progressive in districts with more college graduates and more blue-collar workers, and they are more conservative in districts with higher poverty.

Turning to the Republican estimates in Figure 4.11, we find a similar pattern among the causal parameters. The presidential vote is again inversely related to ideal points, with greater Republican voting causing more conservative Republican candidates under the causal assumptions. We also find a positive controlled direct effect of district-party ideology in the stage 2 model, indicating that Republicans run more conservative campaigns in more conservative districts. The district-party ideology effect in the Republican Party is larger than the effects in the Democratic Party on the scale of the raw data, with a coefficient of 0.27 (versus 0.13). In standardized coefficients, the effects are more similar, with a Republican coefficient of 0.06 versus the Democratic coefficient of 0.04. This is because district-party ideology is more similar across Republican groups than Democratic groups, with a standard deviation in ideal point means of 0.11 versus 0.17, meaning that one standardized unit increase is a smaller increase in absolute terms among Republicans than among Democrats.

Benchmarking these cross-party comparisons using raw or standardized coefficients has consequences for the conclusions we can draw about representation in the two parties. One

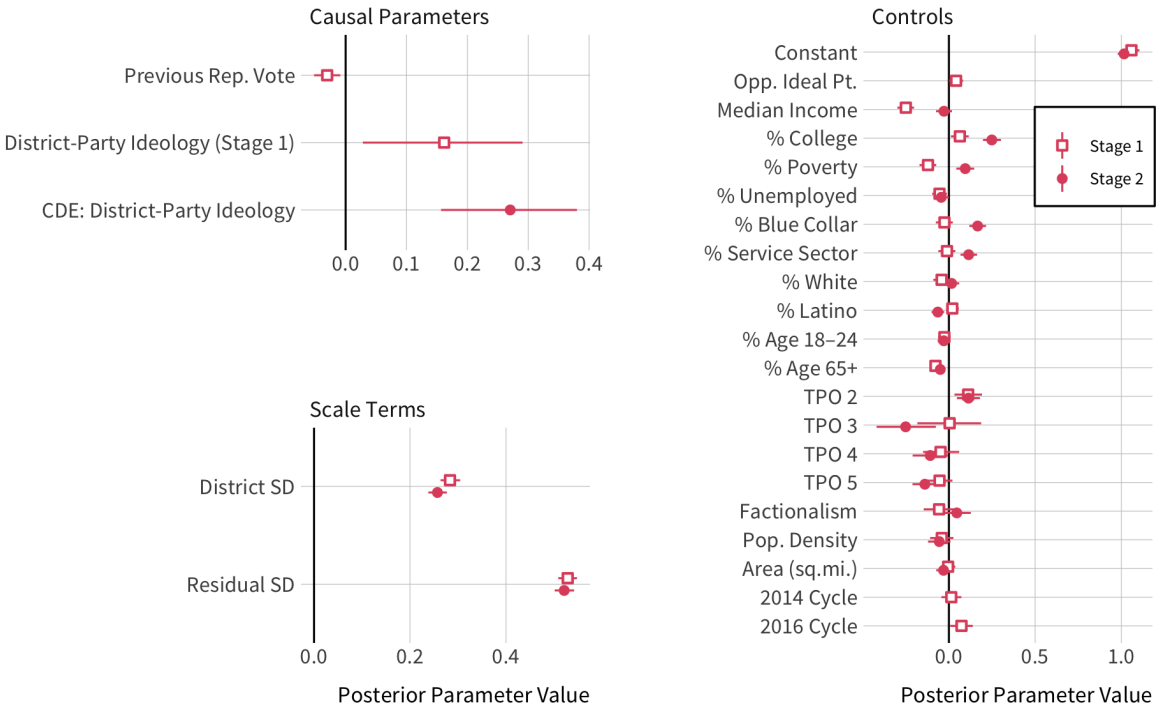


Figure 4.11: Sequential *g* results for Republican candidates. Points and intervals are variational point estimates and 90 percent intervals from the approximate posterior distribution.

interpretation places greater emphasis on the standardized picture: the initial appearance of greater responsiveness among Republicans using raw coefficients (e.g. Clinton 2006) makes an incorrect assumption that the scope of ideological conflict is the same among Democrats and Republicans, when in fact Republican voters and elites are much more ideologically cohesive than Democrats (Lelkes and Sniderman 2016). But conditioning on the actual scope of ideological conflict in the two parties, Republican and Democratic candidates respond to their constituencies with proportional adjustments to their campaign positions. Another interpretation places greater emphasis on the raw coefficients: if the scope of ideological conflict within the Democratic Party is greater, this is essential to keep in mind

for understanding how candidates position themselves in relation to voters. If the Democratic Party contains a larger variety of conflicting group interests, and elite party members have to reign these interests into a coherent platform, the result could be an elite party that appears to be insensitive to the big-tent heterogeneity in voters' policy preferences.

4.3.1 Effect modification: incumbency status and primary rules

This section examines the results for models estimated within strata defined by candidate incumbency status and state primary rules. It is an important to note that conditional effects do not necessarily represent the causal effects of stratum membership, a distinction that is often overlooked in examinations of heterogeneous causal effects (Kam and Trussler 2017). Differences in the CDE across incumbency or primary rules reflect causal heterogeneity, but the sources of heterogeneity could come from factors that confound the the link between strata and outcomes. As such, these results should be seen from an "effect modification" point of view. In order to claim that incumbency or primary rules *cause* heterogeneity in the effects of district-party ideology requires additional assumptions that incumbency or primary rules are ignorable, which are tasks large enough to fill separate dissertations altogether.

I first examine effects by stratifying across candidate status, estimating separate models for incumbents, challengers of incumbents, and open seat candidates. Prevailing literature that examines incumbency and candidate positioning have diverging predictions for these results. Ansolabehere, Snyder, and Stewart (2001) find that incumbents take more moderate campaign positions than non-incumbent candidates, with challengers taking more extreme than both incumbents and open seat candidates. This reflects a perspective that incumbency selects for candidates that take median positions that ensure their reelection, while challengers must take more innovative and extreme stances to garner attention. Burden (2004) finds instead that incumbents take more partisan positions, perhaps because incumbency provides other advantages that give incumbents more security to take positions that don't match district

preferences as closely.

The findings in this study contain a mixture of evidence for against each of these views. Figure 4.12 plots the mediator effect and controlled direct effect for incumbents, challengers, and open seat candidates. Among Democrats, incumbents appear to follow patterns of positioning dictated by the SPD. Coefficients for the presidential vote and district-party ideology are both positive, indicating that Democrats run as more progressive candidates in districts with higher Democratic presidential voting and with more progressive partisan constituents. Challenger positioning is inversely related to the presidential vote but positively related to district-party ideology, indicating that challengers target the partisan constituency at the expense of the general election constituency. This is consistent with an account of incumbent–challenger dynamics where challengers take extreme positions to garner attention to themselves against a more moderate, or that the responsiveness of incumbents results from the selection of well-fitting candidates by district voting (Ansolabehere, Snyder, and Stewart 2001). Open seat candidates behave more similarly to incumbents than to challengers, which contradicts the finding by Ansolabehere, Snyder, and Stewart (2001) that open seat candidates may be most "out of step" with district preferences, and may be more consistent with the Burden (2004) argument that non-incumbents position themselves most aggressively to match voter preferences because they have fewer build-in advantages to overcome any ideological mis-fits.

Among Republicans, the results suggest that incumbent candidates are responsive to within-party preferences, but essentially unrelated to aggregate district voting. This is consistent with a Clinton (2006) view that Republicans position themselves to fit their partisan constituencies, and correlation with district voting is incidental. This also fits with the Burden argument that incumbency provides insurance that lets incumbents deviate from district-optimal preferences, which is also consistent with the finding that Republican challengers are more responsive to district voting than Republican incumbents. Results among Republican

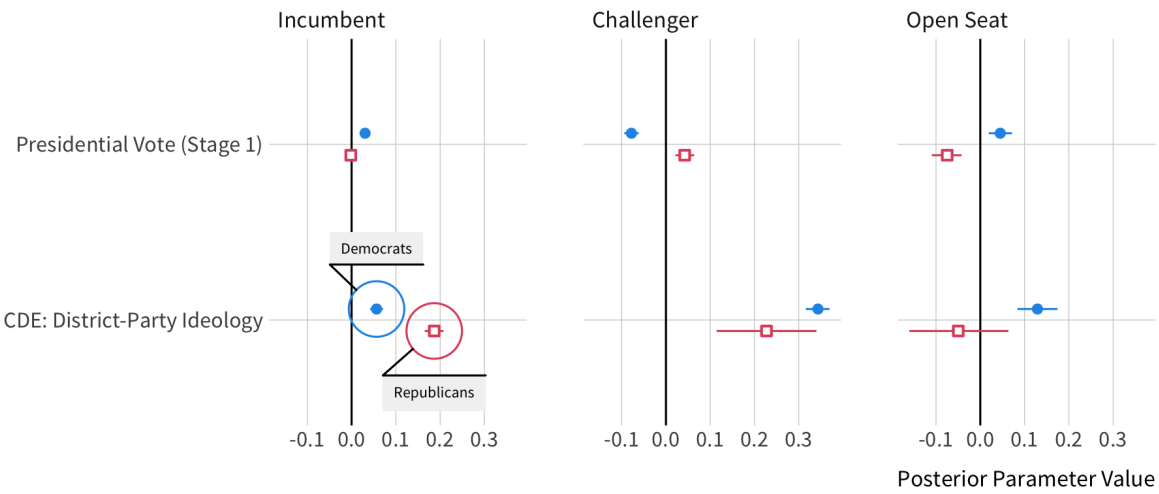


Figure 4.12: Sequential *g* results for incumbents, challengers, and open seat candidates. Points and intervals are variational point estimates and 90 percent intervals from the approximate posterior distribution.

incumbents are difficult to interpret, since the results show a negative effect of Republican voting and no clear relationship between district-party conservatism and candidate conservatism. Overall, these results suggest that nearly candidates of all incumbency status are responsive on average to district-party ideology, which supports an SPD view of primary competition. Among incumbents, the greater degree of general-election responsiveness by Democrats than Republicans tracks Clinton (2006), but the non-uniform findings among challengers and incumbents lack a clear interpretation.

We now turn to an examination of primary rules. Past studies of primary rules typically find no clear effects of different rules either on candidate positioning or vote choice. These studies never measure district-party ideology, however, so has never been clear whether candidates are in fact more responsive to district-party ideology in less-open primary systems. My results in Figure 4.13 suggest that the inclusion of district-party preferences is consequential for understanding primary openness. Among Democrats, the controlled direct effect

of district party ideology is strongest in closed systems and weakest in open systems, with semi-closed systems in the middle. This monotonic relationship is exactly consistent with the conventional wisdom on primary openness and candidate positioning. The results among Republicans are not monotonic—the strongest relationship is among semi-closed rather than closed systems—but generally support a conclusion that candidates in open systems are less responsive to district-party ideology, where the relationship between district-parties and candidates is actually negative.

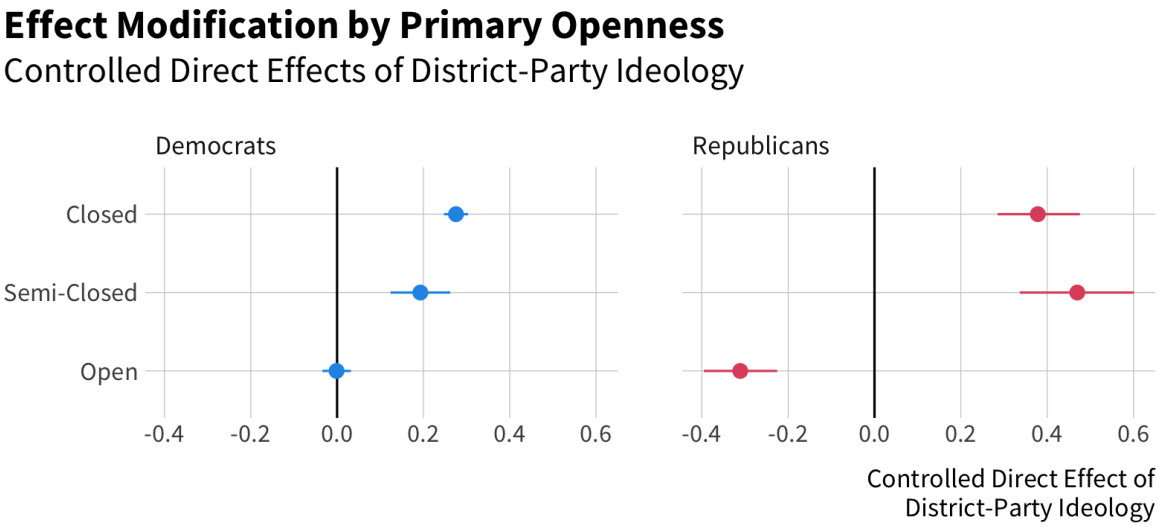


Figure 4.13: Sequential *g* results for candidates in states with closed, semi-closed, and open primary systems. Points and intervals are variational point estimates and 90 percent intervals from the approximate posterior distribution.

It is worth reiterating that these patterns across primary rules should not be interpreted as causal, although it notable that they diverge from the majority of recent research on the effects of primary rules. Furthermore, even though candidates may position themselves in accordance with the strategic environment created by primary institutions, this is no guarantee that primary voters recognize which candidate best fits their ideological preferences conditional on a slate of primary candidates. This is a question I revisit in Chapter 5.

4.4 Discussion

This study contained three notable contributions for the study of the strategic positioning dilemma and primary candidate positioning. First, I employ novel measures of district-party ideology, a concept that is essential to the SPD theory but never operationalized in existing studies to date. Second, I advance the causal credibility of these studies by positing a causal estimand and implementing a statistical approach that estimates it under explicitly stated assumptions. And third, I demonstrate how to bring this method into a Bayesian causal framework, incorporating measurement error in the key independent variable as a prior distribution and deriving a posterior distribution that summarizes uncertainty in all model parameters. I find that primary candidates do position themselves to fit district-party policy preferences, even when controlling for aggregate district voting as an intermediary causal mechanisms. Furthermore, I find that candidate responsiveness is generally greater in states with stricter primary participation rules, a hypothesis that is common among political punditry but doubted among primary elections research.

The contributions of this study provide a structure for improving the research further. Causal diagrams are valuable for understanding how theories of primaries other than the SPD could manifest in the data. Consider, for instance, the notion that candidates position themselves to earn the good favor and resources of policy-demanding groups in the party network. Whether this theoretical perspective jeopardize the causal interpretation of the preceding analysis depends on where policy-demanding are located in a causal diagram. If intermediary groups are an outgrowth of the social, economic, and other demographic bases of the district—so they are descendants of variables x_d—and partisan voters take opinion cues from these groups' activities in the district, then failing to control for intermediary group influence may lead to bias in the estimated effects of district-party ideology.

This study could also be improved by collecting more years of data and setting up a

different research design based on within-unit variation, such as a difference-in-differences panel design. Unfortunately, most routine implementations of difference-in-differences modeling are problematic under data with variation in treatment timing, an issue that econometricians are only just starting to understand (Goodman-Bacon 2018). Panel data approaches to causal inference in political science either generalize the "demediation" routine already employed in this analysis (Blackwell and Glynn 2018) or use fixed-effects models that require particular assumptions how past treatments and outcomes are allowed to affect future treatments and future outcomes (Imai and Kim 2019). Modernizing the primary representation literature for a new generation of causal approaches will not be a simple, one-off endeavor.

Even if researchers stick to the single-stage demediation approach, they could relax modeling assumptions by moving away from the linear modeling laid out by Acharya, Blackwell, and Sen (2016) and implemented in this chapter. Because the identification assumptions, demediation function, and ACDE are nonparametrically defined in terms of expectations about data, the method could be implemented using more flexible estimators of conditional expectations such as matching or machine learning methods.

— 5 —

District-Party Ideology and Primary Election Outcomes

How does district-party ideology affect primary outcomes? The strategic positioning dilemma theory predicts that candidates position themselves as a compromise between the district electorate and the partisan primary electorate. They care about the primary electorate to fend off competition in primaries, either by deterring primary competitors from running at all or by being the "best fitting" candidate to represent partisan constituents in the primary election. I find in Chapter 4 that district-party policy preferences do affect candidate positioning, even after controlling for aggregate partisan voting in the district. This chapter asks if this effort by candidates to position themselves toward their partisan constituency ultimately matters for primary election outcomes. Do more conservative districts nominate more conservative nominees?

This chapter argues that this research question presents problems for statistical modeling and causal inference that come from the same basic limitation of the data: district-party ideology does not vary across candidates in the same district. While it is possible to measure the correlation between district-party ideology and the CF score of the primary winner, this method selects on the dependent variable. We already know from Chapter 4 that candidates are generally more conservative when the electorate is more conservative, so the simple

correlation between district-party ideology and nominee ideology does not capture whether more conservative electorates prefer more conservative candidates, *conditional on the set of available candidates* in the primary. By conditioning on a primary race, however, we also condition on district-party ideology, removing all statistical (and therefore causal) variation in district-party ideology across candidates in the same district. If we want to understand the role of district-party ideology in shaping primary outcomes, we must frame the research question and causal quantity to be identifiable using data on candidates in the same primary.

I confront these statistical and causal challenges using an augmented conditional logit modeling approach. Traditionally, conditional logit is a model that predicts discrete choice based on covariates that vary across *alternatives* within a choice set (candidates), with chooser data (the electorate) held constant within the choice set. This modeling limitation means that chooser features, such as district-party ideology, cannot directly affect candidate choice, but they can indirectly affect candidate choice. I discuss the nature of these indirect effects below, and I devise a statistical model that flexibly estimates a related causal quantity: the causal effect of candidate ideology on primary outcomes, with heterogeneous effects that vary across primary electorates with different district-party ideologies.

Using this modeling approach, I find a noisy effect of candidate ideology on primary outcomes. It appears that primary candidates are less likely to win primary elections when their CF scores are especially centrist or especially extreme, but the estimates are not especially precise. Furthermore, I find no evidence that this effect varies across primary electorates. Although candidates appear to position themselves strategically to fit the particular partisan constituency in their district (Chapter 4), I find no clear evidence that different partisan constituencies reward these strategic maneuvers by candidates.

5.1 Spatial Voting and Candidate Choice

How do the ideal points of candidates and electorates affect primary elections? Spatial voting models argue that primary candidates are more likely to win the nomination when they position their candidacies closer in ideological space to the median primary voter (Aldrich 1983; Downs 1957). This is an essential mechanism underlying the strategic positioning dilemma theory, which states that a candidate must strike a balance between the median partisan voter and the district median voter in order to win both the primary and the general election (Brady, Han, and Pope 2007; Burden 2001). This intuition appears to hold in general elections for U.S. House: candidates who are too progressive or too conservative perform worse than candidates who are "just right" (Canes-Wrone, Brady, and Cogan 2002; Simas 2013). Figure 5.1 plots the key claim of spatial voting models: a candidate is most appealing to a constituency when the candidate's ideological location (represented on a left–right ideological continuum) matches the constituency's preferred ideological outcome. The candidate is less appealing, or provides less *utility* (or "value") to the constituency, when the ideological distance between the candidate and the constituency grows larger. This utility loss occurs whether the candidate is too progressive or too conservative.

One important shortcoming of existing primary elections research is the inability of empirical models to capture this "optimal positioning" in primaries. Studies often measure the relationship between candidate "extremity" and their performance in primary elections—finding that more extreme candidates are more likely to win primary elections (King, Orlando, and Sparks 2016) or that this effect is limited to extreme Republicans (Nielson and Visalvanich 2017)—but extremity is allowed only a constant or monotonic effect on the candidate's primary performance (Hall and Snyder 2015; King, Orlando, and Sparks 2016; Nielson and Visalvanich 2017). Without the possibility of non-monotonicity in the extremity–victory relationship, these empirical models do not reflect their underlying theoretical models. Furthermore,

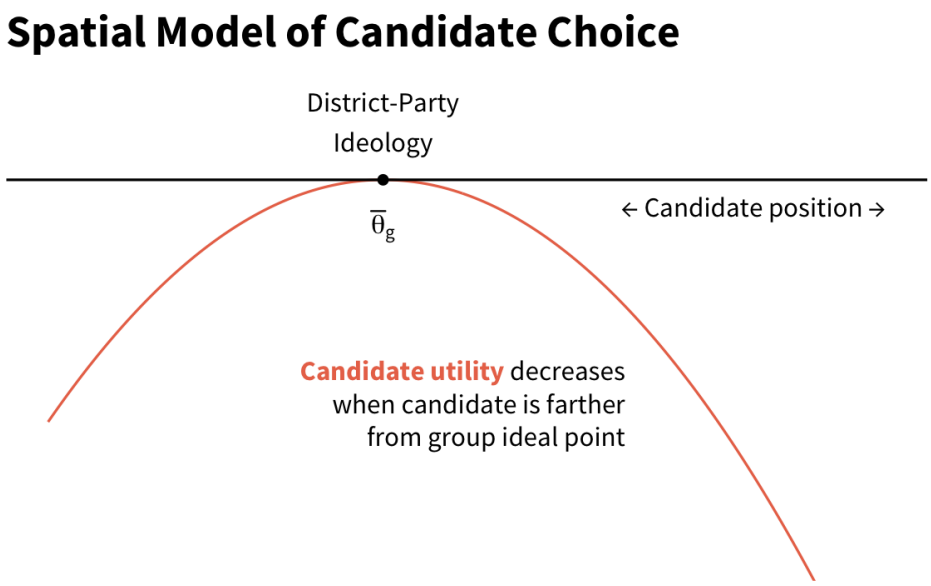


Figure 5.1: A spatial voting model's description of candidate utility (value) as a function of candidate position and district-party ideology. Candidate value is maximized at the group ideal point $\bar{\theta}_g$ and decreases in either direction. The example in this plot assumes quadratic utility loss.

without a measure of the partisan constituency's ideal point (district-party ideology), these studies have no way to know whether the optimal candidate ideology is different in more conservative or more progressive electorates. Because my project measures district-party ideology, I can estimate the optimal candidate ideology in different districts with different partisan constituency ideologies.

Another important limitation of many existing studies is that the factors affecting primary choice cannot be inferred from studying only incumbent members of congress or primary nominees, which is selecting on the dependent variable (e.g. Brady, Han, and Pope 2007; Hirano et al. 2010; Kujala 2020; McGhee et al. 2014). Without somehow accounting for the menu of candidates that a primary electorate can choose from, we cannot infer that whether candidates with certain ideological positions are actually preferred over candidates with other ideological positions. The analysis in this chapter confronts this problem by modeling primary candidate choice using a conditional logit approach, similar to other recent studies

of primary choice (Porter and Treul 2020).

As I discuss in Chapter 1, there are several reasons to doubt that House primary voting predominantly follows a spatial model. Few voters are likely to be aware of candidate positioning in contexts where the party label does not provide differentiating information between candidates (Norrander 1989). Voters do respond to policy differences between candidates if they are made aware of those differences (Lelkes 2019), but learning about the issue positions of House primary candidates is costly. Voters may also be strategic, preferring moderate or incumbent candidates that they believe are most likely to win the general election, even if those candidates are not closest to their ideal points (Simas 2017). Non-ideological traits like incumbency, an "outsider" reputation (Porter and Treul 2020), early fundraising (Bonica 2020), gender (in Democratic races, see Thomsen 2020), or other "valence" features (Nyhuis 2018) may be easier for primary voters to detect and act upon than candidates' ideological stances. It is also possible that candidate ideology's effect on primary elections is mainly a selection function, deterring moderate candidates from entering a primary at all, leaving only the ideologically faithful candidates who present less ideological contrast to voters (Thomsen 2014).

5.1.1 Causal and statistical identifiability

This project is interested in understanding district-party ideology and how it shapes primary elections. An essential constraint of this chapter's analysis is that the "effect of district-party ideology on primary election outcomes" is not a convenient causal quantity to work with. This is because district-party ideology is constant across all primary candidates who compete in a primary contest, therefore it has no direct effect on the probability that any one candidate wins. It can only have indirect effects that interact with other characteristics of the candidates. This section discusses these indirect effects, how they interact with the modeling constraints for primary election data, and how we define causal estimands under these constraints.

Consider a primary race r containing $n_r > 1$ primary candidates, each candidate indexed i. Let $y_r = i$ signify that candidate i wins race r, with the probability that i wins r given by ψ_{ir}. Choice settings such as this, where one chooser must select among several alternatives in the choice set, is traditionally modeled using a conditional logit likelihood (McFadden 1973). Conditional logit has been employed to study candidate choice in U.S. primaries by Ansolabehere, Hirano, and Snyder (2004), Culbert (2015), Simas (2017), and Porter and Treul (2020). Conditional logit supposes that the chooser—in this case, a district-party group in which r takes place—selects a candidate i by comparing the utility they receive from each candidate in the race. Suppose that this utility ω_{ir} contains a systematic component u_{ir} and a stochastic component e_{ir}.

$$\omega_{ir} = u_{ir} + e_{ir} \tag{5.1}$$

The probability that i is chosen is defined as the probability that ω_{ir} is greatest among the alternatives in r. Because the error term e_{ir} is unknown and idiosyncratic to the chooser-choice pairing, conditional logit makes a distributional assumption for the error term and calculates the probability that ω_{ir} is greatest given the distribution of the error term. This probability is calculated as the softmax function of the systematic components,

$$\begin{aligned} p\left(y_r = i\right) &= \psi_{ir} \\ \psi_{ir} &= \frac{\exp\left(u_{ir}\right)}{\sum_{i \in r}^{n_r} \exp\left(u_{ir}\right)} \end{aligned} \tag{5.2}$$

which follows from the assumption that e_{ir} is distributed Gumbel, as in logistic regression. The distinguishing feature of conditional logit is that *chooser* attributes, utility shocks that are specific to the chooser and thus common across all choices, do not have identifiable effects on the resulting choice. This is because the chooser attribute introduces the same utility shock to every u_{ir} term in Equation (5.2).[1] As a result, researchers using conditional logit tend not

[1]This only holds when chooser attributes affect the choice utility *additively*. This project will relax this constraint to insert chooser features in the model.

to include chooser attributes at all in their model of the utility function. They instead model the choice problem as a function of the alternatives only, holding the chooser attributes fixed.

In the case of primary elections, choosers are primary electorates, and district-party ideology is fixed for a given electorate.[2] This means that district-party ideology, $\bar{\theta}_{g[r]}$ for group g in which r takes place, cannot *directly* affect which candidate is nominated by holding all else fixed. This, in fact, is consistent with the spatial model intuition from Figure 5.1: shifting the district-party ideal point $\bar{\theta}_{g[r]}$ left or right affects utility only because it changes the distance between $\bar{\theta}_{g[r]}$ and the candidate location, so the interaction between district-party ideology and candidate location is key. More generally, chooser-level features can be included in conditional logit models as long as there is a cross-level interaction with choice-level data for statistical identification (Fox et al. 2012).[3] Building a statistical model that enables this interactivity is an important contribution of this research design.

This conditional logit model's identifiability constraint matters for causal inference as well, because it affects which causal quantities are feasible to estimate. Consider the potential outcome $\omega_{ir}\left(\mathrm{CF}_{ir}, \bar{\theta}_g\right)$, the candidate utility resulting from a given candidate ideology and district-party ideology.[4] Imagine that we intervene on district-party ideology and measure the average utility effect[5] of setting $\bar{\theta}_g = \theta$ versus some other value θ',

[2]District-party groups are not perfectly synonymous with primary electorates, since some constituents who belong to the district-party do not vote in the primary, and some primary voters may not identify with the party. While this conceptual gap could be explored in future research projects, this project tolerates the inconsistency because the most recent evidence on the representativeness of primary electorates finds that they resemble the demographic profile and policy attitudes of the district-party public (Sides et al. 2020). This analysis contains more years of data and relies on fewer modeling assumptions than analyses that conclude that primary electorates are more polarized than district-parties. (Hill 2015; Jacobson 2012).

[3]I use "interaction" in this context to mean a function of both district-party ideology and candidate location. It does not necessarily imply a multiplicative "interaction term" that is more common generalized linear modeling, multiplicative interactions are a specific example of such a function.

[4]For notational convenience, let g imply $g[r]$.

[5]For the current discussion, we consider the effect on utility instead of the effect on win probability. This is because win probability is complicated by the presence of other candidates, whereas utility is a straightforward function of chooser and choice features. It is important to understand the relationship between the causal model structure and the outcome scale because treatments can have different effects on different scales (VanderWeele 2009).

$\mathbb{E}\left[\omega_{ir}\left(\mathrm{CF}_{ir}, \theta\right) - \omega_{ir}\left(\mathrm{CF}_{ir}, \theta'\right)\right]$. This effect does exist for individual candidates: changing district-party ideology affects the primary electorate's candidate utility by increasing or decreasing the ideological distance between the district-party and the candidate. But because conditional logit model does not provide an easy interface for modeling chooser-level effects, it is impractical to condition on other district-level characteristics to render district-party ideology ignorable.

It is much simpler, instead, to consider the average effect of candidate positioning on candidate utility. Conditioning on candidate features is more straightforward with conditional logit, so causal identification of alternative-level effects is more analytically straightforward as well. The conditional average effect of CF scores on candidate utility would thus be $\mathbb{E}\left[\omega_{ir}\left(\mathrm{CF}, \theta\right) - \omega_{ir}\left(\mathrm{CF}', \theta\right) \mid C_{ir} = c, r\right]$, for a comparison of two values CF and CF′, fixing the district-party ideology at θ and conditioning on other candidate-varying attributes $C_{ir} = c$ and the race r.[6]

Because this project is focused on the added value of my district-party ideology measure, I go one step further to model effect heterogeneity over district-party ideology instead of holding it constant. Because identifying ignorable variation in $\bar{\theta}_g$ is a challenge in conditional logit, I approach this heterogeneity from an effect modification perspective. This means that effect heterogeneity over district-party ideology not reflect the causal effect of district-party ideology. Instead, it reflects only the causal effects of CF scores conditional on a given district-party ideology value. This means rewriting potential outcome as $\omega_{ir}(\mathrm{CF}_{ir})$, removing the causal effect of $\bar{\theta}_g$ from the notation. Formally, we say that district-party ideology is an "indirect modifier" if the CF score effect (CF versus CF′) varies across levels of district-party ideology (θ versus θ') (VanderWeele and Robins 2007), conditional on stratum c and race r. In other words, the conditional average effect of candidate ideology is heterogeneous over

[6]Conditioning on the race, which defines the choice set, is inherent to conditional logit. Conditioning on the choice set is what makes undermines the identifiability chooser-level effects without cross-level interactions.

district-party ideology if the following quantity is not zero:

$$\mathbb{E}\left[\omega_{ir}\left(\mathrm{CF}\right) - \omega_{ir}\left(\mathrm{CF}'\right) \mid \bar{\theta}_g = \theta, c, r\right] - \mathbb{E}\left[\omega_{ir}\left(\mathrm{CF}\right) - \omega_{ir}\left(\mathrm{CF}'\right) \mid \bar{\theta}_g = \theta', c, r\right]. \tag{5.3}$$

Figure 5.2 plots a causal graph of the system under consideration. The causal effect of candidate position CF_{ir} on candidate utility ω_{ir} is unidentified without conditioning on pre-treatment candidate features C_{ir}. District-party ideology is included as an indirect modifier of the CF score effect $\mathrm{CF}_{ir} \rightarrow \omega_{ir}$, represented with the path $\bar{\theta}_g \rightarrow \mathrm{CF}_{ir}$ and no direct path between $\bar{\theta}_g$ and ω_{ir} (VanderWeele and Robins 2007). Because district-party ideology is included as an indirect modifier instead of as a joint treatment, back-door paths that connect district-party ideology and candidate utility through unobserved variables U are allowed to exist without confounding the CF score effect or the effect modification interpretation (VanderWeele 2009). They do confound the causal effects of district-party ideology, however, which effect heterogeneity is not causally attributed to district-party ideology.

5.2 Modeling Causal Heterogeneity with Continuous Interactions

This section describes a statistical model for primary candidate choice that achieves two key objectives. First, the model is designed to capture the heterogeneous causal effect of candidate positioning, conditional on district-party ideology. That is, the model contains appropriate interactions to include chooser-level attributes in the conditional choice model. And second, the model contains the flexibility to capture non-monotonic effects of candidate positioning: utility losses for candidates that position themselves too far from the district-party ideal point. The model detailed below achieves these objectives using two tactics. The first tactic: I model candidate utility using a linear combination of CF scores and district-party ideology. This linear combination projects CF scores and district-party ideology into a common space that can be interpreted like as "ideological distance" metric, allowing candidate utility to increase or decrease as a function of the distance metric. The second tactic: The distance metric's

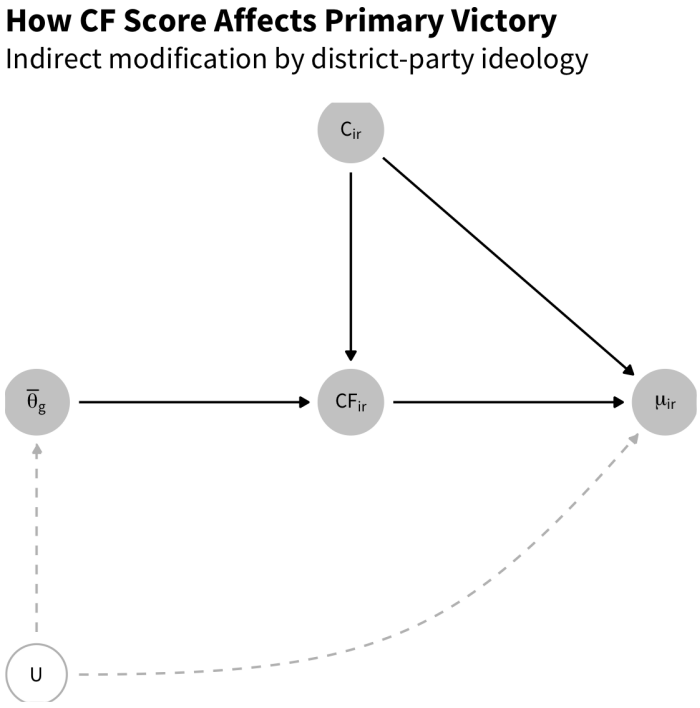


Figure 5.2: Causal Diagram of CF score effect on win probability. District-party ideology is an indirect modifier because has no direct effect on primary outcomes except through candidate proximity. Unobservables U are uncontrolled, so the effect of district-party ideology is not identified. The CF score effect is identified conditioning on C and district-party ideology.

effect on candidate utility is modeled with a spline function. The spline function serves the dual purpose of capturing nonlinearities in candidate utility—an essential component of the spatial voting model—and preserving the interaction between chooser and choice data through those nonlinearities. This strategy enables the effect of candidate positioning on candidate choice to be heterogeneous across candidates with different CF scores and heterogeneous across districts with different district-party ideology values.

The conditional logit model begins by defining the probability that candidate i is chosen in race r as a softmax function of u_{ir}, the systematic component of a candidate's utility

conditional on the choice set.

$$
\begin{aligned}
p\left(y_r = i\right) &= \psi_{ir} \\
\psi_{ir} &= \frac{\exp\left(u_{ir}\right)}{\sum_{i \in r}^{n_r} \exp\left(u_{ir}\right)} \\
u_{ir} &= f\left(\mathrm{CF}_{ir}, \bar{\theta}_{g[r]}\right) + \mathbf{c}_{ir}^{\top}\gamma
\end{aligned}
\tag{5.4}
$$

I use $f()$ to represent a flexible function of candidate i's CF score and the district-party public ideology $\bar{\theta}_{g[r]}$ for group g in which race r is held. I include a vector of candidate-level variables $\mathbf{c}_{ir}$ and regression coefficients γ. Causal inference requires the assumption that conditioning on candidate features renders CF scores ignorable among the candidates in r.

I then construct $f()$ as a flexible spline function of CF scores and district-party ideal points. Although CF scores and district-party ideology both represent ideal points, the two measures are not constructed in the same ideal point space, so calculating the absolute or squared distance between ideal points (e.g. Adams, Bishin, and Dow 2004) isn't immediately available. To rectify this, I create a function that maps these two measures into a common space. Let Δ_{ir} be a linear combination of CF_{ir} and $\bar{\theta}_g$,

$$
\begin{aligned}
\Delta_{ir} &= \alpha \mathrm{CF}_{ir} + \beta \bar{\theta}_{g[r]} \\
\alpha^2 + \beta^2 &= 1
\end{aligned}
\tag{5.5}
$$

which represents an assumption that CF scores and district-party ideology space are affine transformations of one another, similar to the way Aldrich–McKelvey scaling estimates an affine mapping between ideology spaces (Aldrich and McKelvey 1977; Hare et al. 2015). Another way to interpret Δ_{ir} is that the common ideal point space is a weighted average of CF scores and district-party ideology, which weights that are estimated from the data. The second line of (5.5) restricts the coefficients to have a norm of 1, which is an identifiability restriction on the location and scale of the Δ space, which would otherwise be arbitrary. The restriction implies a direct mapping between CF space and $\bar{\theta}_g$ space, since β is defined in

terms of α,

$$\begin{aligned} 1 &= \alpha^2 + \beta^2 \\ \beta &= \pm\sqrt{(1-\alpha^2)} \end{aligned} \tag{5.6}$$

which clarifies how the linear transformation is estimating essentially a scale factor between the two ideal point spaces, parameterized by α only. Because Δ_{ir} is a linear transformation of CF scores and district-party ideology, it has the algebraic interpretation of a "distance measure" of the candidate's CF score and the district-party public ideology in the Δ space. For convenience, I therefore refer to Δ_{ir} as "ideological distance."[7]

I then create a function that lets candidate utility be a nonlinear function of the ideal point distance Δ_{ir}. This captures the spatial voting intuition: shocks to either CF scores or district-party ideology change the ideal point distance Δ_{ir}, which has a nonlinear effect on candidate utility depending on whether the shock moves the ideal point distance toward or away from the optimal distance. I create this nonlinear effect using b-splines. I construct a set of basis functions of Δ_{ir} using a degree-3 polynomial basis with 30 knots across the range of Δ_{ir}.[8] Let $b_k(\Delta_{ir})$ be the k^{th} basis function out of K total, each with a coefficient ϕ_k. The function $f()$ from (5.4) results then in a spline regression on Δ_{ir},

$$\begin{aligned} u_{ir} &= f\left(\text{CF}_{ir}, \bar{\theta}_{g[r]}\right) + \mathbf{c}_{ir}^{\top}\gamma \\ f\left(\text{CF}_{ir}, \bar{\theta}_{g[r]}\right) &= \sum_k b_k\left(\Delta_{ir}\right)\phi_k \end{aligned} \tag{5.7}$$

[7] It is important to note here that my use of "distance" refers more generally to vector spaces than it does to ideal point "differences." The "difference" $(x - z)$ is a special case of the distance $\alpha x + \beta z$ where $\alpha = 1$ and $\beta = -1$. For a linear regression of y on $(\alpha x + \beta z)$, regression predictions for y would be invariant to any nonzero combination of α and β values. So although I refer to Δ_{ir} as an ideal point "distance," it contains the same information as an ideal point "difference" up to an arbitrary rotation of the Δ space (e.g. Armstrong et al. 2014, xv). Restricting the rotation of Δ—for example, by fixing $\beta < 0$—would improve the interpretation of Δ as an ideal point "difference," but it would make Bayesian estimation more difficult by introducing unnecessary boundaries and discontinuities in the posterior distribution over α and β. For ease of estimation, I therefore leave the rotation of the Δ space unrestricted.

[8] The symmetry of the Δ space is ensured by centering CF scores and district-party ideology so that their respective minima and maxima are equidistant from zero. This requires calculating separate Δ spaces for each model, as CF scores and district-party ideology take different values for Republicans and Democrats. It also means that the knot locations vary depending on α and β parameters that define the range of Δ space.

The spline regression enables a continuous interaction effect between CF scores and district-party ideology. The basis functions are nonlinear transformations of district-party ideology, so the chain rule ensures that the derivative of u_{ir} with respect to CF scores (the instantaneous effect of CF scores) is a function that contains district-party ideology $\bar{\theta}_g$.

By specifying the interaction between chooser- and choice-level data in this way, I sidestep the identifiability limitation of a simpler conditional logit model, allowing the causal effect of CF score to vary in different electorates with different district-party ideologies. Interacting two continuous variables through the spline function is much more flexible than a multiplicative interaction term between CF scores and district-party ideology, which would fail to capture both the utility optimum predicted by spatial voting models and any other non-constant interactions. Creating the ideal point distance metric also has a generative interpretation that is superior to the multiplicative interaction, because the intuition of a common ideal point metric is a more faithful representation of spatial voting models. A multiplicative interaction has no comparable generative interpretation.

Although the interpretation of Δ as a common ideal point metric is algebraically sensible, a limitation of the approach is that the model does not very accurately identify which α and β values create more plausible common spaces in terms of posterior probability. This is because the spline regression is flexible enough to create sensible regression functions out of the many configurations of Δ space. In other words, if a particular draw of α and β values "compress" the ideal point space in some way, the spline coefficients are able to "stretch" that space back out in order to fit the data. As a result, the posterior distribution of spline functions is identified from the data even if its component parameters—α, β, and spline weights ϕ_k—are not strongly identifiable on their own. This trade-off between global and local identifiability appears in other flexible modeling approaches such as neural networks (Beck, King, and Zeng 2004; MacKay 1992) and is naturally suited to a Bayesian framework, where unidentified or over-paramaterized models pose no special problem for probabilistic

inference (Jackman 2009, 272). In short, while the model sacrifices some interpretability to fit a flexible regression function, the trade-off is worth the ability to capture nonlinear patterns in spatial voting while avoiding specific assumptions about functional forms or ideal point mappings.

5.2.1 Data

The data for this analysis are drawn primarily from two secondary sources, the Database on Interests, Money in Politics, and Elections (DIME, Bonica 2019b) and the Primary Timing Project (PTP, Boatright, Moscardelli, and Vickrey 2017). Cases are organized at the candidate-contest level, with identifiers for each primary contest indexing political party × congressional district × election cycle. Because primary candidates can run unopposed, I restrict the data to primary races containing at least two candidates. I keep only primary races where the number of winning candidates equals 1, which removes any election where the winner lacked a CF score estimate or where primary outcomes are miscoded in the original data sources. I also drop any primary race where the outcome was decided by a convention instead of an election (coded in the PTP), and I drop all blanket and top-two primaries since those races are not limited to candidates in a single party.

The DIME database contains most of the essential data used for this analysis: CF scores and primary outcome indicators. Primary outcomes for the 2016 election cycle were less thoroughly coded than the primary outcomes for 2012 and 2014 cycles, which led to lots of missing data. Missing primary outcomes in the DIME were supplemented with primary outcome data from the PTP. Matching the same candidacy across databases was not easy using candidate identifiers,[9] so I merge the databases using the probabilistic record-linkage algorithm developed by Enamorado, Fifield, and Imai (2018). I link candidates by name, state,

[9] Candidate IDs in the DIME are regenerated with each vintage of the database, creating inconsistencies in the same candidate's IDs over time. As a result, the DIME identifiers that were initially copied into the PTP do not match the DIME identifiers in more recent DIME vintages.

district number, election cycle, and political party. This process matches 72% of candidacies in 2016 and 62% of candidates in the entire dataset. For candidacies where the DIME and the PTP disagree about the outcome of a primary race, I defer to the PTP because its narrower substantive focus on primary elections lends it more credibility.

Predictive data include dynamic CF scores for every candidate and district-party ideal points from the IRT model in Chapter 2. The conditional logit does not identify district-level shocks to candidate utility because these variables are fixed for all candidates in a primary race, so the choice of controls in $\mathbf{c}_{ir}$ differs sharply from the district Chapter 4. Instead of including district-level demographics, economic indicators, or political background characteristics, $\mathbf{c}_{ir}$ contains candidate-level features that could affect their ideological positioning as well their likelihood of winning the primary. I include an indicator variable for female candidate, which is associated with greater progressivism and a slightly higher primary win probability at least among Democrats (Thomsen 2019, 2020; Thomsen and Swers 2017). I also include an indicator for incumbent candidates, who both have more moderate CF scores (seen in Chapter 4) and are more likely to win their primary reelections. I include no additional indicators for challengers and open-seat candidates, since open-seat races only compare open-seat candidates to one another, and non-incumbency implies challenger status for any race containing an incumbent candidate. The standard control specification includes one last covariate for contribution amount that a candidate gives to themselves, which is logged and then standardized. This control is intended to block a back-door path from CF scores to primary victory through candidate wealth, which could affect both the candidate's ideological position and their win probability.

Although there are additional measures of a candidate's campaign fundraising and spending available in the DIME, I do not use these variables as controls to identify the CF score effect. This is because previous research theorizes that candidate ideology is more likely to influence a candidate's fundraising than vice-versa (Barber, Canes-Wrone, and Thrower

Table 5.1: Number of primary races and primary candidates

Party	Subset	Primary Races	Total Candidates
Democrats	Full data	273	762
Republicans	Full data	342	1,068
Democrats	No incumbents	161	428
Republicans	No incumbents	149	463

2016; Stone and Simas 2010; Thomsen and Swers 2017). The utility model underlying CF scores assumes that this is true *ex ante*, by modeling campaign contributions as a function of ideological affinity. Using the same data for measurement and inference is presents problems that political scientists have been aware of, but political scientists are only just beginning to employ modern causal inference approaches to confront these problems (for an application to text analysis, see Egami et al. 2018).

I estimate separate models for Republicans and Democrats because control variables may confound the treatment effect differently for each party. For instance, gender is thought to have a greater impact in Democratic primaries than in Republican primaries (Thomsen 2019, 2020; Thomsen and Swers 2017). It also may be the case that causal effects vary across party, either because Republicans or Democrats are not equally aware of political ideology or because district-party ideology has different modifying effects for Republicans and Democrats. I also estimate the same model with the sample limited to primary contests with no incumbent present, a practice employed by earlier researchers to sidestep the overwhelming likelihood that incumbents win reelection (e.g. Porter and Treul 2020). Table 5.1 displays the number of candidates and primary contests in each of these subsets of data.

5.2.2 Bayesian modeling, priors, and prior simulation

Like other models featured in this project, the Bayesian setup of this model provides several important benefits. The most important benefit is regularization in the spline function. Al-

though the spline function is beneficial because it can fit many complex functions, complex models always run a risk of overfitting. The trade-off between flexibility and overfitting is especially salient for modeling heterogeneous treatment effects because growing the number of possible comparisons will also grow the number of false positives if no additional methodological adjustments are made. This concern has led researchers to use regularized estimators to detect heterogeneous effects, which introduce bias to shrink heterogeneities toward zero (for example with Bayesian regression trees, Hill 2011; Green and Kern 2012).

I use a hierarchical prior for the spline coefficients to penalize the complexity of the resulting spline function. The prior for each basis function's coefficient ϕ_k has a Normal distribution,

$$\phi_k \sim \text{Normal}\left(0, \eta\right) \tag{5.8}$$

where η is another estimated parameter. By estimating an adaptive prior distribution for the spline coefficients, coefficients are shrunk toward zero through partial pooling. This prior is implemented in Stan as using a non-centered parameterization, which decomposes ϕ_k into a standard Normal variable $\tilde{\phi}_k$ and a scale factor η.

$$\begin{aligned} \phi_k &= \tilde{\phi}_k \eta \\ \tilde{\phi}_k &\sim \text{Normal}\left(0, 1\right) \end{aligned} \tag{5.9}$$

The non-centered parameterization stretches a standard Normal distribution in order to create a Normal distribution with a scale of η. This parameterization is valuable for Bayesian estimation because it de-correlates random variables in the posterior distribution, creating an easier posterior geometry for estimation algorithms. I give the scale factor η a Half-T prior with 3 degrees of freedom and a scale of 1.5,

$$\eta \sim \text{Half-T}\left(\nu = 3, \mu = 0, \sigma = 1.5\right) \tag{5.10}$$

which regularizes the scale value toward zero, but has a modestly flat tail to allow strong signals from the data to depart from the prior. This Normal-T mixture is similar to a "horseshoe prior" (Carvalho, Polson, and Scott 2010; Piironen and Vehtari 2017a, 2017b), which is a popular prior for estimating sparse coefficients with regularization.[10] Unlike the horseshoe, which uses a half-Cauchy scale, the Half-T scale places slightly lower probability on extremely large coefficients but doesn't regularize as strongly as a Half-Normal scale. The left-side panel in Figure 5.3 plots a histogram of simulated draws from this prior, which features a spike near zero and flatter tails than a Normal-Normal mixture.[11]

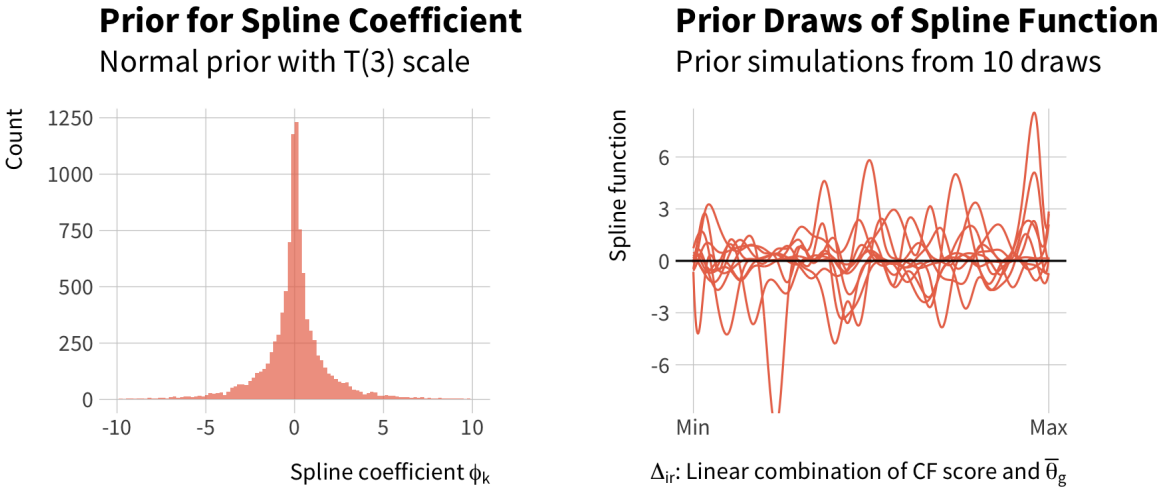


Figure 5.3: Prior draws of spline coefficient and spline function. Left: histogram of prior draws for an individual spline coefficient. Right: draws from the implied prior over spline functions.

The right panel of Figure 5.3 shows 10 prior predictive draws of the spline functions, resulting from 10 coefficient vectors drawn from the hierarchical prior. There are a few important details to note about the construction of this prior. First, most of the "peaks" of

[10] Note that "sparsity" in this context does not imply coefficients of exactly-zero as it does with non-Bayesian L1 regularization (Ratkovic and Tingley 2017b; Tibshirani 1996). Sparse priors may result in posterior *modes* at zero, but posterior intervals will contain non-zero values (Park and Casella 2008).

[11] The tails are long enough that many draws actually fall far outside the region plotted in the figure. These values are much rarer than the values contained in the plotted region, but they are much more probable than they would be under, for example, a Normal-Normal prior.

the spline function are in a neighborhood near zero, especially within the $(-3, 3)$ interval. Although at first this sounds like a very narrow prior, it is important to remember that the spline function is defined on the utility (logit) scale, where small changes in utility can have large and nonlinear effects. For context, a coefficient of 3 on the logit scale would increase the success probability from .5 to 0.95 in a two-candidate choice set, which is a larger effect than almost anything that occurs regularly in elections. Furthermore, a preference for a spline functions near zero is essential for regularization, so this amount of prior information is appropriate for controlling the spline fit. At the same time, there are several peaks that decisively escape the $(-3, 3)$ neighborhood. These larger peaks reflect the flat that the T prior on η has thicker tails that allow larger values to occur more frequently. The shape of the T tail retains enough flexibility to detect a spike in utility even if the center of the prior concentrates spline functions near zero. This plot also shows that 30 knots are more than enough flexibility to capture a utility spike along Δ space.

For the remaining coefficients γ, I specify a weakly informative prior,

$$\gamma \sim \text{Normal}\,(0, 5) \tag{5.11}$$

which rules out explosive coefficient values while still allowing candidate attributes like incumbency to exhibit large correlations with candidate utility. For causal inference, it is important not to regularize confounding effects too much to avoid re-introducing bias into treatment effect estimates (Hahn et al. 2018; Hahn, Murray, and Carvalho 2020).[12]

Because α and β are constrained to have a norm of 1, their values fall on the unit circle. I give these parameters a joint prior that is flat along the unit circle. Stan implements this prior automatically by drawing unnormalized parameters $\tilde{\alpha}$ and $\tilde{\beta}$ from independent standard

[12] For high dimensional problems where regularization cannot be avoided, recent work recommends separate treatment and response models (**???**; Hahn et al. 2018) with a split-sample approach (Ratkovic 2019).

Normal distributions and then dividing by their norm,

$$\begin{aligned} \tilde{\alpha}, \tilde{\beta} &\sim \text{Normal}\,(0, 1) \\ \alpha &= \sqrt{\tilde{\alpha}^2 + \tilde{\beta}^2} \\ \beta &= \sqrt{\tilde{\alpha}^2 + \tilde{\beta}^2} \end{aligned} \tag{5.12}$$

which creates a flat density over the unit circle.[13] The marginal densities for α and β, shown in Figure 5.4 are not exactly flat due to the nonlinear transformation from Cartesian coordinates to polar coordinates.

Prior Draws for Ideal Point Distance Coefficients

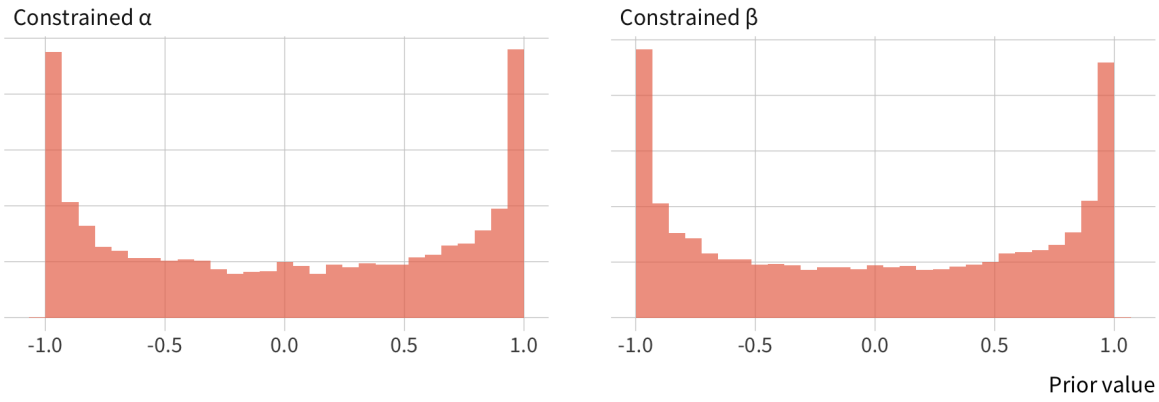


Figure 5.4: Prior draws of coefficients that map CF cores and district-party ideology to the common ideal point distance metric Δ. These priors create a flat prior on unit circle coordinates, even though the marginal priors are not flat.

5.3 Findings

All models were estimated using Stan's full-rank variational inference algorithm, which approximates the posterior distribution as a collection of Normal distributions with a full-rank covariance matrix (Kucukelbir et al. 2015). The main discussion of results focuses on the

[13] Technically, this transformation is undefined if the norm is exactly zero, which realistically never happens.

models estimated using the full datasets. I briefly review the key trends among non-incumbent races in Section 5.3.2.

In order to facilitate the interpretation of the spline model, I first show Figure 5.5, which contains posterior samples the coefficients that map CF scores and district-party ideology into the common ideal point distance measure Δ. Because I identify the latent Δ space by constraining these coefficients to have a norm of 1, all pairs of parameters fall on the unit circle. Points are jittered in the plot to convey which values have greater posterior probability. As mentioned above, many possible ideal point mappings can be rationalized as part of the spline function, so the posterior distribution to does not concentrate very tightly around particular combinations of α and β values. This results in posterior samples that cover all four quadrants of the unit circle. This is not concerning, however, because the common ideal point space is created to facilitate heterogeneous effects, not to be interpreted directly.

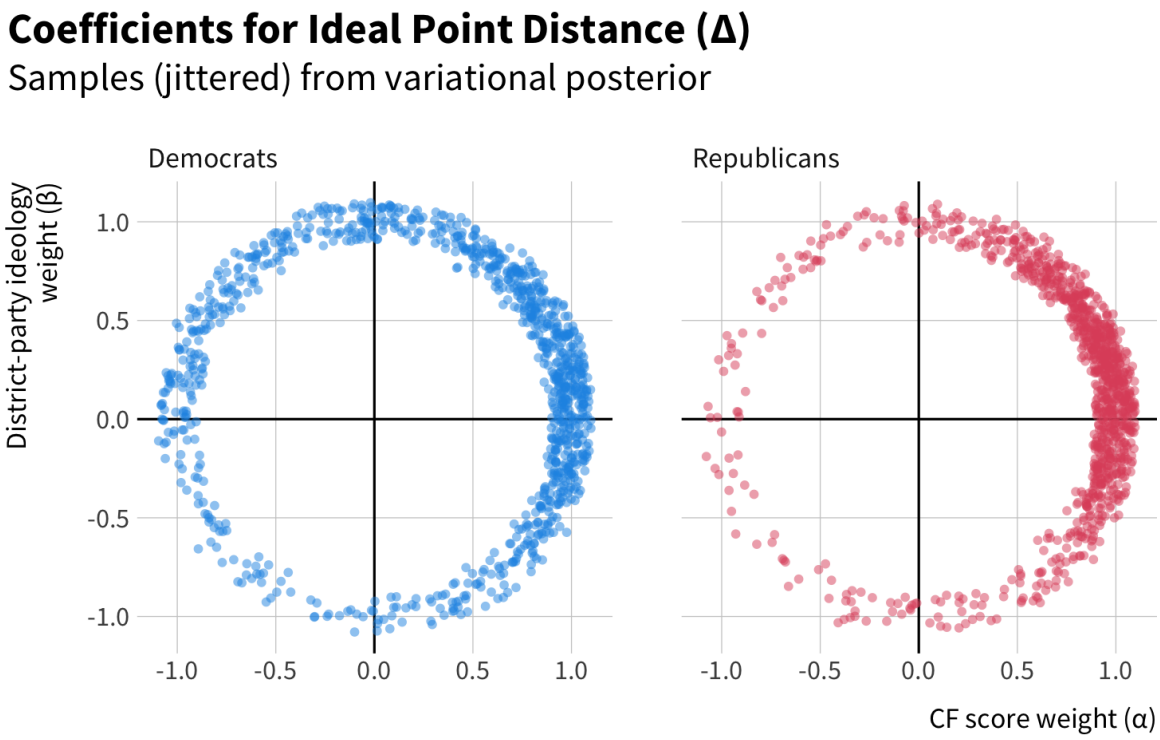


Figure 5.5: Posterior draws of parameters that map CF scores and district-party ideology into Δ space. Points all fall on the unit circle but are slightly jittered to convey posterior density.

Coefficients from the candidate utility model are presented in Figure 5.6. The left panel shows regression coefficients for control variables in $\mathbf{c}_{ir}$: gender, incumbency, and candidate self-fundraising. These coefficients find that gender is positively related to candidate utility among Democrats more than Republicans, a finding that reflects recent evidence by (Thomsen 2019). Unsurprisingly, incumbency has a strong, positive relationship to candidate utility in both parties. Candidate self-fundraising does not strongly relate to candidate utility in either party. This could be because heavier self-funders reflect a mixture of wealthy candidates, who may be advantaged because of their connections to other wealthy funders, and down-on-their luck candidates who rely more heavily on self-fundraising to make up for meager fundraising receipts elsewhere. The right panel shows all spline basis function coefficients and the scale parameter in the smoothing prior for the spline coefficients. Most spline coefficients have posterior point estimates near zero, which is the intended result of the regularizing prior on the coefficients. A few coefficients do depart from the prior, an initial indication that the spline regression detects a smooth function with a small number of "wiggles" rather than a highly variable function with many local peaks and troughs.

The key finding from the conditional logit model is the spline function plotted in Figure 5.7. The spline is a function of the common ideal point metric Δ, which means the spline is a function of both CF scores and district-party ideology. This means that the shape of the spline function comes two signals in the data. First, which Δ values are related to candidate utility, and second, what combination of CF scores and district-party ideology (in terms of α and β values) more strongly affect CF scores. I show candidate CF scores along the horizontal axis, and spline functions holding district-party ideology fixed at different values are plotted on the vertical axis. Solid lines show the spline function conditioned on *average* district-party ideology in each party, while dashed and dotted lines condition the spline function on district-party ideology values one standard deviation above and below the mean. The shaded region shows the 90% posterior interval for the spline function conditioning on

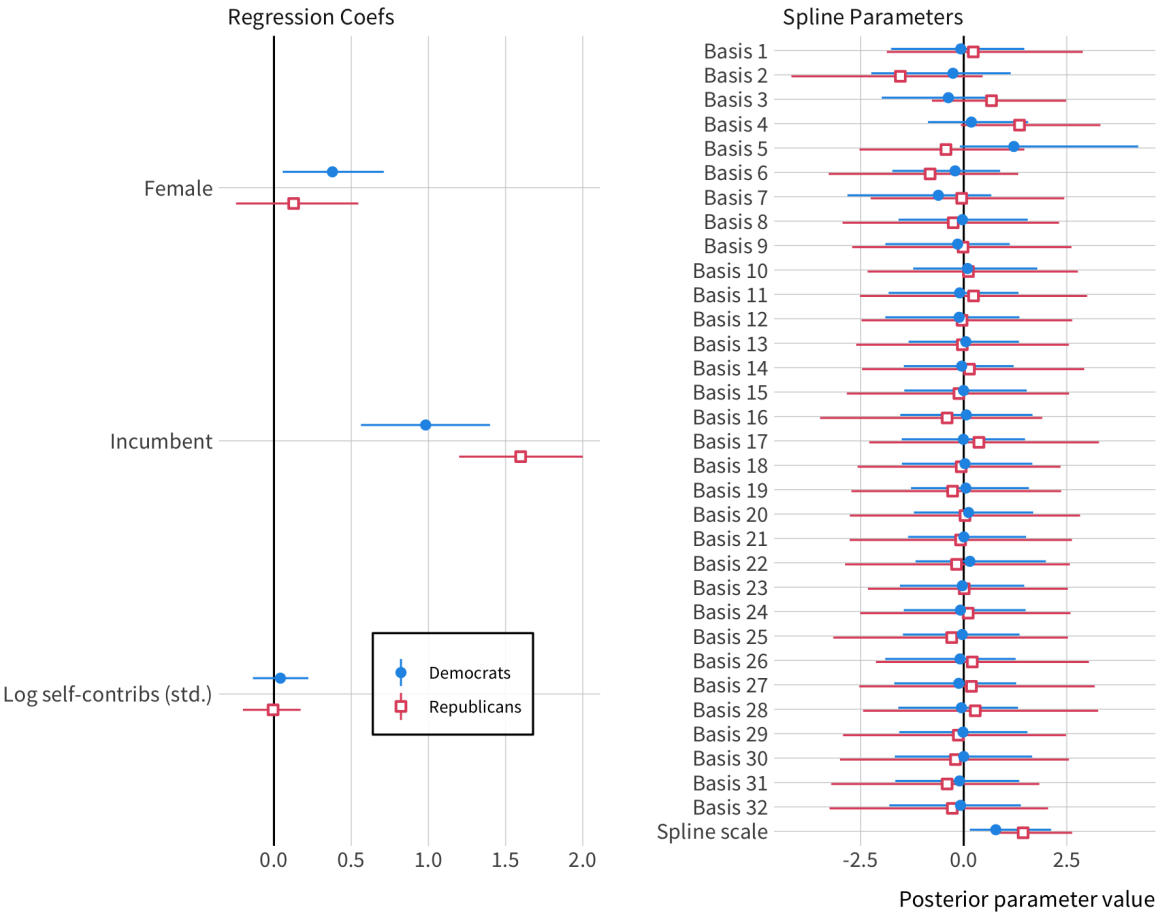


Figure 5.6: Posterior parameters from conditional logit. Points and intervals are variational point estimates and 90 percent quantile intervals from approximate posterior. Left panel shows regression weights for covariates. Right panel shows basis function coefficients and hierarchical scale parameter.

the average district-party ideology, calculated from samples from the variational posterior. A spline function that drifts up and down across the range of CF scores reflects the causal effect of CF scores under the identification assumptions. A spline function that varies for different values of district-party ideology reflects heterogeneous causal effects in districts with different district-party ideologies.

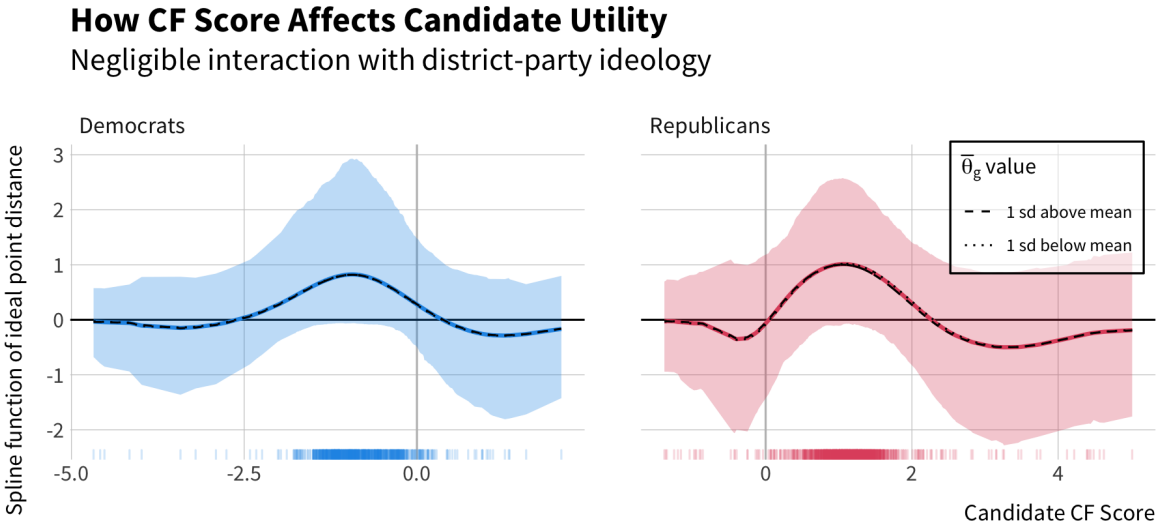


Figure 5.7: CF score effect on candidate utility. Spline function of CF score with district-party ideology held at mean, mean minus one standard deviation, and mean plus one standard deviation.

The estimates for both Democrats and Republicans show that CF scores most strongly affect candidate utility at CF score values near the median partisan candidate, evidenced by the peak in the spline function near the center of each party's CF score distribution. This means that candidates at the ideological center of their party are most valuable to primary electorates, while both centrist and extreme candidates are less valuable. This signal is not overwhelmingly strong—the 90% credible interval covers zero—but the posterior distribution of spline functions generally supports this interpretation that primary voters prefer "representative" candidates. This peak in utility is what we should expect to find under an assumption of spatial utility voting: candidates are most likely to win when they

take an optimal ideological stance, and less likely to win when they are more liberal or more conservative than the optimal stance. This pattern suggests that a null hypothesis of non-ideological voting in primaries is unlikely to be true, because the utility function isn't simply flat over candidate CF scores. The location of the utility peak suggests that primary electorates are not focused by-and-large on electability, which would manifest as a utility peak at centrist CF score values. It also suggests that primary electorates do not place a single-minded emphasis on partisan extremism, which would result in a utility peak at more extreme CF scores. It is also important to note that these estimates control for incumbency, so the utility peak at party-median CF score values *does not* merely reflect the fact that high-value incumbent candidates represent the ideological "core" of the party. Furthermore, results below suggest that a similar utility peak appears even when we remove all non-incumbents from the data.

Although the analysis suggests that candidate ideology does have a causal effect on CF scores, there is no evidence to suggest that this effect is heterogeneous across districts. The utility function for districts with average district-party conservatism is almost identical to the utility functions for districts with above-average and below-average district-party conservatism, shown in the figure with dotted and dashed lines. More conservative candidates are not noticeably more valuable to conservative primary electorates, nor are progressive candidates more valuable to progressive electorates.

How do these results square with the strategic positioning dilemma? The SPD theorizes that candidates position themselves close to the primary constituency in order to defeat possible or actual primary challengers. While candidates may be able to position themselves to target the ideological preferences of their partisan constituents (shown in Chapter 4), these positioning maneuvers are ultimately ineffectual if primary voters do not perceive or act on candidates' ideological stances. While the findings in this do suggest that primary candidates are more highly valued when they position their candidacies near the ideological

"core" of their party, this benefit is very broad. Primary electorates' place less value on very extreme and very moderate candidates than they place on "typical" partisan candidates, but the current data do not support finer conclusions than that. I find no evidence that partisan constituencies with more consistent ideological preferences prefer more consistently ideological candidates, which suggests that primary candidates' ideological maneuvers may have little direct appeal to primary electorates. Either primary constituencies do not perceive fine differences between candidates' ideological positioning, or they do not place much weight on candidate ideology when choosing a candidate.

On a theoretical level, these findings do not necessarily imply that candidate ideological maneuvering is useless to their primary successes. Candidate ideology's effect on primary success can be mediated by other mechanisms. For instance, candidate ideology can affect campaign fundraising (Barber, Canes-Wrone, and Thrower 2016; Stone and Simas 2010; Thomsen and Swers 2017), which could indirectly affect voters' awareness of candidates. Indirect mechanisms such as these are possible, but their reliance on intermediary actors rather is a point of divergence from the strategic positioning dilemma, which construes candidate ideology as a more direct appeal to voters' policy preferences.

On a technical level, a significant benefit of the Bayesian approach is that uncertainty intervals are straightforward to calculate using posterior samples. Non-Bayesian uncertainty estimates would require the researcher to derive a variance estimator analytically or estimate a large number of models on bootstrapped data, both of which are more demanding than generating post-estimation quantities within a Stan program.

5.3.1 Causal effects on primary outcomes

So far, I have discussed model results in terms of candidate utility, a latent scale that does not directly manifest as a primary win or loss. In order to understand how candidate ideology affects primary outcomes, utility must be translated to win probability. This discussion

provides a brief mathematical roadmap to describe causal effects on utility and win probability.

First, I discuss causal effects on candidate utility. The average effect of a change in CF score on candidate utility can't be summarized by a single coefficient because of the nonlinear, interactive form of the spline function $f\left(\text{CF}_{ir}, \bar{\theta}_g\right)$. The causal effect must instead be calculated as the difference in function values at different CF score inputs. Let $\text{CUE}\left(\text{CF}, \text{CF}', \theta, c\right)$ be the conditional utility effect (CUE) of moving to CF from a reference value CF′, which is defined as

$$\begin{aligned}\text{CUE}\left(\text{CF}, \text{CF}', \theta, c\right) = \\ \mathbb{E}\left[f\left(\text{CF}, \bar{\theta}_g\right) \mid \bar{\theta}_g = \theta, C_{ir} = c\right] - \mathbb{E}\left[f\left(\text{CF}', \bar{\theta}_g\right) \mid \bar{\theta}_g = \theta, C_{ir} = c\right]\end{aligned} \tag{5.13}$$

given a fixed value of $\bar{\theta}_g = \theta$ and conditioning on covariates C_{ir}. The Bayesian approach defines a the probability distribution for $\text{CUE}\left(\text{CF}, \text{CF}', \theta, c\right)$, which marginalizes over model parameters:

$$\begin{aligned}p\left(\text{CUE}\left(\text{CF}, \text{CF}', \theta, c\right) \mid \mathbf{y}\right) = \\ \int p\left(f\left(\text{CF}, \theta\right) - f\left(\text{CF}', \theta\right), \alpha, \beta, \boldsymbol{\phi} \mid \theta, c, \mathbf{y}\right) \, \mathrm{d}\alpha \, \beta \, \boldsymbol{\phi}\end{aligned} \tag{5.14}$$

The conditional average effect of CF score on candidate utility can be gleaned by simply contrasting the function values in Figure 5.7 for two different CF score values. I do not plot these causal contrasts directly because they would look identical to the spline function, except the function would be intercept-shifted up or down depending on the reference value CF′.

Instead, I highlight how changes in CF score affect primary win probability. Two intervening factors affect how candidate utility becomes win probability. First, win probability is a nonlinear function of candidate utility, so a utility shock can have different effects on win probability depending on the baseline utility at the reference value CF′. Second, a primary race can feature a variable number of candidates, so a utility shock will have larger effects on win probability when the field of candidates is small, and smaller effects on win probability when the field of candidates is larger. Let $\text{CWE}\left(\text{CF}, \text{CF}', \theta, c, r\right)$ be the conditional *win* effect

(CWE) of moving to CF from a reference value CF′ in race r. The CWE is a function of the race r because each race could have different numbers of candidates with different utilities. Because win probability is modeled as a softmax function of candidate utility, the CWE is the expected difference in softmax functions when CF score takes the value CF versus the reference value CF'.

$$\begin{aligned}&\text{CWE}\left(\text{CF}, \text{CF}', \theta, c, r\right) = \\ &\qquad \mathbb{E}\left[\text{softmax}\left(u_{ir}(CF)\right) \mid \theta, c, r\right] - \mathbb{E}\left[\text{softmax}\left(u_{ir}(CF')\right) \mid \theta, c, r\right]\end{aligned} \tag{5.15}$$

where $u_{ir}\left(CF_{ir}\right)$ represents a candidate utility, holding all variables besides CF score fixed. In turn, the probability distribution for the CWE marginalizes over the parameters that compose the softmax function.

$$\begin{aligned}&p\left(\text{CWE}\left(\text{CF}, \text{CF}', \theta, c, r\right) \mid \mathbf{y}\right) = \\ &\qquad \int p\left(\text{softmax}\left(u_{ir}(CF)\right) - \text{softmax}\left(u_{ir}(CF')\right) \mid \theta, c, r, \mathbf{y}\right)\, \mathrm{d}\alpha\, \beta\, \boldsymbol{\phi}\, \gamma\end{aligned} \tag{5.16}$$

Figure 5.8 visualizes causal effects of CF score on win probability in a two-candidate race. I use the average CF score in each party as the reference value for CF score. Therefore the figure plots the how a candidate's ideological position affects their win probability *compared to holding the average ideological position* for a candidate in their party. I show effects only in races with two candidates total, which conveys an upper bound on the magnitude of causal effects—effects in more crowded primary fields are expected to be smaller. The left-side panels contain causal effects in a two-candidate race with no incumbent. The right-side panel shows causal effects in a two-candidate race when the other candidate is the incumbent.

The first thing to notice is that causal effects are generally negative. This is because the average CF score in each party is very close (but not equal) to the optimal CF score as estimated by the model. As candidates take ideological stances that are more moderate or more extreme than the optimal stance, their win probability generally decreases in all subsets of data. These declines are not monotonically decreasing because the spline function detects

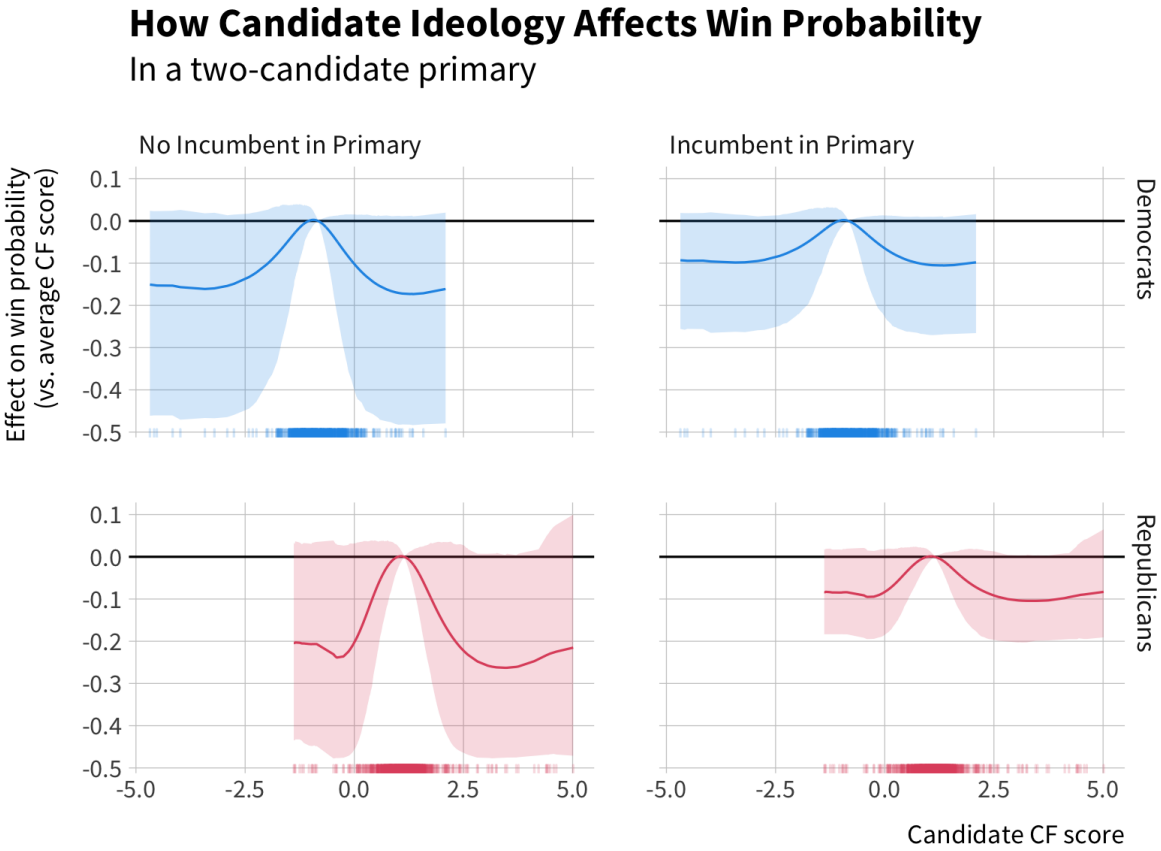


Figure 5.8: CF score effect on primary win probability in a three candidate race. Left panels feature a three-candidate race with no incumbents. Right panels feature a three-candidate race with one incumbent.

noisy effects for CF score with fewer observations. Although the 90% credible intervals contain generally contain effects of zero, the posterior distribution suggests that negative causal effects are most probable.

Another important observation from Figure 5.8 is the way the presence of an incumbent dampens causal effects. Because incumbents are already highly valued in primaries, shocks to challenger utility compose a smaller share of the total utility across all candidates. In terms of win probability, because challengers are already unlikely to win a primary, a challenger with an ill-fitting ideological position can only hurt their chances so much before they confront the zero-bound on their win probability.

Bayesian computation is again valuable for plotting these causal effects with uncertainty. Because each of these effects was calculated using draws from the variational posterior distribution, creating 90% uncertainty intervals from the inner 90% of samples is straightforward.

5.3.2 Races with no incumbents

Because incumbent candidates are likely to win a primary re-nomination, some scholars estimate models of primary candidate choice using only open-seat primaries with no incumbents (e.g. Porter and Treul 2020). I conduct this same analysis by dropping all races that feature an incumbent candidate, and I plot the spline function results in Figure 5.9. I find essentially the same pattern as with the full data: candidates with moderate and extreme ideological stances are less highly values by primary electorates, and the average pattern has no noticeable heterogeneity across districts with different district-party ideologies.

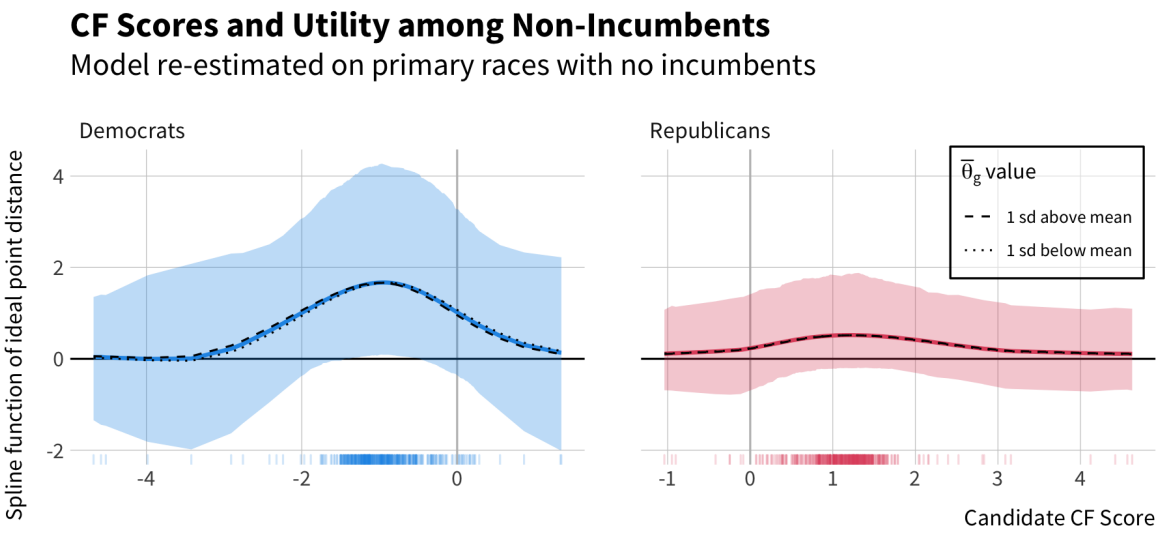


Figure 5.9: CF score effect on candidate utility in races with no incumbent. Spline function of CF score with district-party ideology held at mean, mean minus one standard deviation, and mean plus one standard deviation.

5.4 Discussion: Causal Identification with CF Scores

Causal inference in this analysis depends on a key identifying assumption that candidate CF score values are ignorable, conditional on the observed covariates. Using CF scores as a treatment variable under ignorability could be problematic because CF scores are functions of a measurement process rather than being direct measures of candidate ideology. Because CF scores are affected by patterns of campaign contributions in the parties' campaign finance networks, these factors may play an indirect role in confounding the effect of CF scores.

In this analysis, I find that candidates with "party-typical" CF scores have optimal spatial utility. It could be the case that candidates only receive a party-typical CF score because of their central location in a campaign finance network, which indirectly captures a consensus view that donors find the candidate to be viable. If this is the case, the "viability" of a candidate (as judged by financial contributors) affects their CF score, and this vague sense of viability could confound the relationship between CF scores and primary victory. The use of proxy measures for modern causal inference isn't well understood in political science, and attempts to confront the issue are only just emerging (Egami et al. 2018). But because political science relies heavily on proxy measures in all subfields of quantitative research, learning to conduct causal inference with simultaneous measurement and inferential modeling will be essential for credible causal research in the future.

The fact that party-typical CF scores are most valuable to primary electorates also raises a question of feedback over time. If extreme candidates were the candidates that were most valued in primaries, candidates could perceive this pattern, and over time candidates would re-equilibrate their campaign positions to maximize their win probability. This implies that the optimality of a campaign position causes it to be party-typical, not the other way around. In other words, we may never observe a reality where the optimal campaign stance is not also party-typical, because candidates adjust their ideologies to maximize their benefit

in equilibrium, the same way that a seller of goods adjusts the pricing of their product to meet demand. Political scientists have studied the evolving stances of the parties over time (Carmines and Stimson 1986; Karol 2009; Rubin 2017; Schlozman 2015), but these studies usually arise in the context of interparty polarization and interest group coalitions rather than the context of primary electioneering.

5.5 Closing remarks on the project

How does district-party ideology affect primary outcomes? This question is complicated by statistical and causal identifiability problems. Primary elections are choices between alternative candidates in the same district, so district-party ideology cannot directly affect candidate choice without some interaction with candidate-level variables. I develop a statistical model that addresses statistical and causal identifiability problems, estimating the causal effect of candidate CF scores on candidate primary success with heterogeneous effects across district-parties with different policy ideologies. Under the right causal identification assumptions, I find that candidate ideology does affect primary outcomes: deviating from an optimal ideological position reduces a candidate's chance of winning, penalizing both moderate and extremist candidates. I find no evidence that this effect varies across district-parties, even though district-parties vary in their policy ideologies. This could be because voters do not perceive or do not give much weight to the fine differences between candidates' ideological positioning. It could also be because candidates' ideological stances have diffuse rather than specific effects on their success, increasing their credibility with campaign financiers or other party gatekeepers who can advance their careers, even if voters don't respond to these specific signals as strongly.

This project examined district-party ideology, candidate positioning and primary outcomes as a study of the "strategic positioning dilemma" theory of primary competition.

Chapter 1 justifies the creation of ideal point measures for the partisan preferences of partisan voters in a congressional district, a key independent variable in the theory. I build the scores in Chapter 2 and outline the Bayesian approach to causal inference in Chapter 3, both of which I employ in empirical analyses. Chapter 4 studies the effect of district-party ideology primary candidates' campaign positions. Consistent with the strategic positioning dilemma, I find that district-party ideology *does* affect campaign positions: candidates run as more conservative candidates in conservative district-parties, and vice versa in more progressive district-parties. Importantly, this relationship holds after controlling for district-level voting, suggesting that candidates' responsiveness to their partisan bases doesn't simply reflect the competitiveness of the general election. This chapter studied whether candidates' attempts to target their partisan bases pay off in primary elections. While I do find that primary candidates are more successful in primary elections when they position themselves optimally, I do not find that the optimal campaign position is a narrow enough region of ideological space to have a large effect on primary outcomes. Furthermore, I find no evidence that optimal campaign position is different from one district to the next, suggesting that candidates' efforts to position themselves toward the partisan base is appreciated by voters only in the broadest terms. Primary electorates do not prefer candidates to be too moderate or too extreme, but I find no evidence that partisan ideological variation from one district to the next has a noticeable effect on which candidate ultimately wins the primary.

These results lend partial support to the strategic positioning dilemma. Elite actors in the story—primary candidates—do behave in accordance with theoretical expectations. The mass public's role in the story—judging which candidates best appease their appetites for ideological policy promises—is not supported by the analysis. To be clear, I do not find contradictory evidence either. Contradictory evidence would consist of electorates that tend to prefer moderate candidates, or electorates that have no ideological awareness whatsoever. Instead, I find that primary electorates are "broadly" ideological in the sense

that they nominate candidates who represent the general consensus in the party, with no obvious variation in that pattern across districts. This combination of strategic behavior among elites with satisficing behavior among the public fits a prevailing current in modern US politics research: the mass public knows enough about their *general* ideological tastes to make competent political decisions, but they are not so attuned to politics that they behave as hyper-informed political junkies.

This project has also exemplified other important contributions to contemporary Bayesian analysis and causal inference. The empirical studies in Chapters 4 and 5 demonstrate the use causal graphs to clarify causal assumptions and motivate the choice of causal estimands. The Bayesian approach to causal inference in Chapter 3 justified the use of Bayesian modeling for causal inference, which enabled flexible model-building, pragmatic uses of prior information for model stability and regularization, and principled accounting of statistical uncertainty throughout the causal analyses.

Appendices

Group IRT Model

Table 1: List of policy items used for ideal point estimation and data sources containing these items.

Tag	Text (short)	Source(s)
abort_20wk.ban	20 week ban	CCES 2016, CCES 2014
abort_all.illegal	never allow abortion	CCES 2016
abort_always	always allow abortion	CCES 2016, CCES 2014
abort_employer.insurance	allow employers to decline abortion coverage in insurance	CCES 2016, CCES 2014
abort_hyde	prohibit federal funds for abortion	CCES 2016, CCES 2014
abort_legal	spectrum of abortion legality (fix)	CCES 2012
budget_aid.poor	more/less spending on aid to poor	ANES 2016
budget_child.care	more/less spending on child care	ANES 2016
budget_ed.spend	more federal spending on education	ANES 2016
budget_sci.tech	spending on science on science and tech	ANES 2016
budget_welfare	more/less spending on welfare	ANES 2016
budget_where.balance	priorities for where to balance the budget	CCES 2012

Table 1: *(continued)*

Tag	Text (short)	Source(s)
env_climate.action	should govt take action on climate/warming	CCES 2012, ANES 2016
env_clean.acts	stronger clean air/water acts even if cost jobs	CCES 2016
env_epa.carbon	EPA powers to regulate CO2	CCES 2016
env_fuel.efficiency	raise fuel efficiency	CCES 2016
env_require.renewables	require some minimum of renewables in electricity generation even if cost jobs	CCES 2016
defense_spending	increase or decrease social spending (7pt, 4 mid)	ANES 2016
econ_env.jobs	protect environment or protect jobs (scale)	CCES 2012
econ_guar.jobs	guaranteed jobs/std of living	ANES 2016
econ_min.wage	raise min wage (to $12?)	CCES 2018, ANES 2016
econ_reg.banks	do more to regulate banks	ANES 2016
econ_services	spend more or less on services? (7pt, 4mid)	ANES 2016
econ_tax.wealthy	tax incomes > $1mil to pay for roads, schools	CCES 2018, ANES 2016
econ_inequality	govt action to reduce inequality	ANES 2016
env_spend.more	spend more/less on environment	ANES 2016
gender_equal.pay	require equal pay for men and women	ANES 2016
gun_assault.ban	ban assault weapons	CCES 2016, CCES 2014
gun_back.check	background checks for gun sales	CCES 2016, CCES 2014, CCES 2018
gun_big.mags	ban high capacity magazines	CCES 2014

Table 1: *(continued)*

Tag	Text (short)	Source(s)
gun_easy.cc	make it easier to get a CCW permit	CCES 2016, CCES 2014
gun_publish.registry	should governments be banned/allowed to publish registries	CCES 2016, CCES 2014
gun_stricter	laws covering firearms sales should be stricter/less	CCES 2012, ANES 2016
hc_govt.plan	healthcare: medicare for all/govt insurance	CCES 2018, ANES 2016
hc_govt.spending	increase spending to pay for HC	ANES 2016
hc_repeal.aca	healthcare: repeal ACA	CCES 2018
imm_birth.cit	end birthright citizenship for children of undocumented	CCES 2012, ANES 2016
imm_border.patrol	increase border patrols	CCES 2016, CCES 2012
imm_deport.undoc	ID and deport undocumented	CCES 2016, CCES 2014
imm_dreamers	legal status to children brought here illegally who graduate high school	CCES 2016
imm_fine.businesses	fine firms that hire undocumented imms	CCES 2016, CCES 2012, CCES 2014
imm_imprison.repeats	imprison anybody already once deported	CCES 2018
imm_legal.status	legal status to job-holding, tax-paying (3+ years) no felonies	CCES 2016, CCES 2012, CCES 2014
imm_less.legal	reduce legal immigration	CCES 2018
imm_more.visas	increase overseas worker visas	CCES 2016
imm_police.question	allow police to question anyone suspected of being undocumented	CCES 2012
imm_public.services	bad undocumented from ER or public school services	CCES 2012

Table 1: *(continued)*

Tag	Text (short)	Source(s)
imm_withhold.funds	withhold money from local PDs that don't report undocumented	CCES 2018
law_body.cams	require body cams always	CCES 2016
law_mand.min	eliminate mandatory minimums	CCES 2016
law_weed	should marijuana be legal	ANES 2016
law_more.cops	increase police on street by 10% even at expense of other services	CCES 2016
law_recid.sentences	increase sentences for recidivism	CCES 2016
law_death.pen	favor/oppose death penalty	ANES 2016
lgbtq_marriage	favor/oppose same sex marriage	CCES 2016, CCES 2012, ANES 2016
lgbtq_disc.laws	job discrimination laws	ANES 2016
lgbtq_adoption	allow adoption	ANES 2016
race_aff.act	support/oppose Aff. act.	CCES 2012, ANES 2016
race_aid.blacks	govt aid to blacks	ANES 2016
race_gen.disc	Racial resentment: generations of discrimination	CCES 2012
race_irish.italian	Racial resentment: Irish/Italian	CCES 2012

Colophon

This project is open source and managed with Git. A remote copy of the repository is available at https://github.com/mikedecr/dissertation. Currently, the repository is at the following commit:

```
## Commit:  36213ddb7bf96afaca691842c673395851d78501
## Author:  Michael DeCrescenzo <mgdecrescenzo@gmail.com>
## When:    2020-09-25 18:50:46 GMT
##
##      refine meta graphs, begin writing
##
## 4 files changed, 280 insertions, 31 deletions
## 30_causality.Rmd         | - 4 +153  in 3 hunks
## _book/MGD-thesis.pdf     | - 0 +  0  in 0 hunk (binary file)
## code/03-causality/meta.R | -27 +125  in 6 hunks
## notes/agenda.rmd         | - 0 +  2  in 1 hunk
```

This version of the document was generated on 2020-09-25 18:30:38.

All Bayesian models were estimated using the probabilistic programming language Stan. Front-end interface to Stan and other data management was performed with R. The document

was managed with the bookdown package for R, built to PDF using LaTeX.

References

Abadie, Alberto et al. 2020. "Sampling-based versus design-based uncertainty in regression analysis." *Econometrica* 88(1): 265–296.

Abramowitz, Alan I, and Kyle L Saunders. 1998. "Ideological realignment in the us electorate." *The Journal of Politics* 60(03): 634–652.

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review* 110(3): 512–529.

Adams, James, Benjamin G Bishin, and Jay K Dow. 2004. "Representation in congressional campaigns: Evidence for discounting/directional voting in us senate elections." *Journal of Politics* 66(2): 348–373.

Adolph, Christopher et al. 2003. "A consensus on second-stage analyses in ecological inference models." *Political Analysis* 11(1): 86–94.

Ahler, Douglas J, Jack Citrin, and Gabriel S Lenz. 2016. "Do open primaries improve representation? An experimental test of california's 2012 top-two primary." *Legislative Studies Quarterly* 41(2): 237–268.

Akinc, Deniz, and Martina Vandebroek. 2018. "Bayesian estimation of mixed logit models: Selecting an appropriate prior for the covariance matrix." *Journal of choice modelling* 29: 133–151.

Aldrich, John H. 1983. "A downsian spatial model with party activism." *American Political Science Review* 77(04): 974–990.

Aldrich, John H. 2011. *Why parties?: A second look*. University of Chicago Press.

Aldrich, John H, and Richard D McKelvey. 1977. "A method of scaling with applications to the 1968 and 1972 presidential elections." *The American Political Science Review* 71(1): 111–130.

Alvarez, Ignacio, Jarad Niemi, and Matt Simpson. 2014. "Bayesian inference for a covariance matrix." *arXiv preprint arXiv:1408.4050*.

American Political Science Association, Committee on Political Parties. 1950. *Toward a more responsible two-party system*. Johnson Reprint Company.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Ansolabehere, Stephen et al. 2010. "More democracy: The direct primary and competition in us elections." *Studies in American Political Development* 24(02): 190–205.

Ansolabehere, Stephen, Shigeo Hirano, and James Snyder. 2004. "What did the direct primary do to party loyalty in congress?" In *Process, party and policy making: Further new perspectives on the history of congress*, Stanford University Press.

Ansolabehere, Stephen, Jonathan Rodden, and James M Jr Snyder. 2008. "The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting." *American Political Science Review*: 215–232.

Ansolabehere, Stephen, James M Snyder, and Charles Stewart. 2001. "Candidate positioning in U.S. house elections." *American Journal of Political Science*: 136–159.

Armstrong, David A et al. 2014. *Analyzing spatial models of choice and judgment with r*. CRC Press.

Aronow, Peter M, and Benjamin T Miller. 2019. *Foundations of agnostic statistics*. Cambridge University Press.

Athey, Susan, Guido W Imbens, and Stefan Wager. 2016. "Approximate residual balancing: De-biased inference of average treatment effects in high dimensions." *arXiv preprint arXiv:1604.07125*.

Bailey, Michael. 2001. "Ideal point estimation with a small number of votes: A random-effects approach." *Political Analysis*: 192–210.

Baldi, Pierre, and Babak Shahbaba. 2019. "Bayesian causality." *The American Statistician*: 1–9.

Barber, Michael J. 2016. "Ideological donors, contribution limits, and the polarization of american legislatures." *The Journal of Politics* 78(1): 296–310.

Barber, Michael J, Brandice Canes-Wrone, and Sharece Thrower. 2016. "Ideologically sophisticated donors: Which candidates do individual contributors finance?" *American Journal of Political Science*.

Barber, Michael, and Jeremy C Pope. 2019. "Does party trump ideology? Disentangling party and ideology in america." *American Political Science Review*: 1–17.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data." *Political analysis* 23(1): 76–91.

Barnard, John, Robert McCulloch, and Xiao-Li Meng. 2000. "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica*: 1281–1311.

Bartels, Larry M. 2000. "Partisanship and voting behavior, 1952-1996." *American Journal of Political Science*: 35–50.

Bartels, Larry M. 2009. *Unequal democracy: The political economy of the new gilded age*. Princeton University Press.

Beck, Nathaniel, Gary King, and Langche Zeng. 2004. "Theory and evidence in international conflict: A response to de marchi, gelpi, and grynaviski." *American Political Science Review*: 379–389.

Berger, James. 2006. "The case for objective bayesian analysis." *Bayesian analysis* 1(3): 385–402.

Bernardo, Jose M. 1979. "Reference posterior distributions for bayesian inference." *Journal of the Royal Statistical Society: Series B (Methodological)* 41(2): 113–128.

Betancourt, Michael. 2017. "A conceptual introduction to hamiltonian monte carlo." *arXiv preprint arXiv:1701.02434*.

Betancourt, Michael. 2019. "The convergence of markov chain monte carlo methods: From the metropolis method to hamiltonian monte carlo." *Annalen der Physik* 531(3): 1700214.

Betancourt, Michael. 2018. "Towards a principled bayesian workflow."

Betancourt, Michael, and Mark Girolami. 2015. "Hamiltonian monte carlo for hierarchical models." *Current trends in Bayesian methodology with applications* 79: 30.

Bisbee, James. 2019. "BARP: Improving mister p using bayesian additive regression trees." *American Political Science Review* 113(4): 1060–1065.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Springer.

Black, Duncan. 1948. "On the rationale of group decision-making." *The Journal of Political Economy*: 23–34.

Blackwell, Matthew, and Adam N Glynn. 2018. "How to make causal inferences with time-series cross-sectional data under selection on observables." *American Political Science Review* 112(4): 1067–1082.

Blackwell, Matthew, James Honaker, and Gary King. 2017. "A unified approach to measurement error and missing data: Overview and applications." *Sociological Methods & Research* 46(3): 303–341.

Blackwell, Matthew, and Michael Olson. 2020. "Reducing model misspecification and bias in the estimation of interactions."

Boatright, Robert G. 2013. *Getting primaried: The changing politics of congressional primary challenges*. University of Michigan Press.

Boatright, Robert G, Vincent G Moscardelli, and Clifford Vickrey. 2017. "The consequences of primary election timing." *Primary Timing Project, June*.

Bonica, Adam. 2019a. "Are donation-based measures of ideology valid predictors of individual-level policy preferences?" *The Journal of Politics* 81(1): 327–333.

Bonica, Adam. 2019b. "Database on ideology, money in politics, and elections: Public version 1.0."

Bonica, Adam. 2013. "Ideology and interests in the political marketplace." *American Journal of Political Science* 57(2): 294–311.

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

Bonica, Adam. 2020. "Why are there so many lawyers in congress?" *Legislative Studies Quarterly* 45(2): 253–289.

Brady, David W, Hahrie Han, and Jeremy C Pope. 2007. "Primary elections and candidate ideology: Out of step with the primary electorate?" *Legislative Studies Quarterly* 32(1): 79–105.

Broockman, David E. 2016. "Approaches to studying policy representation." *Legislative Studies Quarterly* 41(1): 181–215.

Broockman, David E, and Christopher Skovron. 2018. "Bias in perceptions of public opinion among political elites." *American Political Science Review* 112(3): 542–563.

Brunell, Thomas L. 2006. "Rethinking redistricting: How drawing uncompetitive districts eliminates gerrymanders, enhances representation, and improves attitudes toward congress." *PS: Political Science and Politics* 39(1): 77–85.

Brunell, Thomas L, Bernard Grofman, and Samuel Merrill. 2016. "Components of party polarization in the us house of representatives." *Journal of Theoretical Politics* 28(4): 598–624.

Bullock, Will, and Joshua D Clinton. 2011. "More a molehill than a mountain: The effects of the blanket primary on elected officials' behavior from california." *The Journal of Politics* 73(3): 915–930.

Burden, Barry C. 2004. "Candidate positioning in u.s. Congressional elections." *British Journal of Political Science* 34(02): 211–227.

Burden, Barry C. 2001. "The polarizing effects of congressional primaries." *Congressional Primaries and the Politics of Representation*: 95–115.

Burden, Barry C, Gregory A Caldeira, and Tim Groseclose. 2000. "Measuring the ideologies of us senators: The song remains the same." *Legislative Studies Quarterly*: 237–258.

Butler, Daniel M, and Eleanor Neff Powell. 2014. "Understanding the party brand: Experimental evidence on the role of valence." *The Journal of Politics* 76(2): 492–505.

Buttice, Matthew K, and Benjamin Highton. 2013. "How does multilevel regression and poststratification perform with conventional national surveys?" *Political Analysis* 21(4): 449–467.

Bürkner, Paul-Christian, and others. 2017. "Brms: An r package for bayesian multilevel models using stan." *Journal of Statistical Software* 80(1): 1–28.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica* 82(6): 2295–2326.

Campbell, Angus et al. 1960. New York: John Wiley and Sons 77 *The american voter*.

Canes-Wrone, Brandice, David W Brady, and John F Cogan. 2002. "Out of step, out of office: Electoral accountability and house members' voting." *American Political Science Review* 96(01): 127–140.

Canes-Wrone, Brandice, William Minozzi, and Jessica Bonney Reveley. 2011. "Issue accountability and the mass public." *Legislative Studies Quarterly* 36(1): 5–35.

Carlson, David. 2020. "Estimating a counter-factual with uncertainty through gaussian process projection."

Carmines, Edward G, and James A Stimson. 1986. "On the structure and sequence of issue evolution." *American Political Science Review* 80(03): 901–920.

Carpenter, Bob et al. 2016. "Stan: A probabilistic programming language." *Journal of Statistical Software* 20: 1–37.

Carvalho, Carlos M, Nicholas G Polson, and James G Scott. 2010. "The horseshoe estimator for sparse signals." *Biometrika* 97(2): 465–480.

Caughey, Devin, James Dunham, and Christopher Warshaw. 2018. "The ideological nationalization of partisan subconstituencies in the american states." *Public Choice* 176(1-2): 133–151.

Caughey, Devin, and Christopher Warshaw. 2015. "Dynamic estimation of latent opinion using a hierarchical group-level irt model." *Political Analysis* 23(2): 197–211.

Caughey, Devin, and Christopher Warshaw. 2018. "Policy preferences and policy change: Dynamic responsiveness in the american states, 1936–2014."

Caughey, Devin, and Christopher Warshaw. 2019. "Public opinion in subnational politics." *The Journal of Politics* 81(1): 352–363.

Chernozhukov, Victor et al. 2017. "Double/debiased/neyman machine learning of treatment effects." *American Economic Review* 107(5): 261–65.

Chernozhukov, Victor et al. 2018. "Double/debiased machine learning for treatment and structural parameters."

Chipman, Hugh A, Edward I George, and Robert E McCulloch. 2010. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics* 4(1): 266–298.

Clinton, Joshua D. 2006. "Representation in congress: Constituents and roll calls in the 106th house." *Journal of Politics* 68(2): 397–409.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98(02): 355–370.

Cohen, Marty et al. 2009. *The party decides: Presidential nominations before and after reform*. University of Chicago Press.

Cox, Gary W. 1990. "Centripetal and centrifugal incentives in electoral systems." *American Journal of Political Science*: 903–935.

Cox, Gary W, and Mathew D McCubbins. 2005. *Setting the agenda: Responsible party government in the us house of representatives*. Cambridge University Press.

Culbert, Gar. 2015. "Realizing 'strategic' voting in presidential primaries." *Rationality and Society* 27(2): 224–256.

DeCrescenzo, Michael G, and Kenneth R Mayer. 2019. "Voter identification and nonvoting in wisconsin—evidence from the 2016 election." *Election Law Journal: Rules, Politics, and Policy* 18(4): 342–359.

Devezer, Berna et al. 2020. "The case for formal methodology in scientific reform." *bioRxiv*.

Doherty, David, Conor M Dowling, and Michael G Miller. 2019. "Do local party chairs think women and minority candidates can win? Evidence from a conjoint experiment." *The Journal of Politics* 81(4): 1282–1297.

Downs, Anthony. 1957. *An economic theory of democracy*. New York: Harper; Row.

Duane, Simon et al. 1987. "Hybrid monte carlo." *Physics letters B* 195(2): 216–222.

Duck-Mayr, JBrandon, Roman Garnett, and Jacob Montgomery. 2020. "GPIRT: A gaussian process model for item response theory." In *Conference on uncertainty in artificial intelligence*, PMLR, p. 520–529.

Egami, Naoki et al. 2018. "How to make causal inferences using texts." *arXiv preprint arXiv:1802.02163*.

Ellis, Christopher, and James A Stimson. 2012. *Ideology in america*. Cambridge University Press.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2018. "Using a probabilistic model to assist merging of large-scale administrative records." *Available at SSRN 3214172*.

Enns, Peter K, and Julianna Koch. 2013. "Public opinion in the us states: 1956 to 2010." *State Politics & Policy Quarterly* 13(3): 349–372.

Epstein, Lee et al. 2007. "The judicial common space." *Journal of Law, Economics, and Organization* 23(2): 303–325.

Feldman, Stanley, and John Zaller. 1992. "The political culture of ambivalence: Ideological responses to the welfare state." *American Journal of Political Science*: 268–307.

Fenno, Richard F. 1978. *Home style: House members in their districts*. Pearson College Division.

Fienberg, Stephen E. 2006. "Does it make sense to be an" objective bayesian"?(Comment on articles by berger and by goldstein)." *Bayesian Analysis* 1(3): 429–432.

Fiorina, Morris P, Samuel J Abrams, and Jeremy C Pope. 2005a. *Culture war? The myth of a polarized america*. Pearson Longman New York.

Fiorina, Morris P, Samuel J Abrams, and Jeremy C Pope. 2005b. *Culture war? The myth of a polarized america*. Pearson Longman New York.

Foster-Molina, Ella. 2016. "Legislative and district data 1972–2013." https://doi.org/10.7910/DVN/26448.

Fowler, Anthony, and Andrew B Hall. 2016. "The elusive quest for convergence." *Quarterly Journal of Political Science* 11: 131–149.

Fowler, Linda L. 1982. "How interest groups select issues for rating voting records of members of the us congress." *Legislative Studies Quarterly*: 401–413.

Fox, Jean-Paul. 2010. *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Fox, Jeremy T et al. 2012. "The random coefficients logit model is identified." *Journal of Econometrics* 166(2): 204–212.

Free, Lloyd A, and Hadley Cantril. 1967. "The political beliefs of americans."

Freeman, Jo. 1986. "The political culture of the democratic and republican parties." *Political Science Quarterly* 101(3): 327–356.

Gabry, Jonah et al. 2019. "Visualization in bayesian workflow." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182(2): 389–402.

García-Pérez, Miguel Ángel. 2019. "Bayesian estimation with informative priors is indistinguishable from data falsification." *The Spanish journal of psychology* 22.

Geer, John G. 1988. "Assessing the representativeness of electorates in presidential primaries." *American Journal of Political Science*: 929–945.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. "Philosophy and the practice of bayesian statistics." *British Journal of Mathematical and Statistical Psychology* 66(1): 8–38.

Gelman, Andrew et al. 2007. "Rich state, poor state, red state, blue state: What's the matter with connecticut?" *Quarterly Journal of Political Science* 2(4): 345–357.

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. "The prior can often only be understood in the context of the likelihood." *Entropy* 19(10): 555.

Gelman, Andrew et al. 2013. *Bayesian data analysis*. Chapman; Hall/CRC.

Gelman, Andrew, and Yuling Yao. 2020. "Holes in bayesian statistics." *arXiv preprint arXiv:2002.06467*.

Gerring, John. 2001. *Social science methodology: A criterial framework*. Cambridge University Press.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups." *American Journal of Political Science* 57(3): 762–776.

Gilens, Martin, and Benjamin I. Page. 2014. "Testing theories of american politics: Elites, interest groups, and average citizens." *Perspectives on Politics* 12(3).

Gill, Jeff. 2014. 20 *Bayesian methods: A social and behavioral sciences approach*. CRC press.

Gill, Jeff. 1999. "The insignificance of null hypothesis significance testing." *Political research quarterly* 52(3): 647–674.

Goodman-Bacon, Andrew. 2018. *Difference-in-differences with variation in treatment timing*. National Bureau of Economic Research.

Green, Donald, Bradley Palmquist, and Eric Schickler. 2002. *Partisan hearts and minds*. New Haven, CT: Yale University Press.

Green, Donald P, and Holger L Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees." *Public opinion quarterly* 76(3): 491–511.

Green, Donald P et al. 2016. "The effects of lawn signs on vote outcomes: Results from four randomized field experiments." *Electoral Studies* 41: 143–150.

Greenland, Sander, Judea Pearl, and James M Robins. 1999. "Causal diagrams for epidemiologic research." *Epidemiology*: 37–48.

Grimmer, Justin. 2011. "An introduction to bayesian inference via variational approximations." *Political Analysis* 19(1): 32–47.

Grossman, Matthew, and David A. Hopkins. 2016. Oxford University Press *Asymmetric politics: Ideological republicans and group interest democrats*.

Grosz, Michael P, Julia M Rohrer, and Felix Thoemmes. 2020. "The taboo against explicit causal inference in nonexperimental psychology."

Hacker, Jacob S, and Paul Pierson. 2005. *Off center: The republican revolution and the erosion of american democracy*. Yale University Press.

Hahn, P Richard et al. 2018. "Regularization and confounding in linear regression for treatment effect estimation." *Bayesian Analysis* 13(1): 163–182.

Hahn, P Richard, Jared S Murray, and Carlos M Carvalho. 2020. "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects." *Bayesian Analysis*.

Hall, Andrew B. 2015. "What happens when extremists win primaries?" *American Political Science Review* 109(01): 18–42.

Hall, Andrew B, and James M Snyder. 2015. "Candidate ideology and electoral success. Working paper: Https://dl. Dropboxusercontent.com/u/11481940/hall snyder ideology.pdf."

Hall, Andrew B, and Daniel M Thompson. 2018. "Who punishes extremist nominees? Candidate ideology and turning out the base in us elections." *American Political Science Review* 112(3): 509–524.

Hare, Christopher et al. 2015. "Using bayesian aldrich-mckelvey scaling to study citizens' ideological preferences and perceptions." *American Journal of Political Science* 59(3): 759–774.

Henderson, John A. 2016. "An experimental approach to measuring ideological positions in political text." *Available at SSRN 2852784*.

Henry, Patrick J, and David O Sears. 2002. "The symbolic racism 2000 scale." *Political Psychology* 23(2): 253–283.

Hernán, Miguel A. 2018. "The c-word: Scientific euphemisms do not improve causal inference from observational data." *American journal of public health* 108(5): 616–619.

Hill, Jennifer. "Multilevel models and causal inference." In *The SAGE handbook of multilevel modeling*, SAGE Publications Ltd, p. 201–220. https://doi.org/10.4135/9781446247600.n12.

Hill, Jennifer L. 2011. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics* 20(1): 217–240.

Hill, Seth J. 2015. "Institution of nomination and the policy ideology of primary electorates." *Quarterly Journal of Political Science* 10(4): 461–487.

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.

Hill, Seth J, and Chris Tausanovitch. 2015. "A disconnect in representation? Comparison of trends in congressional and public polarization." *The Journal of Politics* 77(4): 1058–1075.

Hinne, Max, Marcel AJ van Gerven, and Luca Ambrogioni. 2019. "Causal inference using bayesian non-parametric quasi-experimental design." *arXiv preprint arXiv:1911.06722*.

Hirano, Shigeo et al. 2010. "Primary elections and party polarization." *Quarterly Journal of Political Science* 5: 169–191.

Hirano, Shigeo, and Michael M Ting. 2015. "Direct and indirect representation." *British Journal of Political Science* 45(3): 609.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81(396): 945–960.

Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. "Designing and analyzing randomized experiments: Application to a japanese election survey experiment." *American Journal of Political Science* 51(3): 669–687.

Imai, Kosuke et al. 2011. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies." *American Political Science Review*: 765–789.

Imai, Kosuke, and In Song Kim. 2019. "When should we use unit fixed effects regression models for causal inference with longitudinal data?" *American Journal of Political Science* 63(2): 467–490.

Imbens, Guido W, and Donald B Rubin. 1997. "Bayesian inference for causal effects in randomized experiments with noncompliance." *The annals of statistics*: 305–327.

Jackman, Simon. 2009. 846 *Bayesian analysis for the social sciences*. John Wiley & Sons.

Jackman, Simon. 2000. "Estimation and inference are missing data problems: Unifying social science statistics via bayesian simulation." *Political Analysis*: 307–332.

Jacobson, Gary C. 2012. "The electoral origins of polarized politics: Evidence from the 2010 cooperative congressional election study." *American Behavioral Scientist* 56(12): 1612–1630.

Jeffreys, Harold. 1946. "An invariant form for the prior probability in estimation problems." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186(1007): 453–461.

Jeffreys, Harold. 1998. *The theory of probability*. OUP Oxford.

Jordan, Marty P, and Matt Grossmann. 2017. "Correlates of us state public policies: Announcing a new database." *Michigan State University* 23.

Kam, Cindy D, and Marc J Trussler. 2017. "At the nexus of observational and experimental research: Theory, specification, and analysis of experiments with heterogeneous treatment effects." *Political Behavior* 39(4): 789–815.

Karol, David. 2009. *Party position change in american politics: Coalition management*. Cambridge University Press.

Kass, Robert E, and Larry Wasserman. 1996. "The selection of prior distributions by formal rules." *Journal of the American statistical Association* 91(435): 1343–1370.

Kastellec, Jonathan P et al. 2015. "Polarizing the electoral connection: Partisan representation in supreme court confirmation politics." *The journal of politics* 77(3): 787–804.

Kastellec, Jonathan P, Jeffrey R Lax, and Justin H Phillips. 2010. "Estimating state public opinion with multi-level regression and poststratification using r." *Unpublished manuscript, Princeton University*.

Keele, Luke. 2015. "The statistics of causal inference: A view from political methodology." *Political Analysis* 23(3): 313–335.

Keele, Luke, Corrine McConnaughy, and Ismail White. 2012. "Strengthening the experimenter's toolbox: Statistical estimation of internal validity." *American Journal of Political Science* 56(2): 484–499.

Kernell, Georgia. 2009. "Giving order to districts: Estimating voter distributions with national election returns." *Political Analysis* 17(3): 215–235.

Key, Valdimer Orlando. 1955. "Politics, parties, and pressure groups."

Key, V.O. Jr. 1949. "Southern politics in state and nation."

King, Aaron S, Frank J Orlando, and David B Sparks. 2016. "Ideological extremity and success in primary elections: Drawing inferences from the twitter network." *Social Science Computer Review* 34(4): 395–415.

Koger, Gregory, Seth Masket, and Hans Noel. 2009. "Partisan webs: Information exchange and party networks." *British Journal of Political Science*: 633–653.

Kucukelbir, Alp et al. 2015. "Automatic variational inference in stan." In *Advances in neural information processing systems*, p. 568–576.

Kujala, Jordan. 2020. "Donors, primary elections, and polarization in the united states." *American Journal of Political Science* 64(3): 587–602.

La Raja, Raymond, and Brian Schaffner. 2015. *Campaign finance and political polarization: When purists prevail*. University of Michigan Press.

Lattimore, Finnian, and David Rohde. 2019. "Replacing the do-calculus with bayes rule." *arXiv preprint arXiv:1906.07125*.

Lauderdale, Benjamin E. 2010. "Unpredictable voters in ideal point estimation." *Political Analysis*: 151–171.

Lax, Jeffrey R, and Justin H Phillips. 2009. "How should we estimate public opinion in the states?" *American Journal of Political Science* 53(1): 107–121.

Layman, Geoffrey C., and Thomas M. Carsey. 2002. "Party Polarization and "Conflict Extension" in the American Electorate." *American Journal of Political Science* 46(4): 786. http://www.jstor.org/stable/3088434?origin=crossref (Accessed February 22, 2015).

Layman, Geoffrey C et al. 2010. "Activists and conflict extension in american party politics." *American Political Science Review*: 324–346.

Leavitt, Thomas. 2020. "Causal inference in difference-in-differences designs under uncertainty in counterfactual trends."

Lebo, Matthew J, Adam J McGlynn, and Gregory Koger. 2007. "Strategic party government: Party influence in congress, 1789–2000." *American Journal of Political Science* 51(3): 464–481.

Lelkes, Yphtach. 2019. "Policy over party: Comparing the effects of candidate ideology and party on affective polarization." *Political Science Research and Methods*: 1–8.

Lelkes, Yphtach, and Paul M Sniderman. 2016. "The ideological asymmetry of the american party system." *British Journal of Political Science* 46(4): 825–844.

Lemm, Jörg C. 1996. "Prior information and generalized questions." In *Massachusetts institute of technology, artificial intelligence laboratory and center for biological and computational learning, department of brain and cognitive sciences*, Citeseer.

Levendusky, Matthew. 2009. *The partisan sort: How liberals became democrats and conservatives became republicans*. University of Chicago Press.

Levendusky, Matthew S, Jeremy C Pope, and Simon D Jackman. 2008. "Measuring district-level partisanship with implications for the analysis of us elections." *The Journal of Politics* 70(3): 736–753.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. "Generating random correlation matrices based on vines and extended onion method." *Journal of multivariate analysis* 100(9): 1989–2001.

Liao, Shirley X. 2019. "Bayesian causal inference for estimating impacts of air pollution exposure." PhD thesis.

Liao, Shirley X, and Corwin M Zigler. 2020. "Uncertainty in the design stage of two-stage bayesian propensity score analysis." *Statistics in Medicine*.

Link, William A., and Mitchell J. Eaton. 2011. "On thinning of chains in MCMC." *Methods in Ecology and Evolution* 3(1): 112–115. https://doi.org/10.1111/j.2041-210x.2011.00131.x.

Linzer, Drew A, and Jeffrey K Staton. 2015. "A global measure of judicial independence, 1948–2012." *Journal of Law and Courts* 3(2): 223–256.

Londregan, John. 1999. "Estimating legislators' preferred points." *Political Analysis* 8(1): 35–56.

MacKay, David JC. 1992. "A practical bayesian framework for backpropagation networks." *Neural computation* 4(3): 448–472.

Maisel, L Sandy, and Walter J Stone. 1997. "Determinants of candidate emergence in us house elections: An exploratory study." *Legislative Studies Quarterly*: 79–96.

Mann, Thomas E. 1978. 220 *Unsafe at any margin: Interpreting congressional elections*. Aei Pr.

Martin, Andrew D, and Kevin M Quinn. 2002. "Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999." *Political Analysis* 10(2): 134–153.

Masket, Seth. 2009. *No middle ground: How informal party organizations control nominations and polarize legislatures*. University of Michigan Press.

Mayhew, David R. 1974. *Congress: The electoral connection*. Yale University Press.

Mayhew, David R. 1986. *Placing parties in american politics: Organization, electoral settings, and government activity in the twentieth century*. Princeton University Press.

Mayo, Deborah G. 2018. *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

Mayo, Deborah G. 2013. "The error-statistical philosophy and the practice of bayesian statistics: Comments on gelman and shalizi: 'Philosophy and the practice of bayesian statistics'." *Br J Math Stat Psychol* 66(1): 57–64.

McCandless, Lawrence C, Paul Gustafson, and Peter C Austin. 2009. "Bayesian propensity score analysis for observational data." *Statistics in medicine* 28(1): 94–112.

McCarty, Nolan, and Howard Poole Keith T. and Rosenthal. 2006. *Polarized america: The dance of ideology and unequal riches*. Cambridge, MA: MIT Press.

McCarty, Nolan, Keith T Poole, and Howard Rosenthal. 2009. "Does gerrymandering cause polarization?" *American Journal of Political Science* 53(3): 666–680.

McClosky, Herbert, Paul J Hoffmann, and Rosemary O'Hara. 1960. "Issue conflict and consensus among party leaders and followers." *The American Political Science Review* 54(2): 406–427.

McElreath, Richard. 2017a. "Bayesian inference is just counting."

McElreath, Richard. 2017b. "Bayesian statistics without frequentist language."

McElreath, Richard. 2020. *Statistical rethinking: A bayesian course with examples in r and stan*. 2nd ed. CRC press.

McFadden, Daniel. 1973. "Conditional logit analysis of qualitative choice behavior."

McGann, Anthony J. 2014. "Estimating the political center from aggregate data: An item response theory alternative to the stimson dyad ratios algorithm." *Political Analysis*: 115–129.

McGhee, Eric et al. 2014. "A primary cause of partisanship? Nomination systems and legislator ideology." *American Journal of Political Science* 58(2): 337–351.

Meager, Rachael. 2019. "Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments." *American Economic Journal: Applied Economics* 11(1): 57–91.

Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3): 760–775.

Navarro, Danielle J. 2019. "Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection." *Computational Brain & Behavior* 2(1): 28–34.

Neal, Radford M. 2012. "MCMC using hamiltonian dynamics." *arXiv preprint arXiv:1206.1901*.

Nielson, Lindsay, and Neil Visalvanich. 2017. "Primaries and candidates: Examining the influence of primary electorates on candidate ideology." *Political Science Research and Methods* 5(2): 397–408.

Norrander, Barbara. 1989. "Ideological representativeness of presidential primary voters." *American Journal of Political Science*: 570–587.

Nyhuis, Dominic. 2018. "Separating candidate valence and proximity voting: Determinants of competitors' non-policy appeal." *Political Science Research and Methods* 6(1): 135.

Oganisian, Arman, and Jason A Roy. 2020. "A practical introduction to bayesian estimation of causal effects: Parametric and nonparametric approaches." *arXiv preprint arXiv:2004.07375*.

Ornstein, Joseph T, and JBrandon Duck-Mayr. 2020. "Gaussian process regression discontinuity."

Pacheco, Julianna. 2011. "Using national surveys to measure dynamic us state public opinion: A guideline for scholars and an application." *State Politics & Policy Quarterly*: 1532440011419287.

Papaspiliopoulos, Omiros, Gareth O Roberts, and Martin Sköld. 2007. "A general framework for the parametrization of hierarchical models." *Statistical Science*: 59–73.

Park, David K, Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian multilevel estimation with poststratification: State-level estimates from national polls." *Political Analysis* 12(4): 375–385.

Park, Trevor, and George Casella. 2008. "The bayesian lasso." *Journal of the American Statistical Association* 103(482): 681–686.

Patil, Anand, David Huard, and Christopher J Fonnesbeck. 2010. "PyMC: Bayesian stochastic modelling in python." *Journal of statistical software* 35(4): 1.

Pearl, Judea. 1995. "Causal diagrams for empirical research." *Biometrika* 82(4): 669–688.

Pearl, Judea. 2009. *Causality: Models, reasoning, and inference*. 2nd ed. Cambridge University Press.

Petrocik, John Richard. 2009. "Measuring party support: Leaners are not independents." *Electoral Studies* 28(4): 562–572. http://linkinghub.elsevier.com/retrieve/pii/S0261379409000511 (Accessed April 16, 2015).

Phillips, Anne. 1995. *The politics of presence*. Clarendon Press.

Piironen, Juho, and Aki Vehtari. 2017a. "On the hyperprior choice for the global shrinkage parameter in the horseshoe prior." In *Artificial intelligence and statistics*, PMLR, p. 905–913.

Piironen, Juho, and Aki Vehtari. 2017b. "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics* 11(2): 5018–5051.

Pitkin, Hanna Fenichel. 1967. *The concept of representation*. Univ of California Press.

Plummer, Martyn. 2015. "Cuts in bayesian graphical models." *Statistics and Computing* 25(1): 37–43.

Poole, Keith T, and Howard Rosenthal. 1997. "Congress: A political-economic history of roll call voting." *New York: Oxford University Press*.

Porter, Rachel A, and Sarah Treul. 2020. "Reevaluating experience in congressional primary elections."

Rahn, Wendy M. 1993. "The role of partisan stereotypes in information processing about political candidates." *American Journal of Political Science*: 472–496.

Rainey, Carlisle. 2014. "Arguing for a negligible effect." *American Journal of Political Science* 58(4): 1083–1091.

Ratkovic, Marc. 2019. "Rehabilitating the regression: Honest and valid causal inference through machine learning."

Ratkovic, Marc, and Dustin Tingley. 2017a. "Causal inference through the method of direct estimation." *arXiv preprint arXiv:1703.05849*.

Ratkovic, Marc, and Dustin Tingley. 2017b. "Sparse estimation and uncertainty with application to subgroup analysis." *Political Analysis* 25(1): 1–40.

Robins, James M. 1997. "Causal inference from complex longitudinal data." In *Latent variable modeling and applications to causality*, Springer, p. 69–117.

Rogowski, Jon C. 2016. "Voter decision-making with polarized choices." *British Journal of Political Science*: 1–22. https://doi.org/10.1017%2Fs0007123415000630.

Rogowski, Jon C, and Stephanie Langella. 2015. "Primary systems and candidate ideology: Evidence from federal and state legislative elections." *American Politics Research* 43(5): 846–871.

Rubin, Donald B. 1978. "Bayesian inference for causal effects: The role of randomization." *The Annals of statistics*: 34–58.

Rubin, Donald B. 1984. "Bayesianly justifiable and relevant frequency calculations for the applies statistician." *The Annals of Statistics*: 1151–1172.

Rubin, Donald B. 2005. "Causal inference using potential outcomes: Design, modeling, decisions." *Journal of the American Statistical Association* 100(469): 322–331.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66(5): 688.

Rubin, Donald B. 1981. "Estimation in parallel randomized experiments." *Journal of Educational Statistics* 6(4): 377–401.

Rubin, Ruth Bloch. 2017. *Building the bloc: Intraparty organization in the us congress*. Cambridge University Press.

Samii, Cyrus, Laura Paler, and Sarah Zukerman Daly. 2016. "Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia." *Political Analysis* 24(4): 434–456.

Schlozman, Daniel. 2015. *When movements anchor parties: Electoral alignments in american history*. Princeton University Press.

Seaman III, John W, John W Seaman Jr, and James D Stamey. 2012. "Hidden dangers of specifying noninformative priors." *The American Statistician* 66(2): 77–84.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12(1): 487–508. http://www.annualreviews.org/doi/abs/10.1146/annurev.polisci.11.060606.135444 (Accessed January 13, 2015).

Shor, Boris, and Nolan McCarty. 2011a. "The ideological mapping of american legislatures." *American Political Science Review* 105(03): 530–551.

Shor, Boris, and Nolan McCarty. 2011b. "The ideological mapping of american legislatures." *American Political Science Review* 105(03): 530–551.

Sides, John et al. 2020. "On the representativeness of primary electorates." *British Journal of Political Science*: 1–9.

Simas, Elizabeth N. 2013. "Proximity voting in the 2010 us house elections." *Electoral Studies* 32(4): 708–717.

Simas, Elizabeth N. 2017. "The effects of electability on us primary voters." *Journal of Elections, Public Opinion and Parties* 27(3): 274–290.

Simpson, Daniel et al. 2017. "Penalising model component complexity: A principled, practical approach to constructing priors." *Statistical science* 32(1): 1–28.

Snyder, James M Jr. 1994. "Safe seats, marginal seats, and party platforms: The logic of platform differentiation." *Economics & Politics* 6(3): 201–213.

Snyder Jr, James M. 1992. "Artificial extremism in interest group ratings." *Legislative Studies Quarterly*: 319–345.

Stimson, James A. 1991. *Public opinion in america: Moods, cycles, and swings*. Westview Press.

Stokes, Donald E. 1963. "Spatial models of party competition." *The American Political Science Review* 57(2): 368–377.

Stone, Walter J, and L Sandy Maisel. 2003. "The not-so-simple calculus of winning: Potential us house candidates' nomination and general election prospects." *The Journal of Politics* 65(4): 951–977.

Stone, Walter J, and Elizabeth N Simas. 2010. "Candidate valence and ideological positions in us house elections." *American Journal of Political Science* 54(2): 371–388.

Tausanovitch, Chris, and Christopher Warshaw. 2017. "Estimating candidates' political orientation in a polarized congress." *Political Analysis* 25(2): 167–187.

Tausanovitch, Chris, and Christopher Warshaw. 2013. "Measuring constituent policy preferences in congress, state legislatures, and cities." *The Journal of Politics* 75(02): 330–342.

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.

Thomsen, Danielle M. 2020. "Ideology and gender in us house elections." *Political Behavior* 42(2): 415–442.

Thomsen, Danielle M. 2019. "Which women win? Partisan changes in victory patterns in us house elections." *Politics, Groups, and Identities* 7(2): 412–428.

Thomsen, Danielle M, and Michele L Swers. 2017. "Which women can run? Gender, partisanship, and candidate donor networks." *Political Research Quarterly* 70(2): 449–463.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1): 267–288.

Tomz, Michael, and Robert P Van Houweling. 2008. "Candidate positioning and voter choice." *American Political Science Review*: 303–318.

Treier, Shawn, and D Sunshine Hillygus. 2009. "The nature of political ideology in the contemporary electorate." *Public Opinion Quarterly* 73(4): 679–703.

Treier, Shawn, and Simon Jackman. 2008. "Democracy as a latent variable." *American Journal of Political Science* 52(1): 201–217.

VanderWeele, Tyler J. 2009. "On the distinction between interaction and effect modification." *Epidemiology* 20(6): 863–871.

VanderWeele, Tyler J, and James M Robins. 2007. "Four types of effect modification: A classification based on directed acyclic graphs." *Epidemiology* 18(5): 561–568.

Vansteelandt, Stijn. 2009. "Estimating direct effects in cohort and case–control studies." *Epidemiology*: 851–860.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical bayesian model evaluation using leave-one-out cross-validation and waic." *Statistics and computing* 27(5): 1413–1432.

Vehtari, Aki et al. 2020. "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC." *Bayesian Analysis*.

Verba, Sidney, and Norman H Nie. 1972. *Participation in america*. Harper & Row.

Voeten, Erik. 2000. "Clashes in the assembly." *International organization*: 185–215.

Wagenmakers, Eric-Jan. 2007. "A practical solution to the pervasive problems of *p* values." *Psychonomic bulletin & review* 14(5): 779–804.

Wager, Stefan, and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113(523): 1228–1242.

Wang, Bo, and D Titterington. 2012. "Convergence and asymptotic normality of variational bayesian approximations for exponential family models with missing values." *arXiv preprint arXiv:1207.4159*.

Warshaw, Christopher, and Jonathan Rodden. 2012. "How should we measure district-level public opinion on individual issues?" *The Journal of Politics* 74(01): 203–219.

Western, Bruce, and Simon Jackman. 1994. "Bayesian inference for comparative research." *American Political Science Review*: 412–423.

Woods, Carol M, and David Thissen. 2006. "Item response theory with estimation of the latent population distribution using spline-based densities." *Psychometrika* 71(2): 281.

Xu, Yiqing. 2017. "Generalized synthetic control method: Causal inference with interactive fixed effects models." *Political Analysis* 25(1): 57–76.

Zigler, Corwin Matthew. 2016. "The central role of bayes' theorem for joint estimation of causal effects and propensity scores." *The American Statistician* 70(1): 47–54.

Zigler, Corwin Matthew, and Francesca Dominici. 2014. "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects." *Journal of the American Statistical Association* 109(505): 95–107.

Zigler, Corwin M et al. 2013. "Model feedback in bayesian propensity score estimation." *Biometrics* 69(1): 263–273.