Do Primaries Work?

Bayesian Causal Models of Partisan 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, **TBD**

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, and 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 isn't 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, Pierson, and others 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.

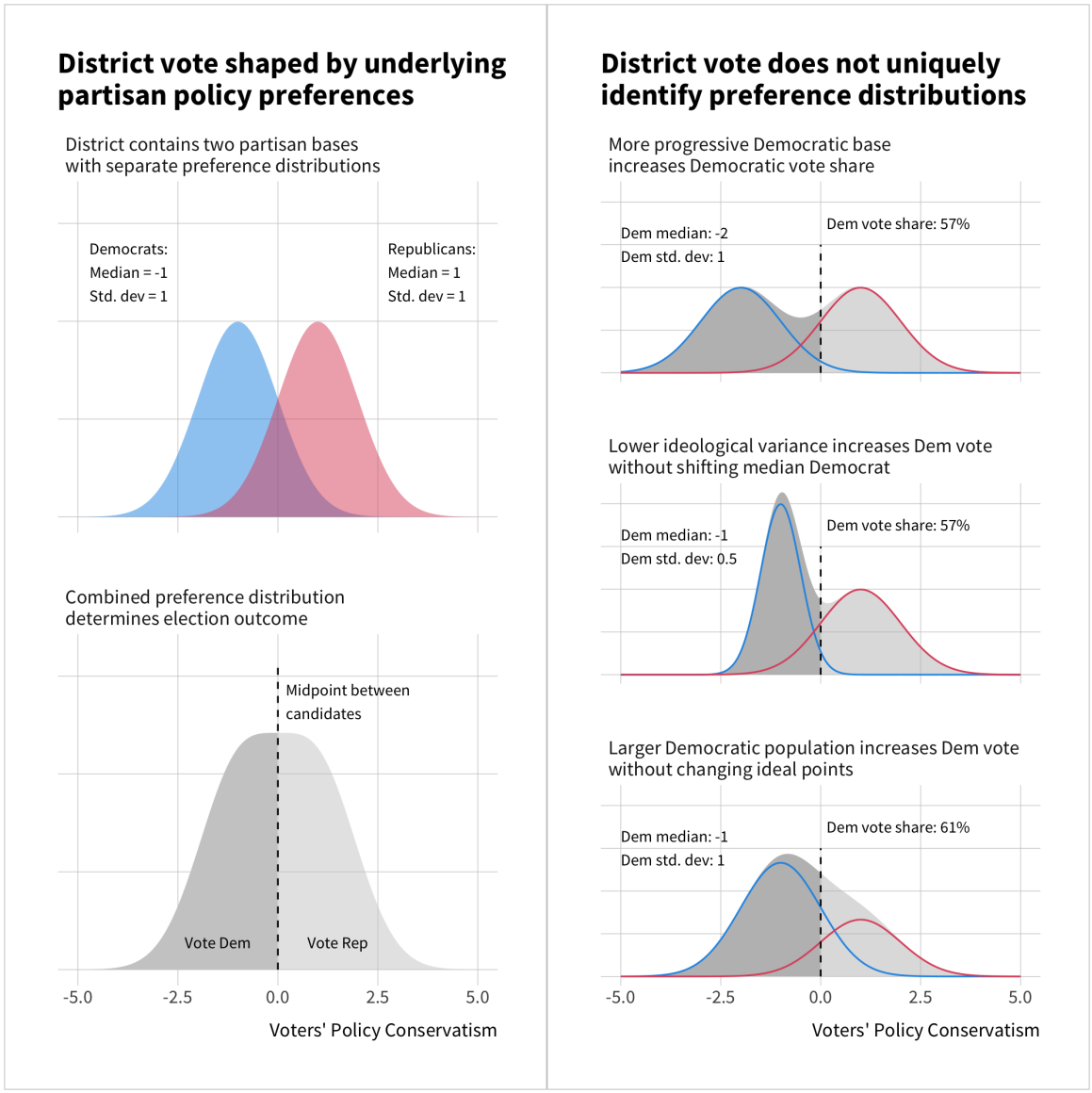


Figure 1.1: Demonstrating how district vote shares from a single election are insufficient to identify underlying policy-ideological features of the district. The left side shows how the policy preference distributions for two parties in a district (top panel) combine to form an aggregate preference 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

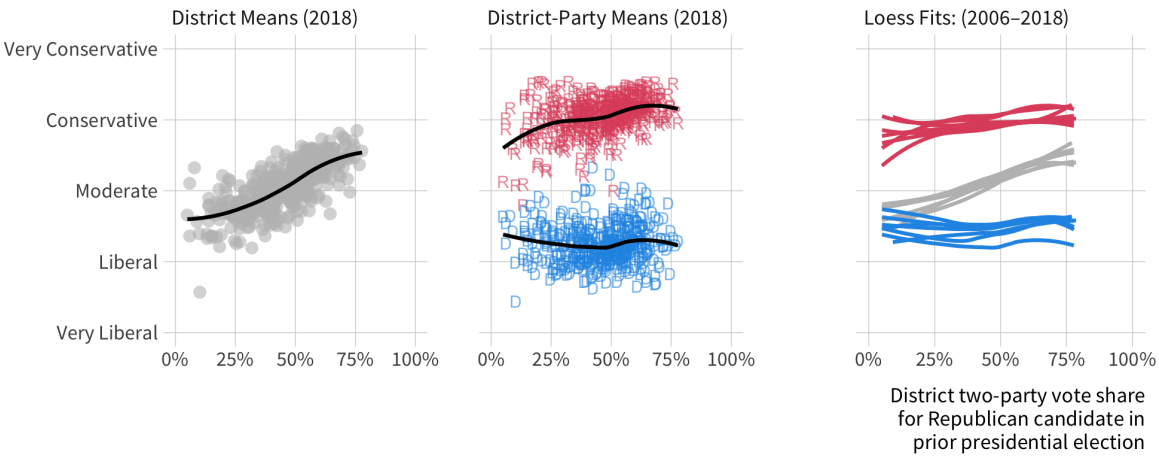


Figure 1.2: Average ideological self placement (vertical axis) and Republican vote share (horizontal axis) in all 435 congressional districts. Mean self-placement is calculated by numerically coding CCES ideological self-placement responses before averaging. The first panel plots average self-placement 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 1978a), 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 2017), 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 (1978a) 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, which can be traced to spatial models of policy choice from classic formal theory work in American political science such as Downs (1957). 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 of prior distributions, and model parameterizations that expedite estimation with Markov chain Monte Carlo (MCMC). I begin with a static model for one time period, and then I describe a dynamic model that smooths estimates across time using hierarchical priors for model parameters (Caughey and Warshaw 2015). I test both the static and the dynamic models by fitting them to simulated data and determining how well they recover known parameter

values.

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. Depending on the structure of the game, these actors often face constrained choices; they can't achieve their most-preferred policy, so they settle on something that is as close as they can get.

Figure 2.1: An Actor and two policy outcomes (Left and Right) represented as locations in ideological/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, indicated by their locations on the line. The Actor has a location herself, which corresponds to her most-desired policy outcome. There is no policy located exactly at the Actor's preferred location, but the Actor is closer to the Right policy than to the Left.

Supposing that the Actor could make an error-free choice over which policy to implement, it appears she would prefer the Right outcome to the Left outcome.

Formal models are more careful to specify the assumptions governing these scenarios, which can be complicated in many cases. For example, suppose that locations along the left-right continuum can be assigned values on the real number line. Figure 2.1 shows a one-dimensional number line, but policies can be generally represented as locations in multidimensional $\mathbb{R}^d$ space. The Actor's location is synonymous with her most preferred policy: her "ideal point." This is the point where the Actor's utility, in an economic utility model, is maximized with respect to policy considerations. Utility implies that the Actor has a utility function that is defined over the policy space, which depends on the distance between her ideal point and a potential policy outcome. Outcomes nearer to the Actor's ideal point are generally more preferred than farther outcomes, but this too is subject to assumptions about the shape of the Actor's utility function. Typically utility functions are assumed to be single-peaked and symmetric around an Actor's ideal point, so a closer policy is always more preferred, all else equal. The notion of an ideal point is similar to a "bliss point" in microeconomics: the optimal quantity of a good consumed such that any more or less consumption would result in decreased utility. Whether an Actor can choose the closest policy to herself depends on the structure of the game: the presence and strategy profiles of other Actors, the sequence of play, and the presence of other non-policy features of Actors' utility functions.

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* a formal model as given 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. Researchers are typically interested in parameter estimates for the Actors' ideal points, although sometimes the parameters about the policy alternatives are substantively interesting.

Having distinguished formal and statistical models, I now show a derivation of a statistical model from a formal model. This exercise model will serve as a theoretical basis for the class of statistical models explored in this dissertation. I begin with notation to describe an arbitrary number of actors indexed $i \in \{1, \ldots, n\}$ making an arbitrary number of policy choices (bills, survey items, etc.) indexed $j \in \{1, \ldots, J\}$. Every Actor has an ideal point, or a location in the policy space, represented by θ_i. Every task is choice between a Left policy located at L_j and a Right policy located at R_j.

The utility that an Actor receives from a Left or Right choice is a function of the distance between her ideal point and the respective choice location. Utility is maximized if an Actor can choose a policy located exactly on her ideal point, and utility is "lost" for choices farther and farther from her ideal point. The functional form of utility loss is an assumption made by the researcher. Some scholars assume that utility loss follows a Gaussian curve, while others choose a quadratic utility loss (Clinton, Jackman, and Rivers 2004). For this analysis, I assume a quadratic utility loss.[1]

The choice of quadratic loss implies a utility function over the *squared distance* between an Actor and a choice location. The utility Actor i receives from choosing Left or Right are given by utility functions $U_i\left(L_j\right)$ and $U_i\left(R_j\right)$, respectively. With quadratic utility loss, these utility functions take the form

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

[1]Researchers typically avoid linear losses for technical reasons: a linear utility loss function is non-differentiable at the ideal point because function comes to a point. This prevents the researcher from using differential calculus to find a point of maximum utility.

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.

With these utility functions laid out, Actor i's decision can be a comparison of the utilities received by choosing Right or Left. Let y_{ij} indicate the Actor's choice of Right or Left, where Right is coded 1, and Left is coded 0. The model so far implies that $y_{ij} = 1$ (Actor chooses Right) if their utility is greater for Right than for Left.

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

To visualize this choice, I represent the deterministic components of Equation (2.2) in Figure 2.2, omitting the stochastic utility terms. The parabola represents i's fixed utility loss for any choice along the ideological continuum, owed to her distance from that choice. The vertex of the parabola is at the Actor's location, indicating that she would maximize her spatial utility if she could choose a policy located exactly at her ideal point. 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 she receives greater utility (or, less utility *loss*) by choosing Right instead of Left.

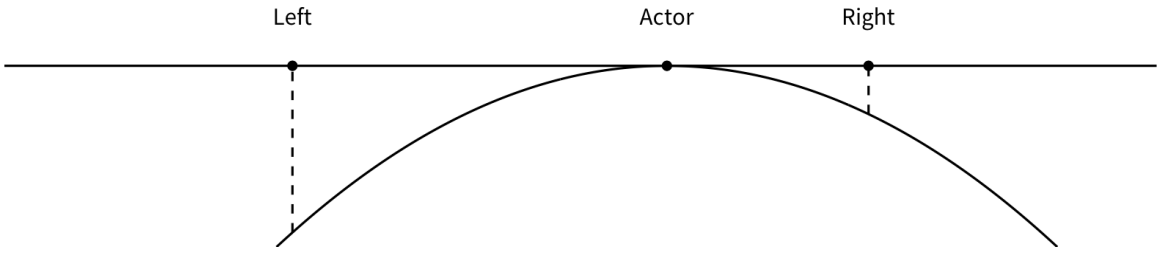


Figure 2.2: A representation of quadratic utility loss over policy choices

It is important to remember that Figure 2.2 shows only the deterministic component of choice task j; random error components u_{ij}^{R} and u_{ij}^{L} are omitted. With idiosyncratic utility error incorporated, Equation (2.2) implies that even though the Actor's distance to Right is

smaller than her distance to Left, there remains a nonzero probability that i chooses Left. This probability depends on the instantiated values of the idiosyncratic 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:

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

The intuition for Equation (2.3) is that the Actor will choose the policy alternative that is nearest to her *unless* idiosyncratic or non-policy factors overcome her ideological considerations. Supposing that the Actor is closer to Right than to Left, $\left(\theta_i - L_j\right)^2$ will be greater than $\left(\theta_i - R_j\right)^2$, capturing the Actor's deterministic inclination to prefer Right over Left. The only way for i to choose Left would be if the idiosyncratic utility of Left over Right exceeded the Actor's deterministic inclinations.

Equation (2.3) can be rearranged to reveal an appealing functional form for i's choice probability. First, expand the polynomial terms on the left side of the inequality…

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

From here, there are two factorizations that reveal convenient expressions for important

constructs in the model.

$$\begin{aligned}\Pr\left(y_{ij}=1\right) &= \Pr\left(2\theta_i R_j - 2\theta_i L_j + L_j^2 - R_j^2 > u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right) \\ &= \Pr\left(2\theta_i R_j - 2\theta_i L_j + \left(R_j - L_j\right)\left(R_j + L_j\right) > u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right) \\ &= \Pr\left(2\left(R_j - L_j\right)\left(\theta_i - \frac{R_j + L_j}{2}\right) > u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right)\end{aligned} \tag{2.5}$$

The first manipulation is to decompose $L_j^2 - R_j^2$ into the two factors $\left(R_j - L_j\right)\left(R_j + L_j\right)$. The second manipulation is to factor $2\left(R_j - L_j\right)$ out of the left-side of the inequality. We perform these manipulations because the resulting terms are appreciably more interpretable than before. First, note that $\frac{R_j+L_j}{2}$ is a formula for the midpoint between the Left and Right locations. This means that the expression $\theta_i - \frac{R_j+L_j}{2}$ intuitively conveys which policy alternative is closer to the Actor. If the Actor is closer to Right than to Left, θ_i will be greater than the midpoint, and vice versa if she were closer to Left. Second, the $2\left(R_j - L_j\right)$ term captures how far apart the policy alternatives are from one another, increasing as the distance between Right and Left increases. Together, the left side of the inequality succinctly describes the deterministic component of the Actor's ideological choice: is she closer to the Left or Right policy, and by how much?[2]

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

$$\begin{aligned}\Pr\left(y_{ij}=1\right) &= \Pr\left(2\left(R_j - L_j\right)\left(\theta_i - \frac{R_j + L_j}{2}\right) > u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right) \\ &= \Pr\left(\iota_j\left(\theta_i - \kappa_j\right) > \varepsilon_{ij}\right),\end{aligned} \tag{2.6}$$

This results in the "discrimination parameter" $\iota_j = 2\left(R_j - L_j\right)$, the "midpoint" or "cutpoint" parameter $\kappa_j = \dfrac{R_j + L_j}{2}$, and a joint error term $\varepsilon_{ij} = u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}$.[3] Parameterizing the model in

[2] Ansolabehere, Snyder, and Stewart (2001) and Burden (2004) use candidate midpoints as predictors in regression analyses to estimate the impact of candidate ideal points 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. I discuss the connection between this model and the IRT model in the next section.

this way expresses the utility comparison in a simpler, linear form. Similar to Equation (2.5) above, $\theta_i - \kappa_j$ shows how far the Actor is from the midpoint between Left and Right, 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 the following section.

A complete statistical model is obtained by making a parametric assumption for the distribution of ε_{ij}. Assuming that ε_{ij} is a draw from a standard Normal distribution,[4] Equation (2.6) implies a probit regression model for the probability that Actor i chooses Right on choice j:

$$\begin{aligned} \Pr\left(y_{ij} = 1\right) &= \Pr\left(\iota_j\left(\theta_i - \kappa_j\right) > \varepsilon_{ij}\right) \\ &= \Pr\left(\iota_j\left(\theta_i - \kappa_j\right) - \varepsilon_{ij} > 0\right) \\ &= \Phi\left(\iota_j\left(\theta_i - \kappa_j\right)\right), \end{aligned} \tag{2.7}$$

where $\Phi(\cdot)$ is the cumulative Normal distribution function. Many IRT models assume that ε_{ij} follows a standard Logistic distribution, (for example Londregan 1999), resulting in a logistic regression model rather than a probit model.[5] As I show below, the probit model facilitates the group-level model much more easily than the logit model.

[4]This implies that $\mathrm{E}\left(u_{ij}^{\mathrm{L}}\right) = \mathrm{E}\left(u_{ij}^{\mathrm{R}}\right)$ and that $\mathrm{Var}\left(u_{ij}^{\mathrm{L}} - u_{ij}^{\mathrm{R}}\right) = 1$. For a given choice j, imposing a scale restriction on the error variance is not problematic because the ideological scale is latent and can be arbitrarily stretched. Any non-unit variance for item j can be compensated for by scaling the discrimination parameter ι_j (i.e. multiplying both sides of the inequality established in Equation (2.6) by some scale factor). The important assumption is that the error variance of a choice j is equal *across individuals*: $s_{ij} = s_j$ for all i.

[5]A technical point of difference between the probit and logit model is the way parameters are scaled to yield the final line of Equation (2.6). If it is assumed that ε_{ij} is a Logistic draw with scale s_j, this implies that $\mathrm{Var}\left(u_{ij}^{\mathrm{L}} - u_{ij}^{\mathrm{R}}\right) = \frac{s_j^2 \pi^2}{3}$, where it is assumed that $s_j = 1$ for the standard Logistic model.

2.2 Item Response Theory

Scholars of ideal point models have noted their similarity to models developed under item response theory (IRT) in psychometrics (for example, Londregan 1999). IRT models have a similar mission as ideal point models: measuring latent features in the data given individuals' response patterns to various stimuli. The canonical psychometric example is in education testing, where a series of test questions is used to measure a student's latent academic "ability" level. This section connects ideal point models to IRT in order to explain their important theoretical and mathematical intuitions.

2.2.1 Latent Traits

The first important feature to note about IRT models is that they are *measurement models*. The goal of a measurement model is to use observed data $\mathbf{y}$ to estimate some construct of theoretical interest $\boldsymbol{\theta}$, supposing that there is a distinction between the two. 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 directly observed. We can represent a measurement model with general notation $\mathbf{y} = f(\boldsymbol{\theta}, \boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ represents some vector of auxiliary model parameters to be estimated in addition to $\boldsymbol{\theta}$ by fitting the model to observed data.

In an educational testing context, students take standardized tests intended to measure their academic "ability" levels. Analysts who score the tests cannot observe a student's ability directly—it is unclear how that would be possible. They do, however, observe the student's answers to known test questions. IRT models provides a structure to infer abilities from the student's pattern of test answers. The context of policy choice is similar. It is impossible to observe any individual's political ideology directly, but we theorize that it affects their responses to survey items about policy choices. The IRT setup lets us summarize an individual's policy preferences by analyzing the structure of their responses to various policy

choices.

It is crucial to note that the only way to estimate a latent construct from observed data is for the model to impose assumptions about the functional relationship between the latent construct and the observed data. In this sense, the estimates can be sensitive to the model's assumptions. While this is always important to acknowledge, it is also valuable to note that model-dependence is an ever-present consideration even for simpler measurement strategies, such as an additive index that sums or averages across a battery of variables. In fact, additive indices are special cases of measurement model where key parameters are assumed to be known and fixed, which is problematic if there is any reason to suspect that item responses are correlated across individuals.[6] In this way, measurement models *relax* the assumptions of simpler measurement strategies, even if the underlying mathematics are more intensive.

2.2.2 Item Characteristics and Item Parameters

Measurement models relax assumptions about the data's functional dependence on the construct of interest. Item response theory focuses this effort on the items to which subjects respond. Different items may reveal different information about the latent construct; the design of the model governs how those item differences can manifest (see Fox 2010 for a comprehensive review of IRT modeling).

Consider a simple model where a 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:

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

[6] Midpoint and discrimination parameters would be sources of this correlation. Additive indices are similar to a model where all all midpoint and discrimination are respectively equal to 0 and 1 by assumption.

where θ_i is scaled such that the probability of a correct/conservative response is 0.5 at $\theta_i = 0$. This model makes the implicit assumption that knowing θ_i is sufficient to produce exchangeable response data; there are no systematic differences in the difficulty level of the test questions or the ideological nature of the policy choices that would affect the propensity of subjects to answer correctly/conservatively on average. This implicit assumption is often unrealistic. Just as some test questions are naturally more difficult than others, some policy questions present more extreme or lopsided choices than others, leading citizens with otherwise equivalent θ values to vary systematically in their response probability across items. Although the "ability-only" model seems unrealistic when posed as such, political science is replete with additive measurement scales that omit all item-level variation: indices of policy views, the racial resentment scale, survey-based scales of political participation, and more.

Rather than assume that all items behave identically for all individuals, IRT explicitly models the systematic variation at the item level using *item parameters.* IRT models have different behaviors based on the parameterization of the item effects in the model. The simplest IRT model is the "one-parameter" model,[7] which includes an item-specific intercept κ_j.

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

IRT parlance refers to the κ_j parameter as the item "difficulty" parameter. In the testing context, if a student has higher ability than the difficulty of the question, the probability that they answer the test item correctly is greater than 0.5. This probability goes up for students with greater ability relative to item difficulty, and it goes down for items with greater difficulty relative to student ability. In a policy choice context, the difficulty parameter is better understood as the "cutpoint" parameter, the midpoint between two policy choices where the respondent is indifferent between the choice of Left or Right on item j. These

[7] One-parameter logit models are often called "Rasch" models, whereas their corresponding probit models are often called "Normal Ogive" models (Fox 2010).

cutpoints are allowed to vary from item to item; some policy choices present alternatives that are, on average, more conservative or liberal than others. For instance, the choice of *how much* to cut capital gains taxes will have a more conservative cutpoint than a question of whether to cut capital gains taxes at all. If there were no systematic differences across items, it would be the case that $\kappa_j = 0$ for all j, and the one-parameter model would reduce to the simpler model in Equation (2.8).

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.

$$\Pr\left(y_{ij} = 1\right) = \Phi\left(\iota_j\left(\theta_i - \kappa_j\right)\right), \tag{2.10}$$

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.[8]

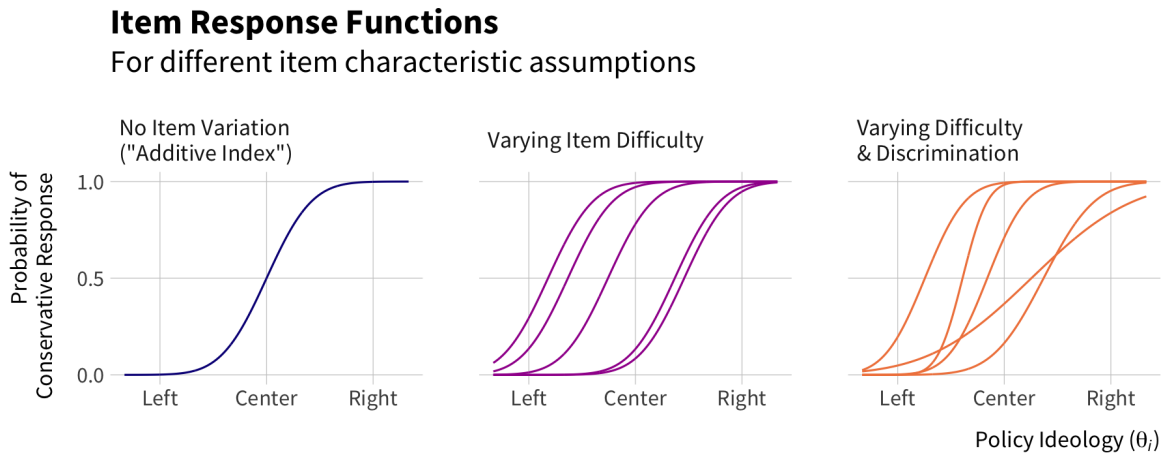


Figure 2.3: Examples of item characteristic curves under different item parameter assumptions

[8]Two-parameter IRT models are sometimes written with ι_j is distributed through the equation: $\iota_i\theta_i + \alpha_j$, where $\alpha_j = \iota_j\kappa_j$. Although this parameterization more closely follows a linear slope-intercept equation, it loses the appealing interpretation of κ_j as the midpoint between policy choices.

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 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 convey which value of θ yields a correct response with probability 0.5, but they do not affect the *elasticity* of the item response function to changes in θ. 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.

2.2.3 IRT Interpretation of the Ideal Point Model

How do we interpret our statistical model of ideal points in light of item response theory? Recall the statistical model that we derived from the utility model above. An Actor i faces policy question j, with a Right alternative located at R_j and a Left alternative located at L_j. The Actor chooses the alternative closest to her ideal point θ_i, subject to idiosyncratic utility shocks summarized by ε_{ij}. Letting y_{ij} indicate the outcome that Actor i chooses the Right position on policy question j, the probability that $y_{ij} = 1$ is given by the two-parameter model in Equation (2.10) (or (2.6) above).

The behavior of the item parameters can be understood by remembering that they are functions of the Left and Right choice locations. For instance, the cutpoint parameter κ_j represents an intercept shift for an items response function and is equal to $\frac{R_j + L_j}{2}$. Suppose that $\theta_i - \kappa_j = 0$, which occurs if the item cutpoint falls directly on an Actor's ideal point. In

such a case, the Actor would be indifferent (in expectation) to the choice of Left or Right, and the probability of choosing Right would collapse to 0.5.[9] 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.[10] 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 behaves as a "coefficient" on the distance between the Actor ideal point and the cutpoint, meaning that the Actor's choice is more elastic to her policy preferences as ι_j increases.[11] Because $\iota_j = 2(R_j - L_j)$, the discrimination parameter grows when the distance between the Right and Left alternatives grows larger, which happens when R_j increases or L_j decreases.

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.

2.2.4 IRT in Political Science

A section reviewing IRT in political science:

[9] This holds in logit and probit models, since $\text{logit}^{-1}(0)$ and $\Phi(0)$ are both equal to 0.5.

[10] 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 ι_j is constrained to be greater than or equal to zero

[11] 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.

- ideal points (Poole and Rosenthal, CJR, Londregan, Jeff Lewis, Michael Bailey, Martin and Quinn)
- citizen ideology with survey data (Seth Hill multinomial and individual, Tausanovitch and Warshaw small area, Caughey and Warshaw groups within areas)
- other latent modeling (Levendusky, Pope, and Jackman 2008)

2.3 Modeling Party-Public Ideology in Congressional Districts

This section outlines my group-level ideal point model for party publics. It begins by describing the connection between the individual-level IRT model and the group-level model and its implication for the parameterization of the model. I then lay out the hierarchical model for party-public ideal points in its static form (Section 2.3.1) and its dynamic form (Section 2.4). I discuss technical features of model implementation, including choices for model parameterization, model identification, prior distributions, and model testing methods such as prior predictive checks and posterior predictive checks.

So far we have modeled individual responses to policy items according to their own individual ideal points, but this project is concerned with the average ideal point of a *group* of individuals. In the group model, we assume that individual ideal points are distributed within a group g, where groups are define as the intersection of congressional districts d and political party affiliations p.

As before, we observe a binary response from individual i to item j, which we regard as a probabilistic conservative response with probability π_{ij}, which is given a probit model.

$$y_{ij} \sim \text{Bernoulli}(\pi_{ij}) \tag{2.11}$$

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

Following Fox (2010) and 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 $\sigma_j = \iota_j^{-1}$ and rewriting the model as

$$\pi_{ij} = \Phi\left(\frac{\theta_i - \kappa_j}{\sigma_j}\right). \tag{2.13}$$

Caughey and Warshaw (2015) describe the dispersion parameter as introducing "measurement error" in π_{ij} beyond the standard Normal utility error from ε_{ij} above.

The group model begins with the notion that there is a probability distribution of ideal points within a group g, where a "group" is a partisan constituency within a congressional district. Supposing that individual deviations from the group mean are realized by the accumulation of a large number of random forces, we can represent an individual ideal point as a Normal draw from the group,[12]

$$\theta_i \sim \text{Normal}\left(\bar{\theta}_{g[i]}, \sigma_{g[i]}\right) \tag{2.14}$$

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

While it is possible to continue building the model hierarchically from (2.14), 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 rewrite the model by aggregating individual-level survey response data to the group level, expressing the grouped outcome data as a function of the group parameters. Let $s_{gj} = \sum_{i \in g}^{n_{gj}} y_{ij}$, 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. Supposing these trials were collected independently across groups and items (an assumption that is relaxed later),

[12]Notation for Normal distributions will always describe the scale parameter in terms of standard deviation σ instead of variance σ^2. This keeps the notation consistent with the way Normal distributions are expressed in Stan code.

we could model the grouped outcome as a binomial random variable,

$$\begin{aligned} s_{gj} &\sim \text{Binomial}\left(n_{gj}, \bar{\pi}_{gj}\right) \\ \bar{\pi}_{gj} &= \Phi\left(\frac{\bar{\theta}_g - \kappa_j}{\sqrt{\sigma_g^2 + \sigma_j^2}}\right), \end{aligned} \tag{2.15}$$

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 any random individual's ideal point relative to the group mean is included in the model as group-level variance term. If individual ideal points are Normal within their group, this within-group variance can simply be added to the probit model as another source of measurement error, with larger within-group variances further attenuating $\bar{\pi}_{ij}$ toward 0.5. Caughey and Warshaw (2015) derive this result in the supplementary appendix to their article.

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 (2.14)), I implement a weighting routine that downweights information from repeated-subject observations while adjusting for nonrepresentative sample design, as I will describe in Section 2.4.2.

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, a single survey will not contain a representative sample of all congressional districts, and

certainty not a representative sample of partisans-within-districts. I specify a hierarchical model for the group parameters in order to stabilize the estimates in a principled way. The hierarchical model learns how group ideal points are related to cross-sectional (and eventually, over-time) variation in key covariates, borrowing strength from data-rich groups to stabilize estimates for data-sparse groups, and even imputing estimates for groups with no survey data at all. This section describes the multilevel structure using traditional notation for hierarchical models; later in Section 2.5 I describe how I parameterize the model for the estimation routine.

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 $d[g]$, or the g^{th} value of the vector $\mathbf{d}$. Similarly, g's party is $p[g]$. For higher levels such as g's state, I write $s[g]$ as shorthand for the more-specific but more-tedious $s[d[g]]$.

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 the distribution of groups whose hypermean is predicted by a regression on geographic-level data with parameters that are indexed by political party. This regression takes the form

$$\bar{\theta}_g \sim \text{Normal}\left(\mu_{p[g]} + \mathbf{x}_{d[g]}^{\top}\beta_{p[g]} + \alpha^{\texttt{state}}_{s[g]p[g]}, \sigma_p^{\texttt{group}}\right) \tag{2.16}$$

where $\mu_{p[g]}$ is a constant specific to party p,[13] $\mathbf{x}_d$ is a vector of congressional district-level covariates with party-specific coefficients β_p. State effects $\alpha^{\texttt{state}}_{sp}$ are also specific to each party. The benefit of specifying separate parameters for each party is that geographic features (such as racial composition, income inequality, and so on) may be related to ideology in ways that are not identical across all parties. This is an important departure from the structure laid out

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

by Caughey and Warshaw (2015), which estimates the same set of geographic effects for all groups in the data.

The state effects are regressions on state features as well,

$$\alpha_{sp}^{\texttt{state}} \sim \text{Normal}\left(\mathbf{z}_s^\top \gamma_p + \alpha_{r[s]p}^{\texttt{region}}, \sigma_p^{\texttt{state}}\right), \tag{2.17}$$

where state-level covariates $\mathbf{z}_s$ have party-specific coefficients γ_p. Each state effect is a function of a party-specific region effect $\alpha_{[s]rp}^{\texttt{region}}$ for Census regions indexed r, which is a modeled mean-zero effect to capture correlation within regions.

$$\alpha_{rp}^{\texttt{region}} \sim \text{Normal}\left(0, \sigma_p^{\texttt{region}}\right) \tag{2.18}$$

2.4 Dynamic Model

lol tbd

2.4.1 Identification Restrictions

Ideal point models, as with all latent space models, are unidentified without restrictions on the policy space. The model as written can rationalize many possible estimates for the unknown parameters, with no prior basis for deciding which estimates are best. A two-parameter model such as this requires some restriction on the polarity, location, and scale of the policy 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 latent scale could be reversed. We could flip the sign of every ideal point, and the response probability would be unaffected if we also flip the sign of every item parameter.

These properties are present with every statistical model, but covariate data typically provide the restrictions necessary to identify a model.[14] Because the response probability is a function of the interaction of multiple parameters in a latent space, however, data alone do not provide the necessary restrictions on the space to provide a unique solution. Absent any natural restriction from the data, I provide my own restrictions on the polarity, location, and scale of the policy space.

The polarity of the space is fixed by coding all items such that conservative responses are 1 and liberal responses are 0. This ensures that increasing values on the link scale always lead to an increasing probability of a *conservative* item response. Additionally I impose a restriction that all discrimination parameters are positive, which implies that shifting any ideal point farther to the *right* of an item cutpoint increases the probability of a conservative response, all else equal.

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

$$\kappa_j = \tilde{\kappa}_j - \frac{\sum_{j=1}^{J} \tilde{\kappa}_j}{J}, \tag{2.19}$$

which is performed in every iteration of the sampler. This restriction on the sum of the cutpoint parameters also implies a restriction on the mean of the cutpoints, since $\frac{0}{J} = 0$.

Lastly, I set the scale of the latent space by restricting the product of the J discrimination parameters to be equal to 1. I implement this by restricting the log discrimination parameter to

[14] We could imagine shifting, stretching, or reversing the sign of a covariate to reveal the same mathematical behaviors. All of these transformations would result in the same predictions as long as the parameters are also transformed to compensate.

have a sum of 0, which achieves an equivalent transformation.[15] Letting $\tilde{\iota}_j$ be the unrestricted discrimination parameter, we obtain the restricted ι_j as follows.

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

$$\log\left(\iota_j\right) = \log\left(\tilde{\iota}_j\right) - \frac{1}{J}\sum_{j=1}^{J}\log\left(\tilde{\iota}_j\right) \tag{2.21}$$

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

2.4.2 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. This section describes an approach for weighting group-level data that adjusts for both issues. The corrections are lifted from Caughey and Warshaw (2015) with slight modifications.

First, the sample size in each group-item "cell" gj 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

$$n^*_{g[i]j} = \sum_{i=1}^{n_{g[i]j}} \frac{1}{r_i d_{g[i]}}, \tag{2.22}$$

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

[15] A quick demonstration using three unknown values a, b, and c. If $a \times b \times c = 1$, then $\log(a) + \log(b) + \log(c) = \log(1)$, which is equal to 0.

from groups that contain greater variation in their survey design weights. It is defined as

$$d_{g[i]} = 1 + \left(\frac{\text{sd}_{g[i]}\left(w_i\right)}{\text{mean}_{g[i]}\left(w_i\right)} \right)^2, \tag{2.23}$$

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

$$\bar{y}^*_{g[i]j} = \frac{\sum_{i=1}^{n_{g[i]j}} \frac{w_i y_{ij}}{r_i}}{\sum_{i=1}^{n_{g[i]j}} \frac{w_i}{r_i}}. \tag{2.24}$$

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

$$s^*_{gj} = \min\left(n^*_{gj}\bar{y}^*_{gj}, n^*_{gj}\right). \tag{2.25}$$

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. Ordinarily this would be a problem for modeling a Binomial random variable, since a Binomial is an integer-valued count of successes given an integer number of trials and a success probability. For this reason, many statistical programs will return an error if a floating-point argument is bassed to the Binomial probability mass function. Whereas Caughey and Warshaw (2015) calculate

$\lceil n^*_{gj} \rceil$ and $\lfloor s^*_{gj} \rceil$ to obtain integer data for estimation, I instead implement a custom Binomial quasi-likelihood function that returns log probabilities for weighted data (see Section 2.5.2).

2.5 Bayesian Estimation and Computation

I implement the model using Stan, a programming language for high-performance Bayesian analysis that extends and interfaces with `C++` (Carpenter et al. 2016). Stan implements an adaptive variant of Hamiltonian Monte Carlo (HMC), an algorithm that efficiently collects posterior samples by "surfing" a Markov proposal trajectory along the gradient of the posterior distribution. Because the algorithm uses the posterior gradient to generate proposals, the algorithm concentrates proposals in regions of high transition probability and performs better in high dimensions than conventional Gibbs sampling algorithms. Although it possible to estimate Stan models using front-end software packages such as `brms` for R (Bürkner and others 2017), complicated models must be programmed with raw Stan code, which can be intensive. This section describes instances where the model *as programmed in Stan* departs from the model *as written* above. Although these alterations do not change the statistical intuitions of the model, they are essential for the model's computational stability by protecting against biased MCMC estimation and floating point arithmetic errors. These contributions should be highlighted here because they are crucial to ensuring valid inferences and are substantive improvements on previous software implementations of group-level IRT models.

2.5.1 Non-Centered Parameterization

Hierarchical models have posterior distributions whose curvatures present difficulties for sampling algorithms (Betancourt and Girolami 2015; Papaspiliopoulos, Roberts, and Sköld 2007). To improve the estimation in Stan, I program the hierarchical models using a "non-centered" parameterization rather than a "centered" parameterization. Whereas the centered

parameterization considers $\bar{\theta}_g$ as a random draw from a hierarchical distribution (Equation (2.16) above), the non-centered parameterization defines $\bar{\theta}_g$ as a deterministic function of its conditional hypermean and a random variable.

$$\bar{\theta}_g = \mu_{p[g]} + \mathbf{x}_{d[g]}^{\top}\beta_{p[g]} + \alpha_{s[g]p[g]}^{\text{state}} + u_g\tau_{p[g]}, \tag{2.26}$$

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

where $u_g\tau_{p[g]}$ behaves as a group-level error term. It is composed of a standard Normal variate u_g and a scalar parameter τ_p that controls the scale of the error term. The non-centered model is algebraically equivalent to centered model in the likelihood, but it factors out (or "unnests") the location and the scale from the hyperprior. The non-centered parameterization improves MCMC sampling by de-correlating the parameters that compose the hierarchical distribution. Hierarchical models using a centered parameterization, on the other hand, are vulnerable to estimation biases due to poor posterior exploration (Betancourt and Girolami 2015).[16] This is a crucial extension to the estimation approach developed by Caughey and Warshaw (2015), whose model implements all hierarchical components using the centered parameterization.

Equation (2.27) is an incomplete implementation of the non-centered form; to complete the parameterization, I apply it too all hierarchical components in the regression, including

[16] Stan's HMC algorithm is programmed to diagnose poor posterior exploration by detecting "divergent transitions" during sampling. Because Stan's HMC algorithm uses the gradient of the posterior distribution to propose efficient transition trajectories through the parameter space, it adaptively builds expectations about the probability density of the next Markov state. Areas of high curvature in the posterior gradient can lead to "divergences" in the HMC algorithm: transitions where the log density of a state differs substantially from what Stan anticipated when it proposed the transition. Markov chains with many divergent transitions have a high risk of being severely biased, since the divergences indicate that the Markov chain is failing to efficiently navigate the parameter space (Betancourt and Girolami 2015). The non-centered parameterization smooths out these problematic regions of posterior density, safeguarding against biased MCMC estimates.

the state and region effects.

$$\begin{aligned}\bar{\theta}_g = \mu_{p[g]} + \mathbf{x}_{d[g]}^{\top}\beta_{p[g]} + u_g^{\text{group}}\tau_{p[g]}^{\text{group}} \\ + \mathbf{z}_{s[g]}^{\top}\gamma_p + u_{s[g]p[g]}^{\text{state}}\tau_{p[s]}^{\text{state}} \\ + u_{r[g]p[g]}^{\text{region}}\tau_{p[g]}^{\text{region}}\end{aligned} \tag{2.28}$$

The full model places the hypermean regressions and error terms for groups, states, and regions in one deterministic equation. It contains "error terms" for each level of hierarchy—groups, states, and regions—where all parameters are indexed by party.

2.5.2 Log Likelihood for Weighted Data

One important implementation consideration for the group IRT model is the presence of weighted, non-integer response data. As described in Section 2.4.2, grouped data require reweighting to account for nonrepresentative sample designs and repeated observations within individual members of a group. The resulting data are likely to take non-integer values, which would cause the built-in Binomial likelihood function to fail. Whereas Caughey and Warshaw (2015) round their data to conform to Stan's Binomial likelihood function, my approach rewrites the likelihood function to accept non-integer data. This allows me to maintain the precision in the underlying data while still correcting the issue at hand.

To explain how this works in Stan, some context on Bayesian computation is helpful. It is usually sufficient for Bayesian estimation with Markov chain Monte Carlo to calculate the posterior density of model parameters only up to a proportionality constant,

$$p(\Theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \Theta)p(\Theta), \tag{2.29}$$

where Θ and $\mathbf{y}$ generically represent parameters and observed data. For computational stability, especially in high dimensions where probability densities get very small, these

calculations are done on the log scale.

$$\log p(\Theta \mid \mathbf{y}) \propto \log p(\mathbf{y} \mid \Theta) + \log p(\Theta), \tag{2.30}$$

MCMC algorithms calculate the right-side of this proportionality at each iteration of the sampler to decide if proposed parameters should be accepted into the sample or rejected. In Stan, this calculation is passed to the *log density accumulator*, a variable containing the sum of the log likelihood and log prior density at every sampler iteration (Carpenter et al. 2016).

In the current case, it is the probability of the data $\log p(\mathbf{y} \mid \Theta)$ that presents a problem. The Binomial log likelihood function as written in Stan will not accept non-integer data, so I rewrite the kernel $K(\cdot)$ of the Binomial log likelihood:

$$\log K\left(p\left(s^*_{gj} \mid \bar{\pi}_{gj}\right)\right) = s^*_{gj} \log \bar{\pi}_{gj} + \left(n^*_{gj} - s^*_{gj}\right) \log\left(1 - \bar{\pi}_{gj}\right) \tag{2.31}$$

where the weighted number of trials n^*_{gj} and the weighted number of successes s^*_{gj} can take non-integer values. This is the same approach that Ghitza and Gelman (2013) take in a frequentist maximum likelihood context.[17] I pass the results of Equation (2.31) directly to the log density accumulator. This maneuver accumulates the log probability of the response data, which Stan uses to evaluate the acceptance probability of MCMC samples. Other sampling statements that add prior density to the accumulator are unaffected by my approach to the log likelihood.

2.5.3 Optimized IRT Model

The matrix expansion trick would go here, if I did it.

[17] They describe this as "simply the weighted log likelihood approach to fitting generalized linear models with weighted data" (Ghitza and Gelman 2013, 765).

2.5.4 Prior Distribution

The Bayesian modeling paradigm requires 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. This enables the researcher to stabilize parameter estimates and downweight unreasonable estimates, enabling the researcher to guard against overfitting and smooth estimates from similar groups. This is especially valuable in data-sparse settings where parameters are unidentified or weakly identified, such as in hierarchical models where some groups contain more data than others (Gelman and Hill 2006). Bayesian estimation has fundamental computational advantages for ideal point models, since MCMC generates posterior samples of latent variables just as it would for any other model parameter. This allows the researcher to escape certain pathologies of optimization algorithms in high-dimensional parameter spaces with many incidental nuisance parameters (Lancaster 2000, @clinton-jackman-rivers:2004:ideal). The drawback of Bayesian modeling is that prior specification is additional work for the researcher, which can be complicated especially in situations where a model is sensitive to the choice of prior.

This section provides describes and justifies priors used in the group ideal point model. The discussion here is more detailed than in a typical paper describing a Bayesian ideal point model for several reasons. Firstly, 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). These prior checks allow researchers to explore and demonstrate the consequences of their prior choices in transparent ways, but most Bayesian analyses in political science lack these explicit prior checks. Authors often declare their prior choices without explicitly justifying these choices, which can make prior specification feel opaque or even arbitrary to non-Bayesian readers. Secondly, and more specifically to this project, the nonlinearities introduced by a probit model present particular challenges for specifying

priors. Some of my choices depart from those in previous Bayesian ideal point models for important theoretical and practical reasons that I explain below. Thirdly, model parameterization is important for effective Bayesian computation (see Section 2.5.1), and although reparameterization does not affect the likelihood of data given the parameters, parameterization naturally affects the choice of priors. This exploration of priors is a crucial component of the model-building process for this project and is uncommon in other Bayesian works in political science, so it is important to justify these choices with sufficient detail.

Before discussing priors for the group ideal point model, it is helpful to discuss some general principles for working with prior distributions. They are not *universal* principles, but they are *theoretical* in the sense that they provide a pre-data orientation for prior distributions. They are heuristic principals in the sense that they provide powerful shortcuts to good analysis decisions based on lightweight signals about the problem at hand.[18]

This discussion of priors begins with the orientation laid out by Gelman, Simpson, and Betancourt (2017) that "the prior can often only be understood in the context of the likelihood." Although prior information is generally regarded as information that a researcher has before encountering data—and therefore before making any modeling decisions—in practice it is often the case that priors are chosen with reference to a specific analysis model. For example, we may have prior expectations about the proportion of Republicans who express conservative preferences on a given policy question (e.g. that the proportion is most likely above 50 percent), but if we model the proportion with a probit model, we typically specify priors on regression coefficients rather than the proportion parameter itself. This means that researchers must consider their priors as embedded in the specific data model at hand. This further implies that the *parameterization* of a model can affect the researcher's prior for some ultimate quantity of interest, even if the parameterization does not affect the likelihood of

[18] Gelman (2017) holds that theoretical statistics ought to be the "theory of applied statistics," in the sense that statistical theory ought to be informed by "what we actually do" and should thus work to formalize aspects of workflow that begin merely as "good practice."

the data given the parameters(Gelman 2004). The consequences of model parameterization are explored further in Chapter 3.

- weak information
  - between structural and regularizing. (gabry et al?)
  - They achieve regularization by encoding structural information, plus information befitting a *class* of problems. Short of information tailored to a *specific* problem
- families of priors
  - less reliance on the actual prior param values
  - features of a CDF
  - entropy in the distribution
  - shape of the log probability (normal vs T)
- prediction-focused (gelman et al prior/likelihood)

Subjective v. objective

- the actual degree of belief is zero
- not really the issue
- priors provide practical stability

General thoughts about priors

- WIPS and the evolution of thinking about this
  - likelihood is important weak information
  - constraining values to what is reasonable
  - but not so informed that we're downweighting reasonable values unless when context demands

    * sometimes it does, like identifiability, strong regularization/separation

- parameterization
    - normal(a, b) or a + bu, u ~ normal(0, 1)

Throughout this discussion is a common underpinning that the data model itself provides important structure for choosing priors. This is a pragmatic view. While many discussions of priors focus on the fact that they are philosophically essential for posterior inference, this pragmatic view emphasizes their practical implications for regularization, model stabilization, and computational efficiency and tractability.

2.5.5 Understanding the Probit Model

- priors and likelihoods
    - probit is a nonlinear model
    - data aren't additively related to y, but some other thing
    - the prior is specified in reference to some likelihood

How the probit model works

- there is a latent scale
- covariates typically have linear, additive effects
- error is assumed Normal
- the latent scale is identified as having location zero and scale 1
- this means the predicted probability of a 1 is the probability that the index > 0, which is equivalent to the normal CDF at the predicted index value
- coefficients are uncertain, so uncertainty in posterior data are owed to probabilistic uncertainty about y given an index value, and baseline uncertainty about location on the index given covariates

How Bayesian modeling with probit works:

- because we know the Normal distribution, we know the region of quantile space that reasonably produces our outcome data
- combination of data and parameters shouldn't realistically lead us beyond the quantiles that produce probabilities between 1% and 99% (justify)
- it isn't crazy that some of our predictions are highly determined, so we don't want to be too restrictive, but broadly speaking, a priori, you know (justify better)

Divergence from past work

- Vague priors:
  - Clinton, Jackman, Rivers
  - Treier and Hillygus
  - Tausanovitch and Warshaw
- Caughey and Warshaw
  - noncentering
  - lognormal parameterization (enable pooling by leveraging transformed parameters)

2.5.6 Item Parameters

I specify priors on the unscaled cutpoint and discrimination parameters that are Normal and LogNormal, respectively. In order to model their joint distribution, I specify a multivariate Normal distribution for the cutpoint and logged discrimination parameter,

$$\begin{bmatrix} \tilde{\kappa}_j \\ \log\left(\tilde{\iota}_j\right) \end{bmatrix} \sim \text{Normal}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{2.32}$$

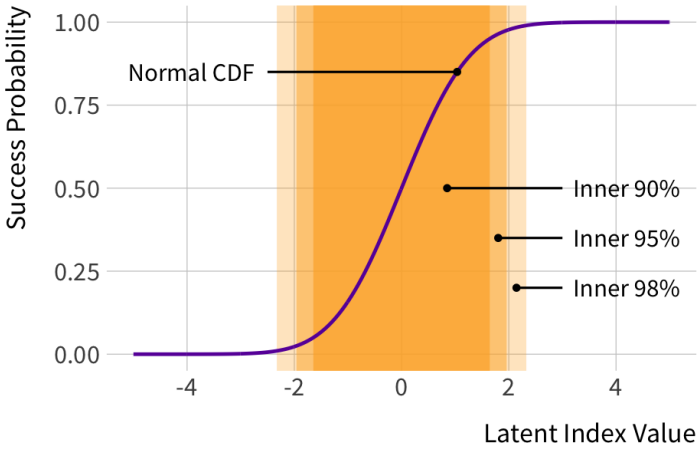


Figure 2.4: The region of the probit model's latent index that maps to response probabilities between 1 and 99 percent.

where $\boldsymbol{\mu}$ is a 2-vector of means and $\boldsymbol{\Sigma}$ is a 2×2 variance–covariance matrix. 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 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. 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.[19]

$$\boldsymbol{\mu} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \tag{2.33}$$

I build a prior for the variance–covariance matrix $\boldsymbol{\Sigma}$ by decomposing it into a diagonal

[19] 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.

matrix of scale terms and a correlation matrix. First I factor out the scale components.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\tilde{\kappa}}^2 & \rho\sigma_{\tilde{\kappa}}\sigma_{\tilde{\iota}} \\ \rho\sigma_{\tilde{\kappa}}\sigma_{\tilde{\iota}} & \sigma_{\tilde{\iota}}^2 \end{bmatrix} \tag{2.34}$$

$$= \begin{bmatrix} \sigma_{\tilde{\kappa}} & 0 \\ 0 & \sigma_{\tilde{\iota}} \end{bmatrix} \boldsymbol{S} \begin{bmatrix} \sigma_{\tilde{\kappa}} & 0 \\ 0 & \sigma_{\tilde{\iota}} \end{bmatrix} \tag{2.35}$$

The resulting matrix $\boldsymbol{S}$ is a 2×2 correlation matrix, meaning it has a unit diagonal and off-diagonal correlation terms (denoted ρ).

$$\boldsymbol{S} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{2.36}$$

I then specify priors for the scale terms, $\sigma_{\tilde{\kappa}}$ and $\sigma_{\tilde{\iota}}$, and the correlation matrix $\boldsymbol{S}$ separately. This approach is also known as a "separation strategy" for covariance matrix priors (Barnard, McCulloch, and Meng 2000).[20] 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 but whose scale is wide enough that the data are likely to dominate the prior. I give $\boldsymbol{S}$ 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).[21]

$$\boldsymbol{S} \sim \text{LKJcorr}(\eta = 2) \tag{2.37}$$

[20] Although inverse-Wishart priors are often chosen for covariance matrices because they ensure conjugacy of the multivariate Normal distribution, recent work by Bayesian statisticians suggests that the separation strategy for covariance matrices is superior. The inverse-Wishart distribution has certain restrictive properties such as prior dependency between scales and correlations (large and small scales imply large and small correlations, respectively) that many Bayesian statisticians find undesirable compared to priors specified using the more flexible separation strategy (Akinc and Vandebroek 2018; Alvarez, Niemi, and Simpson 2014). Furthermore, the conjugacy of the inverse-Wishart is irrelevant for this model because conjugacy does not provide the same computational benefit for Hamiltonian Monte Carlo samplers as it does for Gibbs samplers or analytic posterior computation.

[21] For a matrix $\boldsymbol{S}$ that follows an LKJ distribution with shape parameter η, the density of $\boldsymbol{S}$ is a function of its determinant: $\text{LKJcorr}(\boldsymbol{S} \mid \eta) = c \times \det(\boldsymbol{S})^{\eta-1}$ with proportionality constant c that depends on the dimensionality of $\boldsymbol{S}$.

The LKJ distribution has one shape parameter η, which can be interpreted like a shape parameter for a symmetric Beta distribution. Setting $\eta = 1$ yields a flat prior over all correlation matrices, where increasing values of $\eta 1$ concentrate prior density toward the mode, which is an identity matrix. The chosen value of $\eta = 2$ provides weak regularization against extreme correlations near −1 and +1. Although it would have been sufficient to specify a prior for ρ instead of the entire matrix $\boldsymbol{S}$, this convenience only arises in small (in this case, 2×2) correlation matrices. Larger matrices (such as those that would result from a more complex IRT model specification) would require explicit priors for a larger number of off-diagonal parameters. The LKJ prior can be generally applied to larger correlation matrices, so I choose it for the sake of building a more flexible and extensible model.

Figure 2.5 plots several details of the 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 off-diagonal parameter in the matrix $\boldsymbol{S}$ that controls the covariance of items in the joint prior, generated from the LKJ correlation matrix prior. The bottom panel shows the distribution of item parameter values simulated from the multivariate Normal distribution implied by the joint hierarchical prior. Each point represents a simulated item as a combination of cutpoint values (on the horizontal axis) and log-discrimination values (on the vertical axis). Points are colored according to the number of nearby points, which informally conveys the prior density of items with particular cutpoint and discrimination values.

2.5.7 Ideal Point Parameters

For the hierarchical model that smooths group estimates, the model is parameterized to ease the specification of priors. First I standardize all covariates to have a mean of zero. This ensures that the constant μ_p in the hierarchical model for $\bar{\theta}_g$ (See Equation (2.16)) can be interpreted as a "grand mean" for party p, the average group ideal point for party p when

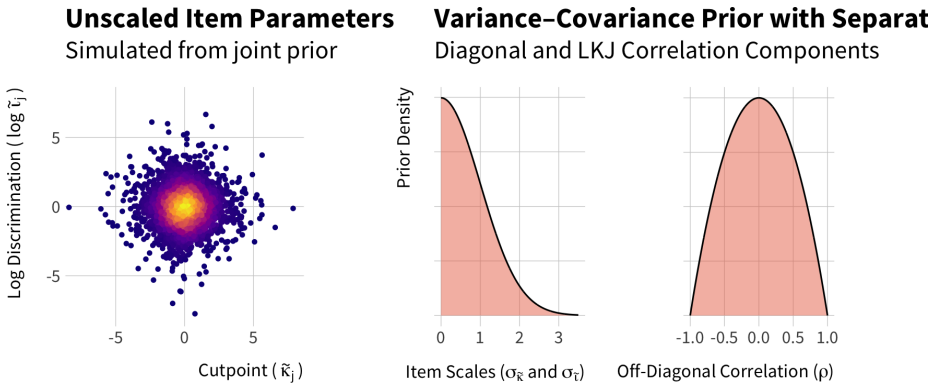


Figure 2.5: Components of the joint hierarchical prior for the unscaled item parameters. Left panel shows prior values for unscaled item parameters from the joint prior. Remaining panels show priors for decomposed covariance matrix components: including the standard deviation that form the matrix diagonal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

all covariates are at their means. I then give this grand mean a Normal $(0, 1)$ prior, which implies a flat prior on the probability scale. Substantively this represents an assumption where the predicted probability of a conservative response for the typical item

average Democratic constituency and the average Republican constituency "could be anything." Because the latent scale is identified by restricting the item parameters, the relaxed prior for the average ideal points prevents the ideal point priors from interfering with the identification of the scale.

I set priors for the coefficients in the hierarchical model by

I give coefficients Normal $(0, 0.5)$ priors. Substantively, this represents a prior where a typical draw, expected to be one standard deviation away from the mean, would change the probability of a conservative response from 0.5 to 0.8413447 (if above) or to 0.3085375 (if below). Constants are given less informative Normal $(0, 1)$ priors, whose density is rather flat when transformed from the link scale to the probability scale, as shown in Figure **??**. Given that 95 percent of the standard Normal distribution falls between the quantiles -1.96 and 1.96, our priors should not give much weight to coefficients large enough to cause the response

probability to leap from one end of that scale to the other.

Within-group standard deviations, as well as the scale parameters in the non-centered error terms (τ), are given LogNormal $(0, 1)$ priors.

2.6 Testing the Model with Simulated Data

2.7 Ideal Point Estimates for District-Party Publics

2.7.1 Data

2.7.2 Posterior Analysis

— 3 —

Bayesian Causal Inference

Before I employ the estimates of party-public ideology obtained in Chapter 2, this chapter discusses a Bayesian view of causal inference. This framework addresses two major themes in the empirical problems that I confront later in the project, as well as several other minor themes.

First, this project views causal inference as a problem of posterior predictive inference. Causal models are tools for inferences about missing data: what we would observe if a treatment variable were set to a different value. The unobserved data are "unobserved potential outcomes" in the Rubin causal framework or "counterfactual outcomes" in the Pearl framework (Pearl 2009; Rubin 2005). Causal inference can be Bayesian if the target of interest is the probability distribution of unobserved potential outcomes (or the probability distribution of any causal estimand), conditioning on the observed data. In this chapter, I will argue that this is what researchers are implicitly trying to obtain, even if implicitly, nearly all of the time.

Second, the Bayesian approach incorporates uncertainties about the key independent variable in this project. Causal estimands (to use the Rubin terminology) are comparisons of potential outcomes under two hypothetical values of a treatment, usually a unit's outcome

under an *observed* treatment versus an *unobserved* treatment. The data in this project frustrate the typical structure of a causal estimand because the treatment value of interest—policy ideology in a district-party public—is not observed. Instead, it is estimated up to a probability distribution specified by the measurement model in Chapter 2. Uncertainty about the effect of setting district-party ideology to some new value θ' therefore contains multiple sources of uncertainty: statistical uncertainty about the estimated causal effect, and measurement uncertainty about the original value of θ before a causal intervention. Bayesian analysis provides the statistical machinery to quantify uncertainty in these causal effects as if they were any other posterior quantity, by marginalizing the posterior distribution over any nuisance parameters.

This chapter unpacks these issues according to the following outline. First, I review 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, conditional on observed data. Because Bayesian modeling remains largely foreign to political science, 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. Finally, I provide examples of Bayesian causal modeling by replicating and extending published studies in political science, showing where priors add value to causal inference.

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 distinguishes the current causal inference movement from other informal approaches. This section reviews causal inference by breaking 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 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

[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.

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; no causal effects are ever simply 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, $Y_i\left(A_i\right)$. 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* causal identification of causal estimands (Keele 2015). The resulting 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. There is often no guarantee that linear regression models, or any parametric models, will correctly model the data and recover causal effects.

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 2017; 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 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 π, 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,

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

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, $p\left(\pi\right)$. 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 data and parameters are both simply 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 feels unscientific. This is because the marginal parameter distribution often represents "prior information" about which parameter values are plausible before conditioning on observed data. The inside view demystifies priors by acknowledging that a prior and a likelihood are the same thing: using a probability distribution to quantify uncertainty about the exact value of a model variable (McElreath 2017b). As data are collected, these probability distributions collapse to known values, in the case of observed data, or concentrate on a narrower range of unknown parameter values that can rationalize the observed data. In the extreme, the only way to make inferences about unobserved data from observed data is by through prior assumptions about the relationship between observed and unobserved variables (Lemm 1996).

This is similar to the fundamental problem of causal inference: causal judgments are never possible without assumptions that are external to the data.

Bayesian updating, From the inside view, means considering a multitude of model configurations and judging which configurations are consistent or inconsistent with the data. The joint prior model, $p(\mathbf{y} \mid \boldsymbol{\pi})p(\boldsymbol{\pi})$ describes an overly broad set of possible assumptions about the world. These assumptions cover a distribution of possible parameters, $p(\boldsymbol{\pi})$, and possible data given parameters, $p(\mathbf{y} \mid \boldsymbol{\pi})$. 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, since there is no formal notion of "plausible parameters given the data" without conditioning parameters on data, which necessitates prior and posterior distributions. As it connects to causal inference, this means there is no formal notion of

"plausible causal effects" 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 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. Suppressing the

unit index i,

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

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 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. We let the systematic component $f(\cdot)$ be a plug-in estimator for $\mathbb{E}\left[Y \mid A = a, X = x\right]$. 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, $\hat{\bar{\tau}}(X = x)$, by differencing these model predictions over the treatment.

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

Where the second line includes $\hat{\bar{\tau}}$ to reflect the fact that $f(\cdot)$ is an estimator of $\bar{\tau}$.

The Bayesian approach, inspired largely by Rubin (1978a), confronts the problem with a joint model for data and parameters: $p(Y, \boldsymbol{\pi}) = p(Y \mid f(A, U, \boldsymbol{\pi}))\, p(\boldsymbol{\pi})$. The data are distributed conditional on the statistical model prediction $f(\cdot)$, which conditions on the model parameters $\boldsymbol{\pi}$. The parameters also have a prior distribution $p(\boldsymbol{\pi})$, 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 prior 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),

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

which is the probability distribution of model contrasts for $A = 1$ versus $A = 0$. This distribution of model contrasts contains two sources of uncertainty: uncertainty about the data given parameters, and uncertainty over the parameters themselves. 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…

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

where we would integrating over the posterior distribution of the parameters, instead of the prior distribution, returns a probability distribution for the CATE $\tilde{\tau}$ that reflects Bayesian updating from data Y.

3.2.1 Why Bayesian causal modeling?

The Bayesian causal approach is sensible for causal inference because it facilitates *direct inference* about treatment effects given the data: which effect sizes are more likely or less likely than others. While 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 positive treatment effects, negligible treatment effects, and more. Making positivistic statements about plausible causal effects is a natural way to think about the scientific aims of any discipline engaged causal inference: "the world probably works in this way, given the evidence." This is the view espoused by Don Rubin himself, the namesake of the Rubin causal model commonly employed in causal political science, who 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). It is not formally coherent to invoke similar language under a non-Bayesian inference paradigm, since we cannot statistically describe the plausibility of a causal effect without a posterior belief, which entails a prior belief. This is why non-Bayesian methods formally conduct inference about the plausibility of *data* given fixed parameters, and inferences about parameters must be done indirectly with an additional layer of decision theory. Statistical decisions under non-Bayesian analysis can be awkward—for instance,

using a p-value to measure whether the data are inconsistent with a null hypothesis that the researcher usually does not find credible at the outset, and usually with no corresponding measure of the relative plausibility of an alternative hypothesis (Gill 1999). Restated more formally, researchers routinely perform statistical inference by estimating $p(\mathbf{y} \mid H_0)$ (for null hypothesis H_0), when they are are probably more interested in $p(\boldsymbol{\pi} \mid \mathbf{y})$ or even $p(H_a \mid \mathbf{y})$ (for alternative hypothesis H_a).

The notion of direct inference is especially valuable when the observed data represent the whole population, which is common for observational causal inference in political science. Under a "frequentist" inference approach, estimators inherit their statistical properties from the sampling distribution of the estimator: the theoretical probability distribution of estimates in an infinite number of independent samples from the population. Many causal questions in political science conflict with this inferential framework, however, because the data come from events that have no possibility of future sampling. Instead, the data come from one-off events in history and often contain the entire population of relevant units, so the epistemic uncertainty about statistical inferences bears little resemblance to frequentist uncertainty about a sampling mechanism (Western and Jackman 1994).[3] The foundations of uncertainty in Bayesian inference are probability distributions that represent imperfect pre-data information about the generative processes underlying the variables in a model. Whether this imperfect information corresponds sampling randomness or other epistemic uncertainty can be subsumed in the Bayesian framework (Rubin 1978a).

Another reason why Bayesian methods make sense for causal modeling is because causal models, at their core, are models for counterfactual data. Because the Bayesian model is

[3]To be sure, there is non-Bayesian theory for statistical uncertainty without resampling, namely "design-based uncertainty" where the source of uncertainty is the treatment assignment mechanism itself (Abadie et al. 2020; Keele, McConnaughy, and White 2012). This framework is rarely invoked in political science except among researchers on the cutting edge of "agnostic" causal inference. For more on agnostic statistics, see Aronow and Miller (2019). It is also common for Bayesian statisticians to be interested in frequentist properties of their methods such interval coverage (Rubin 1984), which partially motivates an interest in "objective Bayesian inference" (Berger and others 2006; Fienberg and others 2006).

a *generative* model for parameters and data, the model contains all machinery required to directly quantify counterfactual potential outcomes using probability distributions. To see this in action, we can "run the model forward" to create a predictive distribution for Y given the model parameters. Denote these simulated observations as $\tilde{Y}$ to distinguish them from the data observed Y. If we marginalize this predictive distribution with respect to the prior parameters, we obtain a "prior predictive distribution"—the distribution of data we would expect under the prior (Gelman et al. 2013).

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

We update this distribution by conditioning on observed data, delivering 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 Bayesian generative models,[4] and they are the basis for counterfactual inference as well. Invoking the causal identification assumptions, we generate counterfactual data as predictive distributions as well, setting the treatment $A = a$ to some other value $A = a'$. Denote these counterfactual predictions $\tilde{Y}'$, which I will subscript i to show that this model implies a probability distribution for individual data points as well as aggregate treatment effects.

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

Stated more simply: if causal models define a space of potential outcomes, then Bayesian causal models are probabilistic representations of the potential outcome space. Probability densities over potential outcomes are defined in the prior and in the posterior, and they can

[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.

be defined all the way to the unit level if the generative model contains a probability statement for unit data.[5] The Bayesian view of causal inference, where the statistical model is nothing more remarkable than a missing data model for unobserved counterfactuals, is at least as old as Rubin (1978a).[6] Bayesian methods for causal inference have appeared in political science only sporadically in the decades since (Green et al. 2016; Horiuchi, Imai, and Taniguchi 2007; Ornstein and Duck-Mayr 2020). Unlike these rare examples of Bayesian causal inference, however, this chapter will contain more practical guidance for thinking Bayesianly about causal modeling, more synthesis of Bayesian causal innovations from other fields, and more examples of Bayesian causal modeling in practice.

It is common for advocates of Bayesian inference to celebrate the fact that the posterior distribution quantifies uncertainty in all parameters simultaneously, but this is especially useful for causal methods that entail multiple estimation steps. These multi-stage procedures include instrumental variables, propensity score weighting, synthetic control, causal mediation approaches, and structural nested mean models (Acharya, Blackwell, and Sen 2016; Angrist and Pischke 2008; Blackwell and Glynn 2018; Imai et al. 2011; Xu 2017). Bayesian approach to these methods combines all estimation stages into one model, estimating the ultimate treatment effect by marginalizing the posterior distribution with respect to the "design stage" parameters (Liao 2019; McCandless, Gustafson, and Austin 2009; Zigler and Dominici 2014). The combined modeling approach diverges from methods that use only

[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. See Section 3.3.8 for explanation and examples.

[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. This has a similar spirit as McElreath (2017b)'s "inside view" of Bayesian stats, where all data and parameters are unified as functions of probabilistic variables in a joint generative model.

point estimators, which ignore design stage uncertainty by conditioning the final-stage effects on design stage estimates or must derive post-hoc variance estimators to account for design stage uncertainty. It is reminiscent of "uncertainty propagation" methods insofar as uncertainty from design estimates are pushed forward to later estimates—for example, the way Kastellec et al. (2015a) simulate measurement uncertainty in MRP estimates of public opinion in later analyses. Unlike uncertainty propagation, which is mechanically similar to a "posterior cut" (Plummer 2015), the complete Bayesian model effectively treats the design stage estimates as priors and updates all model parameters using 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 estimated from a Bayesian measurement model. In order to understand district-party ideologies effects in primary elections, it is essential to pass posterior uncertainty from this measurement model into the causal analyses. However, building a combined model for district-party ideology and all downstream effects would be logistically overwhelming, so I approximate the full model by drawing ideal points in causal analyses from a multivariate prior distribution that reflects the measurement model's posterior samples. For instance, to understand the causal effect of district-party ideal point $\bar{\theta}_g$ on some outcome measure y_g, our causal model contains a model for the outcome,

$$y_g \sim \mathcal{D}\left(\theta_g, \ldots\right), \tag{3.12}$$

where $\mathcal{D}\left(\cdot\right)$ is some distribution, and we place a multivariate Normal prior on the vector of all district-party ideal points $\bar{\boldsymbol{\theta}}$ as

$$\bar{\boldsymbol{\theta}} \sim \text{MultiNormal}\left(\bar{\bar{\boldsymbol{\theta}}}, \bar{\Sigma}\right) \tag{3.13}$$

where $\bar{\bar{\boldsymbol{\theta}}}$ and $\bar{\Sigma}$ are the estimated mean vector and variance–covariance matrix of the ideal point samples from the measurement model. The multivariate Normal prior is justified

on the ground that the original ideal point model smoothed ideal point estimates using a hierarchical Normal prior, and estimating the variance–covariance matrix from the samples is a simple method to summarize systematic relationships between ideal points as a function of their geographic attributes. This approximation avoids the difficulties of writing an overly complicated model while still giving ideal points a prior that resembles their posterior distribution from the IRT model.

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. After all, the estimator exists in the space of real data, unlike a causal estimand that belongs to the hypothetical space of potential outcomes. Being merely an estimator, we could 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. We could characterize the use of Bayesian methods for these purposes as practically valuable but theoretically dispensable, in the sense that researchers might prefer non-Bayesian means to the same ends. For instance, Green and Kern (2012) adapt Bayesian Additive Regression Trees (or BART, Chipman et al. 2010) to measure treatment effect heterogeneity in randomized experiments. (See Hill 2011 for a non-political science introduction to causal inference with BART.) The advantages of a regression tree model for treatment heterogeneity is that it can explore arbitrary interactions among covariates while controlling overfitting, but the fact that BART is Bayesian is an afterthought. We observe a similar pattern in the use of Gaussian process models for fitting the running variable in regression discontinuity designs by Ornstein and Duck-Mayr (2020) and in the development of augmented LASSO estimators for sparse regression models by Ratkovic and Tingley (2017; 2017).[7] These authors use priors to regularize richly parameterized functions, posterior distributions to characterize uncertainty, and MCMC to estimate models, but the theoretical

[7] See Tibshirani (1996) for a general introduction to the LASSO shrinkage estimator using the L1 penalized optimization. Park and Casella (2008) implement a Bayesian LASSO using Laplace priors for regression coefficients.

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 priors for model parameters. Because causal estimates are functions of model parameters, prior densities on model parameters imply that some causal effects can be more or less likely even before considering any data. If the statistical model contains a unit-level data model, which is the case for most regression approaches, this implies that unit-level potential outcomes also have prior probability densities: some potential outcomes at the unit level are more or less likely, marginal of any observed data. This is a decisive philosophical departure from a non-Bayesian approach to causal modeling, where potential outcomes and causal effects are simply defined in a space of outcomes. As with causal effects, the benefit of prior distributions for unit outcomes is that that we can obtain, describe, and conduct direct inference using the posterior distribution for unit counterfactuals. The drawback is that researchers must specify priors for model parameters and understand how they impact the implied priors for unit counterfactuals. A few recent methodology papers in political science have invoked this idea of unit counterfactual estimation in a Bayesian framework, which is especially intuitive for synthetic control estimation (Carlson 2020; see Ratkovic and Tingley 2017 for a regression application). But these papers do not highlight the notion of direct priors for counterfactuals, so skeptical applied researchers have little guidance for understanding what it means to have priors on counterfactual data, either theoretically or practically.

The theoretical notion of prior densities in the potential outcome space is only interesting if it makes sense to see priors as important for causal estimation. I have already argued that priors enable direct probabilistic inference on causal effects and unit counterfactuals, which is the essential modeling goal for causal inference in the first place. It is natural to ask if it is sensible to use flat or uninformative priors achieve to enable direct probabilistic

inference while minimizing the uncomfortable feeling of specifying priors for unobservable counterfactuals, avoiding the need to think hard about priors at all. As I discuss in future sections, however, flat priors obscure rather than avoid these complexities. This is because even simple modeling scenarios that begin with flat priors usually contain quantities of interest whose priors cannot also be flat. As a result, it is the researcher's job to decide how to parameterize a problem and how the chosen priors affect the important quantities in the analysis.

Priors on important modeling constructs do not have to be an inconvenience. There are many scenarios where priors can actually relax assumptions, which 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 assumption holding. If we invoke a measurement model of treatment effects akin to Gerber, Green, and Kaplan (2004),

$$\text{Estimated Effect} = \text{True Effect} + \text{Bias} + \text{Error}, \tag{3.14}$$

such that the estimated effect equals the true effect only if it is estimated with zero random error and zero systematic bias. Identification assumptions imply a prior that the bias is precisely zero, but in many applied research contexts, this assumption may be unreasonable to sustain with 100% certainty. A Bayesian approach to identification assumptions allows the researcher to relax their model of treatment effects by specifying a different prior on the bias term, or the sensitivity parameters that compose the bias term, that is consistent with the researcher's reasonable expectations for the remaining bias in the research design (Oganisian and Roy 2020). This bears resemblance to sensitivity testing approaches in which the researcher evaluates the treatment effect estimates by stipulating a range of fixed values for a key sensitivity parameter, and then evaluating the treatment effect estimate conditional

on each fixed value (Acharya, Blackwell, and Sen 2016; Imai et al. 2011). Constructing these identification assumptions as priors lets the researcher conduct inference about treatment effects by *marginalizing over* their priors for design parameters, rather than conditioning on fixed design parameter values with little thought to which values are plausible or implausible. One recent political science example of this approach is Leavitt (2020), who frames the parallel trends assumption in a difference-in-differences (DiD) analysis as a prior over unobserved trends. This introduces an additional layer of "epistemic uncertainty" into the DiD analysis that would ordinarily be assumed to be zero by design. I elaborate on the value of priors for unidentifiable quantities in Section 3.4.1.

A few papers in political science and related fields invoke Bayesian causal models that occupy more than one level of this causal inference hierarchy at once. One important figure to note is Rubin himself, who has advocated for a "phenomenological Bayes" approach to causal inference since his pioneering papers on causal inference (Imbens and Rubin 1997; Rubin 1978a, 1978b, 2005). The fundamental tenet of the so-called "phenomenological" approach is the use of parametric models to generate posterior distributions for unobserved potential outcomes, which are used to construct causal estimates as functions of these unit-level posterior predictions. Posterior distributions for units are essential to the approach because statistical models may differ in their parameterization, making model parameters difficult to compare, but counterfactual data from the posterior distribution can always be compared to observed data regardless of the model parameterization. This is especially useful for studies with more complex missing data structures, such as studies with noncompliance, because Bayesian methods can always return a posterior distribution even for quantities that are not point-identified from data (Imbens and Rubin 1997). As such, the Rubin example invokes a Bayesian framework for estimation (level three) as well as a philosophical posture that the posterior distribution directly characterizes the plausibility of counterfactual data (level one). A recent political science example invoking a Bayesian approach for missing data under

treatment noncompliance is Horiuchi, Imai, and Taniguchi (2007).

Another important frontier for Bayesian methods in causal inference is meta-analysis. A fundamental modeling dilemma for meta-analysis is the choice between "fixed effects" and "random effects" approaches to meta-analysis (Borenstein et al. 2011). Fixed effects meta-analysis assumes that all studies contain imperfect estimates of the same underlying treatment effect, while random effects admit the possibility of between-study heterogeneity in the true underlying effect. Bayesian approaches to meta-analysis can compromise between the two extremes, relaxing the fixed effects assumption by allowing between-study variation while incorporating prior information that rules out extreme heterogeneity more aggressively than the conventional random effects approach. A common example from Bayesian pedagogy is a meta-analysis of parallel experiments across schools by Rubin (1981). The approach simultaneously estimates study-specific treatment effects, cross-study variance in treatment effects, and a population average treatment effect. Meager (2019) creates a similar model to summarize the experimental evidence on micro-credit expansion in developmental economics. Like the noncompliance approaches above, these meta-analyses engage in levels one and three by constructing the problem explicitly as Bayesian learning of experimental effects and engaging assertively with Bayesian modeling assumptions using hierarchical priors in the meta-analytical model. In political science, Green et al. (2016) perform a fixed effects analysis to aggregate noisy treatment effects, invoking a notion of Bayesian learning from parallel experiments but stopping short of a richer statistical model for cross-study variation. As I show in Section 3.3.8, the researcher's assumptions about cross-study variation is highly consequential for meta-analytic inference, and a Bayesian framework provides inimitable benefits for theorizing about those assumptions and understanding how they affect inferences.

For the remainder of this chapter, I explore several ways Bayesian modeling changes our view of causal inference at all three levels of hierarchy: causal modeling, identification

assumptions, and statistical estimation. Many points discussed below touch on more than one level of the hierarchy, so the chapter cannot be neatly divided according to the hierarchy.

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? This discussion has two main goals: a defensive goal of clearing up misunderstandings and criticisms of priors, and an offensive goal to argue that priors are omnipresent and valuable for thinking coherently about drawing inferences from causal models.

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? Can't the researcher sneak in enough prior information to get any answer they want? This argument accords with an interpretation that Bayesian analysis with informative priors is analytically equivalent to "data falsification," since priors influence the posterior distribution through the same mechanism as additional observations do: 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." Broadly, we can view a prior as any assumption that brings information into a model. This information can take the form of plausible parameter values, but almost all modeling, even so-called "non-Bayesian" modeling, adds information to a

model as assumptions about the residual distribution of observations, functional forms, and so on. These assumptions do not come from data. Priors are "beliefs" in a pragmatic sense: they are "belief functions" only in the sense that the height of the prior represents the support for a parameter value *within the model*, but outside the model, the researcher's degree of belief in a prior is zero, and the researcher is "morally certain" that the model is wrong (Gelman and Shalizi 2013, 19–20) These assumptions represent reasonable approximations that structure a model so that information extraction from data is more efficient. Since priors are often more pragmatic than sincere, it is common for Bayesian statisticians to care about other pragmatic features of modeling like frequency properties, designing noninformative priors for optimal learning from data, (Berger and others 2006; Fienberg and others 2006; Rubin 1984), and invoking workflow practices that "fall outside the scope of Bayesian theory" such as model checking with diagnostic graphics (Betancourt 2018; Gabry et al. 2019; Gelman and Shalizi 2013).

"randomness" is missing information. The connection between randomness and information is a key foundation of information theory, where randomness that cannot be further distilled is thought to be "pure information," and patterns that can be distilled

Information theory cite…

How does any model learn from data? By imposing structure on what data *can* say about parameters.

model structure subtracts randomness from the system, subtracts "pure information," says that "this is information that I already have, and the data won't teach me it." What the data *can* teach me about is the parameter, which is why they are represented as random variables with probability distributions.

The idea that data are drawn from a distribution is a prior, and the extent that we structure that prior represents external information injected into the model.

Generalization is all priors. Similarity of sample and population is akin to a prior that any

residual variation is none. If we allow for the sample to vary from the population in some systematic way that we might correct for either by modeling study effects or post-stratification, these represent information laid overtop data. [free lunch].

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

Formally, a flat prior is a probability distribution that assigns equal probability density to all possible values of a parameter. It is common for researchers to prefer flatter priors in situations where they lack specific prior information about model parameters. In cases where a prior distribution is perfectly flat, the posterior mode will be equivalent to the maximum likelihood estimator regardless of the sample size. For this reason, many researchers who find themselves exploring Bayesian methods use flat or nearly-flat priors as a "default choice". It is likely that a researcher exploring Bayesian modeling for causal inference might make a similar choice, in the interest of "letting the data speak." A common criticism of this practice is that flat priors understate the researcher's actual prior information. Although this sounds obviously true, it is easy to see why an applied researcher would be unbothered by it—what's the harm with a vague prior? Instead, I argue that the default use of flat priors can often *misspecify* prior information in many common modeling scenarios. Researchers must set prior distributions on parameters, but they usually have prior information about outcome data, which are functions of parameters. If researchers are not mindful of the functional relationship between raw parameters and other quantities of interest in the analysis, they will not the pernicious effects that "default priors" can have on model behavior.

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 is a prior density, then $h(\pi)$ has an implied prior density that is affected by the density of π and the composition of 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 $(3, 2)$, because

a linear transformation of a Normal random variable results in another Normal random variable. Figure 3.1 shows the probability density of π and the implied density of $h(\pi)$. Any function of a parameter will an implied probability density, but there is no general guarantee that the implied density will be easy to understand like the example in Figure 3.1.

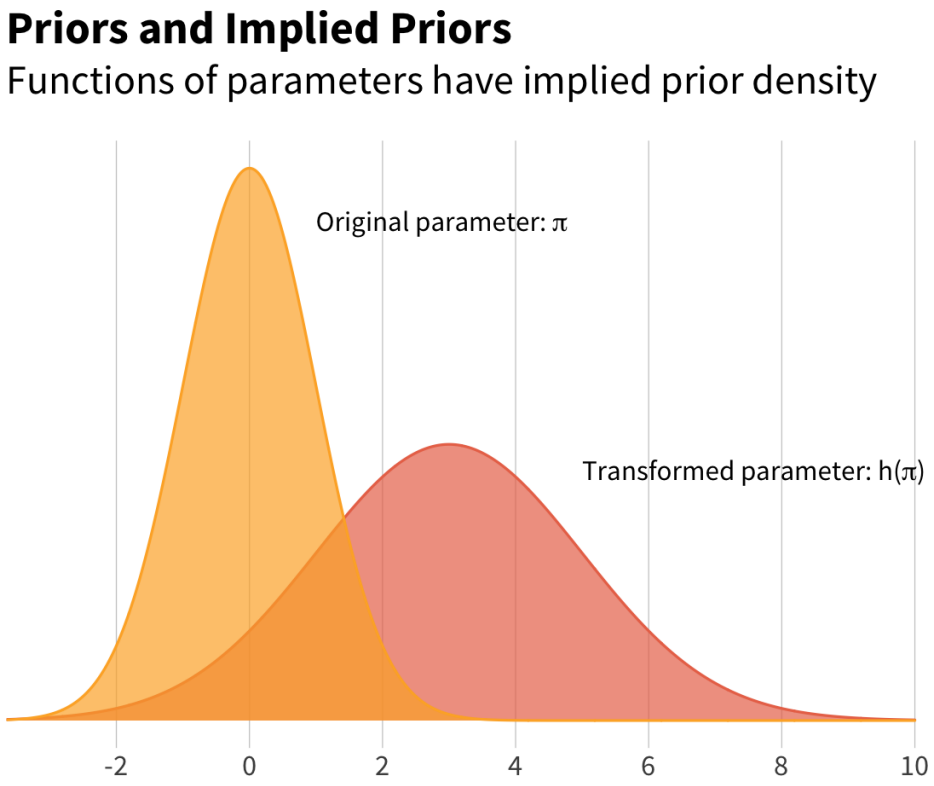


Figure 3.1: Demonstration of centered and non-centered parameterizations for a Normal distribution. The non-centered parameterization is statistically equivalent, but the location and scale are factored out of the distribution.

In order to understand a model, the researcher should understand the consequences of priors on the model's predictive distribution for new data, because prior distributions in parameter space yield implied priors in outcome space. This highlights an ever-present, pragmatic problem in building any model: a researcher has prior information about the *world*, but they must set priors on model parameters. This requires the researcher to solve backward for priors in parameter space that resemble reasonable expectations about the behavior of

data. In order to understand and anticipate the consequences of priors, it is important to understand the data's functional dependence on model parameters and how those functions transform probability densities. It is thus a general principal of model-building that "the prior can often only be understood in the context of the likelihood" (Gelman, Simpson, and Betancourt 2017). The close relationship between priors and the data model expose where flat priors create problematic model behaviors.

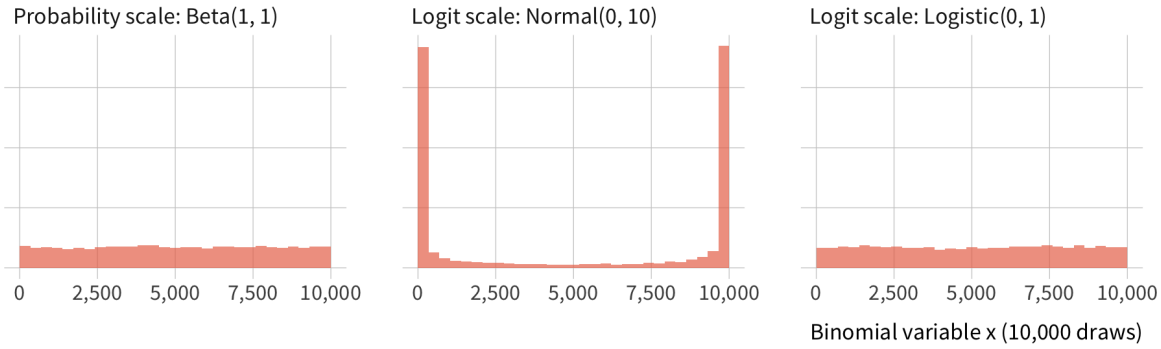
It is helpful to highlight an example where flat priors, believed "by default" to be reasonable and conservative, create problematic or nonsensical results (Seaman III, Seaman Jr, and Stamey 2012). Consider a binomial outcome variable that counts x successes out of n independent trials, each with success probability α. We want to estimate α, and we don't have specific prior information about it. We represent ignorance about α using a flat Beta $(1, 1)$ density. Not consider the identical data but estimating α with a logit model likelihood instead, where $\alpha = \text{logit}^{-1}(\eta)$. In the logit model, α is a deterministic function of the parameter η, so we place the prior on η. We are ignorant about its value as well, so we follow a default instinct and give it a prior with a wide variance, $\text{Normal}(0, 10)$ on the logit scale, and we might have considered scale values larger than 10 as well to represent more ignorance.

If we take both of these models and generate prior simulations for $x \sim \text{Binom}(n, \alpha)$, depicted as histograms in Figure **??**, the implied prior distributions 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 from the logit model, where $\text{logit}(\alpha) = \eta \sim \text{Normal}(0, 10)$, resulting in a very different prior for x where most prior probability is concentrated on very small or very large values of x. Why does this prior for x appear so informative when we used such a vague prior for η? Because η had a wide prior on the logit scale, and the probability scale is a nonlinear transformation of the logit scale, the resulting prior for x is also affected by the nonlinear mapping between parameter spaces. Only a thin range of logit-scale values map

to probabilities that we routinely encounter in political science: logit sclae values between -3 and 3 correspond to probabilities between approximately 0.047 and 0.953. Because the Normal $(0, 10)$ prior places most probability density outside of reasonably values on the logit scale, most of our implied probabilities are unreasonably small or large as well. In order to obtain a flat prior for x using a logit model, we would actually use the prior $\eta \sim \text{Logistic}(0, 1)$, shown in the third panel of Figure **??**. The standard Logistic prior creates a flat density on the probability scale because the inverse link function for a logit model is the cumulative distribution function for a standard Logistic distribution.[8]

Prior Flatness ≠ Prior Vagueness

How transformations of parameter space affect implied priors



- [x] Introduce implied prior
- [x] outcome vs. parameters
- principal: info vs. shape
- example, horiuchi
- lesson for causal inference: we should expect to get different results because thinking explicitly about priors shows us where implied priors aren't flat (and that's fine) or where implied priors ARE flat (and that's batshit).
- we might have vague priors in one space that are informative in another, or one space is easier for specifying priors

[8] This same intuition holds for a standard Normal prior in a probit model, demonstrated in Section 2.5.4.

Flat priors are problematic for many analyses because they lead to implied priors that have strange behaviors that researchers would be uncomfortable confronting.

More complex functions of parameters, especially nonlinear functions, can produce in implied priors with strange shapes that are difficult to anticipate using intuition alone. [horiuchi]

Outcome space != parameter space. Flat priors about outcomes is not the same thing as flat priors about parameters. Information != shape. The mapping from geometric space to statistical information is entirely dependent on the scale of a parameter space, and different model parameterizations invoke different scales.

The primary resistance to Bayesian inference in applied research is the need to set a prior at all. To many researchers, the prior distribution is an additional assumptions that is never feels justified because it is external to the data. Often researchers wish to sidestep this choice altogether, preferring a "flat" prior that prefers all parameters equally.

We have seen so far that the parameterization of a model has consequences for prior specification. Reparameterization may result in an algebraically equivalent likelihood

The incoherence of flatness:

- no universally valid strategy for specifying flat priors because it is always possible to rearrange the data model either by transforming a parameter or otherwise rearranging the likelihood.

Consider an experiment with a binary treatment Z and a binary outcome variable Y. We want to determine the effect of Z by comparing the success probability in the treatment group, π_1, to the success probability in the control group, π_0.

- "no way to conceptualize an uninformative prior because you can always rearrange the problem through a reparameterization or transformation of a parameter"

- examples of transformations having crazy implications/MLE being wild (logit).
- Jeffreys prior: actually a very limited range of priors that satisfy an "invariance" property. My words: such that the "amount of information obtained from data about is invariant to parameterization of the likelihood, for all possible values of the parameter," or, "the only way for the posterior distribution to be exactly the same, given the same data, for all true parameter values (?), is the Jeffreys prior," or, regardless of the data, I will learn the same thing about the generative model regardless of which equivalent parameterization of the generative model is used.
  - is it worth it to think about the theoretical meaning of information
  - how does flatness reflect information in nonlinear scales?

Suppose we have some posterior distribution which relies on some parameter vector $\vec{\alpha}$.

$$p\left(\vec{\alpha} \mid y\right) \propto p(y \mid \vec{\alpha}) p(\vec{\alpha}) \tag{3.15}$$

Consider some alternate parameterization of the likelihood parameterized by $\vec{\beta}$.

Nonlinear transformation of π does not preserve a uniform density over parameters.

Alex meeting takeaways:

- every prior has a "covariant" prior in a different parameterization
- the posteriors will be covariant as well.
- The way you get between them is by transforming the parameter and doing the appropriate Jacobian transformation to the density.
- Jeffrey's priors are a special case of this where the prior is proportional to the determinant of the information matrix. This has the beneficial property of "optimal learning" from the data. For example, flat Beta prior doesn't "hedge toward 50" in quite the same way.

3.3.3 Priors and model parameterization {sec:prior-parameterization}

Priors are defined with respect to a model of the data (the likelihood). We may have priors about the way the world works, but we rarely have priors about model parameters. This is because parameters are an invention in the model. They are mathematical abstractions similar to points and lines, so they only exist when we translate the world into a mathematical language. This means that the mathematical representation of the world is in direct dialog with the choices available to a researcher about how to encode prior information. In the real world, the prior information that I have about the world isn't affected by a mathematical representation of the world. As a researcher, the way I encode prior information depends on the choices I make about that mathematical representation.

One essential feature for understanding prior choices in practice is the *parameterization* of the data model, $p(y \mid \phi)$, for some generic parameter ϕ.[9] We say that a data model has an "equivalent reparameterization" if for some transformed parameter $\psi = f(\phi)$, the function that defines the data model can be rewritten in terms of ψ and return an equivalent likelihood of the data. More formally, the parameterization is equivalent if $p(y \mid \phi) = p(y \mid \psi)$ for all possible y.

In a maximum likelihood framework, equivalent parameterizations are a more benign feature of the modeling framework. Reparameterization may result in likelihood surfaces that have easier geometries for optimization algorithms to explore, but the *value* of the likelihood function is unaffected by the algebraic definition or parameterization of the likelihood function. For instance, a Normally distributed variable x with mean μ could be parameterized in terms of standard deviation σ or in terms of precision $\tau = \frac{1}{\sigma^2}$, but the resulting density is

[9] Bayesian practitioners sometimes refer to the data model as the "likelihood." This can be confusing because the "likelihood function" more traditionally refers to the *product* of the data probabilities under the data model. References to the "parameterization of the likelihood" should be understood as interchangeable with "parameterization of the data model," since the former is determined entirely by the latter.

unaffected.

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\left(\frac{x-\mu}{2\sigma}\right)^2} = \sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau(x-\mu)^2}{2}} \tag{3.16}$$

The consequence for Bayesian analysis, however, is that the parameterization of the data model determines the set of parameters and their functional relationship to the data.

One example of equivalent reparameterization arises with the different possible ways to write a linear regression model. The first form specifies y_i for unit i as a linear function of x_i with a random error that is mean 0 and standard deviation σ.

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \text{where } \varepsilon_i \sim \text{Normal}\,(0, \sigma) \tag{3.17}$$

The second form, more common when viewing linear regression in the framework of generalized linear models, is to express y_i directly as the random variable, with a conditional mean defined by the regression function and standard deviation σ.

$$y_i \sim \text{Normal}\,(\alpha + \beta x_i, \sigma) \tag{3.18}$$

Algebraically, these two models are identical. The difference is only a matter of which component has the distributional assumption. In Equation (3.17), the distribution is assigned to ε_i, so y_i is a random variable only by way of ε_i. In Equation (3.18), we assign the distributional assumption directly to y_i, bringing the regression function into the mean rather than "factoring it out" of the distribution.

The linear regression context is one context where the choice of parameterization appears. These two parameterizations are typically called the "centered" and "non-centered" parameterization for a Normal distribution. In the centered parameterization, the random variable is drawn from a distribution "centered" on a systematic component, whereas the non-centered distribution factors out any location and scale information from the distribution, such that the only remaining random variable is a standardized variate. The equations below describe

a Normal variable v with mean 3 and standard deviation 2.

$$\text{Centered Parameterization:} \quad v \sim \text{Normal}(3, 2) \tag{3.19}$$

$$\text{Non-Centered Parameterization:} \quad v = 3 + 2z, \quad \text{where } z \sim \text{Normal}(0, 1) \tag{3.20}$$

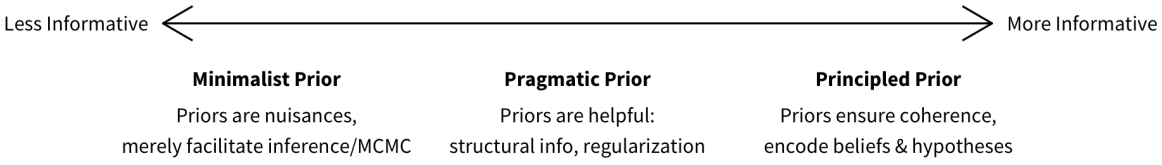


Figure 3.2: A spectrum of attitudes toward priors.

Problems of beliefs:

- No degree of belief.
- Parameterization makes this too challenging.
- Prior might change depending on what I ate for lunch.
- "Elicitation" of priors satisfying the wrong audience, or at the very least can be easily misused. We do not want to elicit priors about arcane model parameters. We want to elicit priors about the *world* (Gill Walker)d

Problems of nuisance prior

- parameterization gets you again
- the MLEs are unstable, overfit
- make the regularization argument in-sample

Pragmatic view of priors

- we're between full information and nuisance prior
- Weak information: structure, regularization, identification

- Structural information about parameters
- regularization toward zero (L1, L2), learning by pooling
- stabilizing weakly identified parameters, separation, etc.

Parameters are a *choice*.

- They are part of the *rhetoric* of a model. Sometimes we make pragmatic choices (something is easy to give an independent prior to, but independence isn't always valuable per se). Sometimes we make principled choices (normality, laplace, etc).
- They deserve scrutiny (else just "excrete your posterior") and are a part of the model that you should check and diagnose.
- They aren't merely a nuisance because we can use them to our benefit,
- sometimes when we parameterize a problem to reveal easier things to place convenient priors on

Models are a tool, set it up so that it works.

Constrained parameters in causal mediation?

For instance, consider a simple experiment with a binary outcome variable y_i and binary treatment assignment $z_i \in \{0, 1\}$. Suppose that the treatment effect of interest is a difference-in-means, $\bar{y}_{z=1} - \bar{y}_{z=0}$, estimated from a linear probability model. This linear probability model might be parameterized in two ways. First is a conventional regression setup,

$$y_i = \alpha + \beta z_i + \varepsilon_i \tag{3.21}$$

where α is the control group mean, $\alpha + \beta$ is the treatment group mean, β represents the difference in means, and ε_i is a symmetric error term for unit i. With the model parameterized in this way, the researcher must specify priors for α and β. Suppose that the researcher gives β a flat prior to represent ignorance about the treatment effect. An equivalent *likelihood*

model for the data would be to treat each observation as a function of its group mean μ_z.

$$y_i = z_i \mu_1 + (1 - z_i)\, \mu_0 + \varepsilon_i \tag{3.22}$$

Although the treatment effect β from Equation (3.21) is equivalent to the difference in means $\mu_1 - \mu_0$ from Equation (3.22), the parameterization of the model affects the implied prior for the difference in means. If the researcher gives a flat prior to both $\mu_{\mathbf{z}}$ terms, the implied prior for the difference in means will not be flat. Instead, it will be triangular, as shown in Figure 3.3. The underlying mechanics of this problem are well-known in applied statistics—if we continue adding parameters, the Central Limit Theorem describes how the resulting distribution will converge to Normality—but it takes the explicit specification of priors to shine a light on the consequences of default prior choices in a particular case. In particular it shows how even flat priors, which are popularly regarded as "agnostic" priors because of their implicit connection to maximum likelihood estimators, do not necessarily imply flat priors about the researcher's key quantities of interest. Rather, flat priors can create a variety of unintended prior distributions that do not match the researcher's expectations. I return to this important idea in the discussion about setting priors for a probit model in Section 2.5.5.

- equivalent parameterizations

3.3.4 Principled and pragmatic approaches to Bayesian modeling

Degrees of informativeness

- merely facilitate posterior inference
- structural priors
- WIPS
- informative priors

Prior strategies

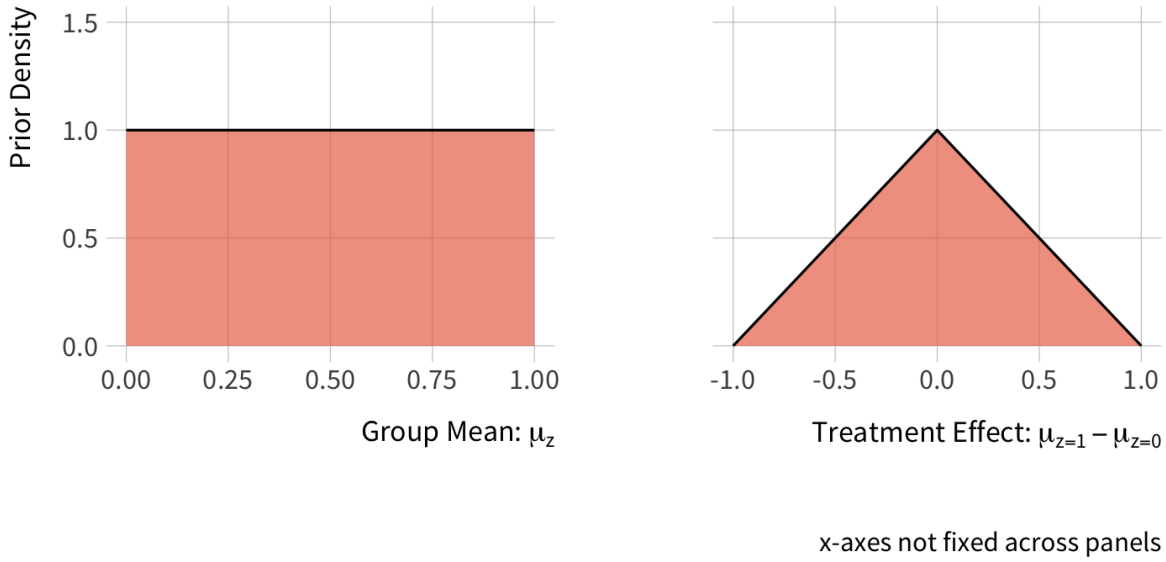


Figure 3.3: Model parameterization affects prior distributions for quantities of interest that are functions of multiple parameters or transformed parameters. The left panel shows that the difference between two means does not have a flat prior if the two means are given flat priors. Note that the x-axes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

- ref
- stabilization
- regularization
- prior knowledge

A general orientation toward priors in this dissertation:

- Not about "stacking the deck" or hazy notions of "prior beliefs"
- information, not belief

- Bayesian view of probability is *more general*, contains information and beliefs. Information is priors, but it's also data. Information is the fundamental unit of uncertainty-quantification
- inference about the thing we care about (counterfactuals)
- structural information when we have it
- Causal inference: "agnosticism" is something valuable generally

Priors are not de-confounders

- downweighting, not upweighting
- contra western/jackman

3.3.5 Statistical Bias

1. posterior isn't about frequency properties (especially in one-off data)
2. What is "bias?"
    - look up in BDA
    - "Requiring unbiased estimates will often lead to relevant information being ignored (as we discuss with hierarchical models in Chapter 5)" (94)

- Why would we want this? Inference makes more sense.
    - What's the probability of a model/hypothesis, given the data
    - vs. What's the probability of "more extreme data" (?) given a model that I do not believe.

- posterior probabilities mean what they say they mean
    - conditioning on data (and implicitly the model), this is the distribution of parameters

  - p-values are the "probability of more extreme results." They condition on the model, but they're only useful if they do not.

- Proper frequentist analysis is violated as soon as you look at the data.
- Frequency properties are still possible for Bayesian estimators, but we view frequency properties as a byproduct of something more essential (MSE).

3.3.6 Structural Priors and Weak Information

Structure (bounds), regularization (L1, L2), hierarchy

p doesn't care about your n.

3.3.7 Understanding Log Prior *Shape*

This is low-key pretty big

3.3.8 Models for Means

In many situations, we can get around a data-level model if we want to.

3.3.9 Generalization, Big and Small

Models (likelihoods) are priors

- we restrict the space of all other models
- could think about "flexible" models, but these are just priors over more spaces

Identification assumptions are priors

Generalization to any population is a prior

- priors are not MY data, but ANY data
- parameters describe ANY data

- Lampinen Vehtari 2001: likelihood as "prior for the data" is the basis for all generalization from any finite model

We are always doing violence, but the framework lets us build out more and more general models to structure our uncertainties

Email to David:

> My first reaction to this is like this: it's probably correct to say that Bayesians may have some shared ideas about how to think about generalizing that might differ systematically from non-Bayesians, but I am not sure how much of that is because of Bayes per se so much as just…the types of modeling someone is willing to do. By "modeling" I mean, functional form assumptions are you willing to make about data, which is different from "Bayes" in that the former is something required of all modeling and the latter is only what you can say about parameters. For example, a functional modeling thing you might do is specify (using some weights or something) how an estimate in one sample might map to another sample, but whether you do that with Bayesian estimation is a separate choice aside from the functional model itself. That being said, even though functional modeling things can be done with Bayes/non-Bayes point of view, it does seem right to say that certain modeling approaches may feel more natural in a Bayesian framework, or that Bayesians might construe a problem slightly differently because they are more used to hierarchical modeling.
>
> That's a pragmatic view of things, but I can give you a more theoretically abstract view, and think of situations where Bayesian lets you do things that non-Bayes can't. It all comes down to how "seriously" you want to take the tenets of Bayesian work and what kind of generalization-based claims you want to make. So I will lay out a series of vignettes that start from a world where Bayes is "not unique"

and then gets into worlds where Bayes gets more and more necessary to say what you want to say about the out-of-sample world. I will use the example of estimating some parameter, but you can translate this into learning about a "mechanism" however Erica and Nick are defining that, and you can think about it non-statistically as well even if I'll use statistical modeling language.

I estimate some parameter μ in a study, but it's one instance of a more general phenomenon. If the study were representative of the world, you can imagine that the effect is one instance of the "population effect" and give it a hierarchical prior as such: $\mu \sim D(\theta, \sigma)$ where D() is some distribution, θ is the "general" effect. If I learn about μ (the estimate and standard error σ), I learn about θ! Choice of distribution depends on your assumptions about the stochastic process at work. naturally, but that logic works basically just like a likelihood function choice. (In fact, Bayes sees priors for parameters as mechanistically no different from likelihoods for data. Which is to say, MLE models like logit are simply hierarchical priors on the data, and regression is estimating the hierarchical parameters of the prior.) This is basically a meta-analysis setup using Bayesian language: if you want tangible examples you can look at the Don Rubin "eight schools experiment" which is about generalizing from parallel studies in schools, or Rachael Meager (sp) has a paper applying this setup to micro-credit experiments in the development econ context: what do we learn about the "overall effect of microcredit expansion" by assimilating information from different studies. In this sense the priors are just ways to structure the meta-analysis.

If the study is unrepresentative or "not externally valid" then it's up to the researcher to specify some approach to modeling the invalidity: $\theta \sim D(f(\theta), \sigma)$, where f() is some function that distorts the representativeness of the study. Which

is to say, $f(\theta)$ is the expectation for a study with these distortions. Researcher's task is then to learn the form of f(). These distortions might be like sample bias, the country where the study was performed, or whatever, and all you're really doing is reweighting or adjusting the estimate to make more sense for the target population. If you can estimate parameters that determine f(), viola you can infer the posterior distribution for the true θ. But this is what I mean when I say that none of this is really EXCLUSIVELY Bayesian. Reweighting/adjusting happens in non-Bayes world all the time; the main thing that is different is how you write the model and your ability to say that the population estimate is a "posterior of the true parameter, given the information learned from the data." One example of this kind of thing is maybe the multilevel regression and poststratification: we use national surveys to model the attitudes of different demographic groups, and then we use those model predictions plus census information to project estimates for smaller units, for example states or counties, based on the demographic composition of those units! This example goes the other direction from where Erica and Nick want to go (from representative to unrepresentative) but the technology sounds similar: estimate something in the data that you have, and map it into a space where you do not have data. MRP in political science comes from Bayes world and feels natural there, but nothing saying that it HAS to be Bayesian in its overall approach.

Now we get to a world where Bayes is more necessary. If there's something TRULY Bayesian that really makes no sense withoutBayes, it is the fact that I actually do not need data about f() in order to estimate θ. This is because I have priors even in the absence of data. If all I have is priors about f(), then even if I collect data about ONLY μ and NOTHING about f(), I nonetheless update my information

about θ. This is because the parameters are functionally related through f(), so if I learn about the subsample then I learn about the population. Stated differently, learning about μ restricts the space of θ because I can basically "solve backward" using my priors about f. This is the stuff that is very natural in Bayes, and I can think of basically no analog in non-Bayes that lets you do something similar (other than picking point estimates for unknowns and simulating, which doesn't have the same theoretical coherence as a prior/posterior distribution). Of course, this means that inference on f() is subject to the priors that go into f(), which is exactly the kind of thing that non-Bayesians are super afraid of despite majorly misconstruing how this works (IMO). For one, the functional form of f() is the kind of thing non-Bayesians would make assumptions about anyway, so that's not unique to Bayes at all. And secondly, the priors that would go into f() would usually be generic enough that researchers aren't "picking their hypothesis" (a common and frankly stupid stereotype) so much as restricting the space of f() to rule out stuff that's frankly impossible. Happy to give you more concrete examples of the kind of "weakly informative priors" that someone would use in a situation like that if it's a route you want to dig into more. It's this kind of stuff that I think non-Bayesians are under-utilizing: how much extrapolation power you get by being willing to place even weak priors on stuff you can't exactly identify with data. And if you REALLY want to give a nod to Bayesian views of extrapolation, this is the area you'd want to dig into, because it's the stuff that doesn't really make sense without Bayes. You can sort of see Ken and I do this in our voter ID paper (which you can find on my website) though we kind of wimp out of fully placing priors on f().

Here's where things get really abstract because, gun to my head, we can be

really scorched-earth and say that all extrapolation falls apart without a Bayesian notion of priors. Think about any model for data: $y \sim D(\theta)$, I think my data come from this distribution, and if I were to go out into the world and collect new data, my estimate for a new data point is characterized by this distribution assumption i.e. this prior for the data. If you try to lay out a formal definition of what "generalization" is, I would say that there is no such thing as generalization without an implicit prior that links your observed data to unobserved things that you want to project into. There are stats theorems out there called "no free lunch" theorems that basically say "all statistical inference is limiting the space of models that link parameters to data, and there is no way to improve your guess for a new data point using a model except to impose prior information on the system by way of the model." So this would be a hard line view of what priors mean in a philosophy of science (not necessarily quantitative or statistical, mind you), but that if you accept that view it trickles down into the more minor examples in a very natural way: the only way to generalize is by using priors to structure the connection between what I do observe and what I do not observe.

3.3.10 Regularization and Prediction Problems

One-Off Data Collection:

1. posterior isn't about frequency properties
2. What is "bias?"
    - look up in BDA
    - "Requiring unbiased estimates will often lead to relevant information being ignored (as we discuss with hierarchical models in Chapter 5)" (94)

- Why would we want this? Inference makes more sense.

  - What's the probability of a model/hypothesis, given the data
  - vs. What's the probability of "more extreme data" (?) given a model that I do not believe.

- posterior probabilities mean what they say they mean
  - conditioning on data (and implicitly the model), this is the distribution of parameters
  - p-values are the "probability of more extreme results." They condition on the model, but they're only useful if they do not.

- Proper frequentist analysis is violated as soon as you look at the data.
- Frequency properties are still possible for Bayesian estimators, but we view frequency properties as a byproduct of something more essential (MSE).

3.3.11 Regularization-Induced Confounding

This is a huge, underappreciated problem in the broad ML-for-causal-inference world

3.4 Bayesian Opportunities

3.4.1 Priors for Imperfect Identifiability

Relatedly: priors over models

"The Bayesian approach also clarifies what can be learned in the noncompliance problem when causal estimands are intrinsically not fully "identified." In par- ticular, issues of identification are quite different from those in the frequen- tist perspective because with proper prior distributions, posterior distribu- tions are always proper." (Imbens and Rubin 1997)

This is where Imbens and Rubin push (also Horiuchi et al. example)

Randomization limits the impact of the Bayesian assumptions

- "Classical random- ized designs stand out as especially appealing assignment mechanisms designed to make inference for causal effects straightforward by limiting the sensitivity of a valid Bayesian analysis." (Rubin 1978)

Casting these identification assumptions as priors has at least two distinct benefits over an approach with fixed sensitivity parameter values. Firstly, it allows the researcher to conduct inference about the treatment effect by *marginalizing over* their priors for design parameters, rather than conditioning on fixed design parameters. By marginalizing over the design parameters, the researcher obtains one posterior distribution that averages over their prior design uncertainty, rather than an arbitrary range of design parameter values that contains no information about which design parameter values are realistic or unrealistic. Secondly, and relatedly, marginalizing over design parameters speaks to the inferential question that is more consistent with the researcher's guiding question: which treatment effects are consistent with the data and our prior information. Sensitivity analysis using fixed parameters, meanwhile, answers a different question that is less useful for the goal of posterior inference about treatment effects: how big of an assumption violation is required in order for treatment effect estimates to be statistically insignificant, with no mention of whether violations of that size are likely or unlikely.

returns a quantity of interest that is more consistent with the researcher's ultimate inferential goal—what treatment effects are consistent with prior information and the data—rather than a quantity

Many identification assumptions can be encoded as functional form assumptions in an estimated model. Consider a regression scenario with outcome Y_i, treatment A_i, and some other covariate U_i.

$$\begin{aligned} \text{True model: } Y_i &= \alpha + \tau A_i + \beta U_i + \varepsilon_i \\ \text{Estimated model: } Y_i &= \hat{\alpha} + \hat{\tau} A_i + e_i \end{aligned} \tag{3.23}$$

If we invoke an identifying assumption that the assignment of A_i is ignorable, then our estimate for the average effect of A_i using the second line of (3.23) is unbiased. This is equivalent to an assumption that the correlation between the treatment A_i and the true error term ε_i is exactly zero. For causal inference purposes, researchers have developed sensitivity tests that recover the treatment effect under ignorability violations, which work by picking a non-zero error correlation value and deriving the treatment under that model (Acharya, Blackwell, and Sen 2016; Imai et al. 2011).

, and building a richer understanding of how inferences depend on modeling assumptions.

Researchers have derived estimators for τ in the presence of a, which can be used in a causal inference context to test the "robustness" or "sensitivity" of inferences to the ignorability assumption (Acharya, Blackwell, and Sen 2016; Imai et al. 2011). If this correlation is zero, the treatment effect estimate is unbiased These are

In a causal inference context, researchers sometimes develop sensitivity tests to test the robustness of their inference to violations of ignorability assumptions (Acharya, Blackwell, and Sen 2016; Imai et al. 2011), exclusion restrictions (Nyhan, Skovron, and Titiunik 2017), and so on. Because identification assumptions are special cases of these more general models with additional nuisance parameters, it would be straightforward in a Bayesian framework to specify prior distributions on nuisance parameters that represent priors for violated assumptions. Unpublished work by Leavitt (2020) explores this approach in a difference-in-difference contexts, introducing an "epistemic" source of variance in his treatment effect estimate by specifying a prior on the parallel trends assumption.

- Oganisian and Roy (2020) "identification" assumptions vs. "statistical" assumptions. Identification assumptions get you to a place where you can express causal effects in terms of expectations about Y given treatment, within confounding strata (perhaps then averaging over confounders). "Statistical" assumptions define how we think we

can build $E[Y]$

3.4.2 Application: Weakly Informed Regression Discontinuity

This section presents a 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 with nonzero probability. We overcome the pathology using weakly informative priors that contain structural information about the dependent variable only, excluding impossible parameters from the prior but uninformative over possible parameter 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. The treatment variable of interest, the ideological extremity of a party's general election nominee, is confounded by several factors. Competitive districts may lead more moderate candidates to run in the first place, creating selection biases for which candidates represent which district. Conversely, voters in electorally "safe" districts may feel freer to nominate more extreme candidates because the likelihood of their party losing the seat in the general election is sufficiently low. Furthermore, the incumbency advantage in general elections confounds this picture if incumbents tend to be more moderate than challengers who are angling to raise their name recognition (Gelman and King 1990).

To identify the effect of candidate ideology, 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 a relative extremist and a relative 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 any greater than 0. If the extremist's primary margin is any less than 0, the moderate advances to the general election instead. This primary margin deterministically assigns congressional candidacies to treatment or control if the extremist wins or loses the primary, respectively. While candidate ideology's effect on general election outcomes may be confounded in the aggregate, the effect can be identified at the threshold (extremist margin of 0). The key identification assumption for a "sharp" regression discontinuity design is that the forcing variable, X_i, and the expected outcome given the forcing variable, $\mathbb{E}\left[Y_i(x) \mid X_i\right]$, are both continuous at the threshold x_0. This assumption identifies *local* treatment as the difference in the limits of the conditional expectations for treatment ($X = 1$) and control ($X = 0$) at the threshold (Calonico, Cattaneo, and Titiunik 2014; Skovron and Titiunik 2015).

$$\lim_{x \downarrow x_0} \mathbb{E}\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow x_0} \mathbb{E}\left[Y_i \mid X_i = x\right] = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = x_0\right] \tag{3.24}$$

Equation (3.24) implies that the difference in potential outcomes can be identified from observed data only by observing that everything else about units is continuous at the threshold except for the realized treatment value.

Hall (2015) applies the RD design by assuming that the effect of candidate ideology on vote share and win probability in the general election are identified locally where the extremist's margin in the primary election crosses 0. For this example, we concentrate on models that predict win probability, since these are the estimates that contain pathological results that we can avoid with Bayesian methods. 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. The local linear regression is justified by the limit intuition of

the key assumption in (3.24); any nonlinear regression function, as long as it is continuous at the cutoff, converges to linearity at the cutoff in the limit. Data were obtained from Hall's replication materials, available on his website.[10] The outcome y_{dpt} is a binary indicator that takes the value 1 if the general candidate running in district d for party p in election year t wins the general election, and it takes 0 if the candidate loses the general election,

$$\begin{aligned} y_{dpt} = \beta_0 + \beta_1(\text{Extremist Wins Primary})_{dpt} + \beta_2(\text{Extremist Primary Margin})_{dpt} \\ + \beta_3(\text{Extremist Wins Primary} \times \text{Extremist Primary Margin})_{dpt} + \varepsilon_{dpt} \end{aligned} \tag{3.25}$$

where *Extremist Primary Margin* is the extremist candidate's margin over the moderate candidate with the highest vote in the primary, and *Extremist Wins Primary* is a binary indicator equaling 1 if that margin exceeds 0, and ε_{dpt} is an error term. When the extremist margin exceeds 0, the candidate representing case dpt is the extremist, otherwise the candidate representing dpt is the moderate. The coefficient β_1 represents the intercept shift associated with the extremist primary win, estimating the treatment effect of candidate extremism at the discontinuity. I replicate this LPM using ordinary least squares, and I also create a Bayesian equivalent using an algebraically reparameterization. The Bayesian parameterization has the same linear form, but instead of specifying two lines an interaction term, I subscript the coefficients by w, which indexes the treatment status (*Extremist Wins Primary*),

$$\begin{aligned} y_{dpt} &= \alpha_{w[dpt]} + \beta_{w[dpt]}(\text{Extremist Primary Margin})_{dpt} + \varepsilon_{dpt} \\ \varepsilon_{dpt} &\sim \text{Normal}\,(0, \sigma) \end{aligned} \tag{3.26}$$

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 will be helpful for extending the model below.

[10] http://www.andrewbenjaminhall.com/, last accessed July 02, 2020.

I plot the OLS win probability estimates around the discontinuity in Figure 3.4. 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). The original publication lacks a graphical depiction of these results. Our visualization of the RD predictions reveal that the confidence set for the parameter estimates that compose the treatment effect contain many values that would be impossible to observe. The point estimate for average moderate candidate win probability at the discontinuity is 0.95, which is a possible number to obtain, but the 95 percent confidence intervals includes values as high as 1.24, which far exceeds the maximum possible value of 1.0.

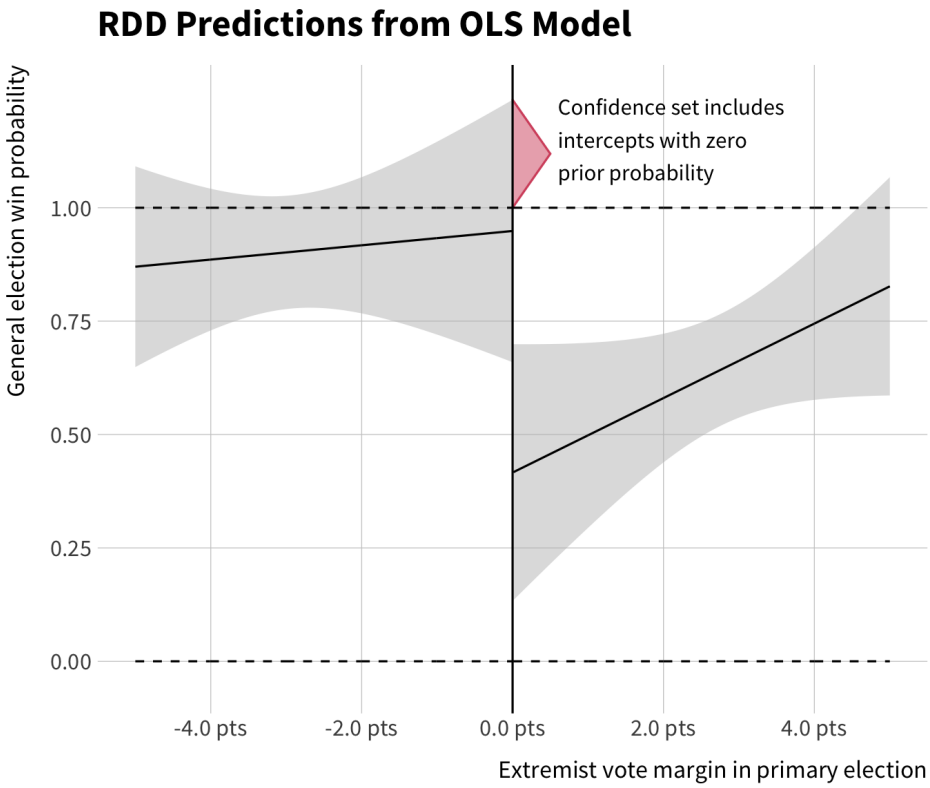


Figure 3.4: OLS estimates of the predicted probability that a candidate wins the general election. Local average treatment effect is estimated at the threshold. Treatment effect is defined by parameter estimates whose confidence sets contain values with no possibility of being true.

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. First, the model isn't fully saturated. Because the design employs a local linear regression, 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. Furthermore, the repeated sampling intuition of the frequentist approach does not guide our inferences because the data in the analysis are the population of interest. We have no ability to repeatedly sample this data generating process, so our uncertainty about our inferences must come from some other mechanism. Most importantly, because the intercept estimates are essential for defining the treatment effect of interest, the degree to which this one estimate 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 3.5 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.

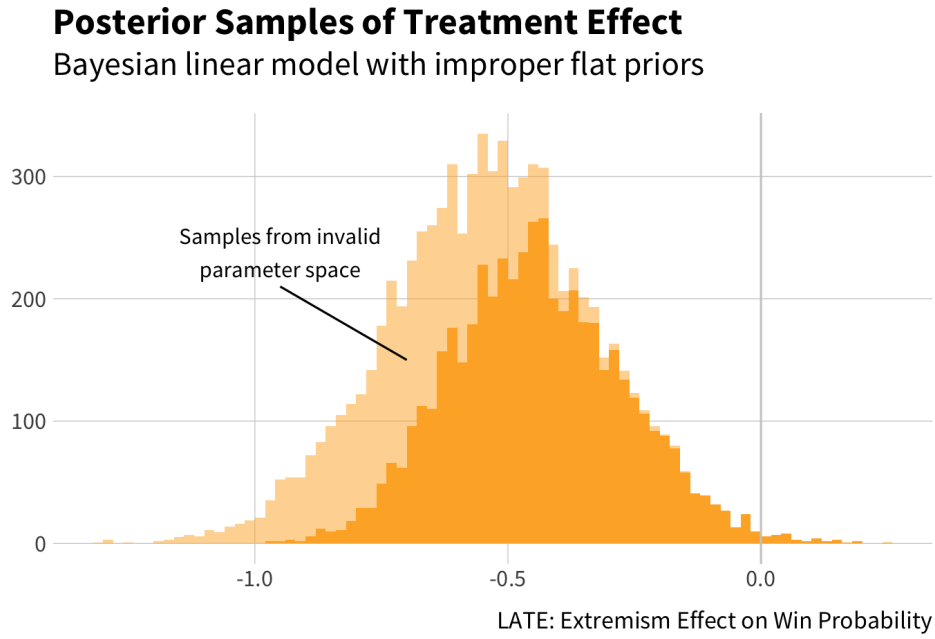


Figure 3.5: Histogram of posterior samples for the treatment effect. Using flat priors for each win probability intercept, a substantial share of the treatment distribution consists of parameters that cannot possibly occur.

The Bayesian approach begins with 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 $[0, 1]$. We remain agnostic as to which values within that interval are more plausible in the prior. The result is a uniform prior over possible win probabilities, which we apply to both intercept parameters.

$$\alpha_{w=0}, \alpha_{w=1} \sim \text{Uniform}\,(0, 1) \tag{3.27}$$

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 concentrates probability density away from treatment effects that cannot be true, while maintaining the local linear

specification that is justified by the limiting intuition of the key identification assumption. 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 despite not being flat.

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

$$\begin{aligned} y_{dpt} &\sim \text{Normal}\left(\alpha_{w[dpt]} + \beta_{w[dpt]}(\text{Extremist Primary Margin})_{dpt}, \sigma\right) \\ \beta_{\text{w}} &\sim \text{Normal}\left(0, 10\right) \\ \sigma &\sim \text{Uniform}(0, 10) \end{aligned} \tag{3.28}$$

The Normal model for the outcome data in the first line is equivalent to the Normal error term defined in (3.26). 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.

A possible retort to this model setup is a Bayesian approach would be entirely unnecessary if instead we employed a binary outcome model like logit or probit regression. These models are typically used to estimate probabilities underlying binary data in other contexts, so we entertain it here as well. Although this model contradicts the limiting intuition that the regression function is instantaneously linear at the discontinuity (as any function is instantaneously linear for an infinitesimal change in its input), I indulge this possible retort by building a Bayesian logit specification as well. This setup considers the binary election result as a Bernoulli variable with a probability parameter specified by a logit model,

$$y_{dpt} \sim \text{Bernoulli}\left(\pi_{dpt}\right) \tag{3.29}$$

$$\text{logit}\left(\pi_{dpt}\right) = \alpha^*_{w[dpt]} + \beta^*_{w[dpt]}(\text{Extremist Primary Margin})_{dpt} \tag{3.30}$$

with parameters denoted α_w^* and β_w^* to distinguish them from the α_w and β_w parameters in the linear setup.

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.

$$\begin{aligned} &\text{logit}\left(\pi_{dpt}\right) = \alpha^*_{w[dpt]}, \text{ at Extremist Primary Margin}_{dpt} = 0 \\ &\text{which implies } \pi_{dpt} = \text{logit}^{-1}\left(\alpha_{w[dpt]}\right) \end{aligned} \tag{3.31}$$

If we want to construct a prior for the treatment effect that is similar to the structural information we encoded in the linear specification, we must specify priors for the extremist and non-extremist win probabilities that are flat over valid probability values at the discontinuity. This requires a prior for α_w^* on the log-odds scale that implies a flat prior for $\text{logit}^{-1}\left(\alpha_w^*\right)$ on the probability scale. To solve this problem, we leverage the logit model's connection to the standard Logistic distribution. The logit function maps values in the $(0, 1)$ interval to any real number, and the inverse logit function maps any real number to $(0, 1)$. We accomplish a flat prior for win probabilities at the threshold using a a standard Logistic prior on the log-odds scale,

$$\alpha_w^* \sim \text{Logistic}\left(0, 1\right) \tag{3.32}$$

which becomes a flat density for $\text{logit}^{-1}\left(\alpha_w^*\right)$. It is startling at first to consider a prior as narrow as $\text{Logistic}\left(0, 1\right)$ as an uninformative prior for a key parameter. But at discussed in Section 3.3.2, the connection between prior vagueness and prior flatness is not absolute.

Flatness is only a shape. The relationship between flatness and informativeness depends on model parameterization and the scale of the data.

Figure 3.6 visualizes how the Logistic prior for the intercept on the log-odds scale becomes a flat prior on the probability scale at the threshold. The left panel shows a histogram of Logistic $(0, 1)$ simulations, and the right panel shows a histogram of the same values after they are converted to probabilities using the inverse logit function. For comparison, I also simulate a Normal $(0, 10)$ prior, which is something a researcher might pick if they wanted to be vague on the log-odds scale. Converting the wide Normal prior to the probability scale, however, shows that greater prior density on logit values far from zero translates to greater prior density over probability values very close to 0 and 1.

The fact that the wide Normal prior has strange behavior on the probability scale does not mean that it shouldn't be used in Bayesian logistic modeling. It could be an appropriate choice for specifying priors for constructs that should be understood directly on the logit scale. For instance, I give this exact prior to the slope parameters in this RDD logit model,

$$\beta_w \sim \text{Normal}\,(0, 10) \tag{3.33}$$

because I want the prior to consider a broader distribution slopes *on the logit-scale*. The lesson with these priors, as with any prior, is that prior distributions should be chosen to suit the modeling context. Elements of that context include link functions, model reparameterization, the scaling of outcome data or covariates, regularization concerns, and so on. Choosing "default priors" that always encode flatness on one scale has no guaranteed behavior for implies priors for important functions of parameters.

These prior interventions in both the Bayesian LPM and the Bayesian logit are minor. They merely encode structural information about the outcome scale. Win probabilities for extremists and non-extremists are constrained to take valid values—between 0 and 1—but the prior is otherwise agnostic about which win probabilities are more likely than others

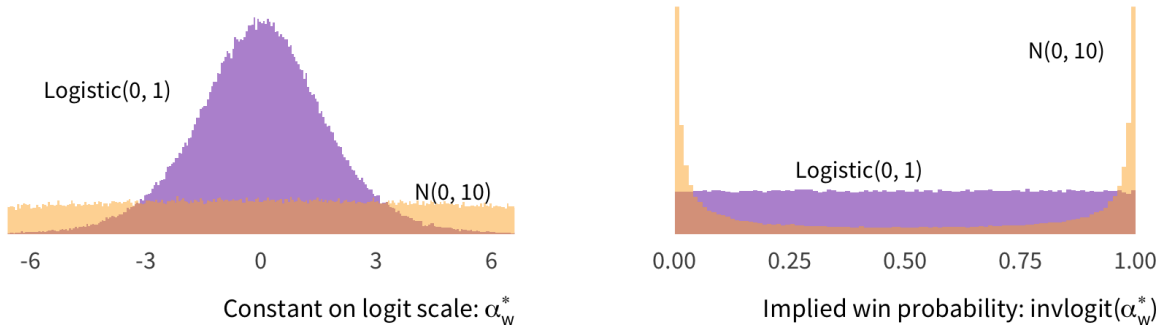


Figure 3.6: Scale invariance of logit model priors. Standard logistic prior on logit scale becomes a flat prior on the probability scale. "Diffuse" priors on logit scale imply priors on probability scale that bias toward extreme probabilities.

before seeing any data. What effect do these minor 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 creates the structural prior using the transformed Logistic distribution. The left panel shows a histogram of posterior MCMC samples for the non-extremist win probability at the threshold. The LPM at the top of the panel included 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 to obtain. The histograms in the second and third rows show the LPM and logit with the structural prior. 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. The asymmetry is

direct indicator of the way Bayesian priors added value to the analysis.

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 nearly all Bayesian modeling approaches, our priors have the effect of shrinking important effects toward 0 and reducing the variance of the effect. 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 , which is a

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 parameters. This manifested earlier in Figure 3.5, which showed that larger treatment effects were more likely to contain pathological parameters than the smaller treatment effects. 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.

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 do not realize how much information they are leaving out of their analyses.

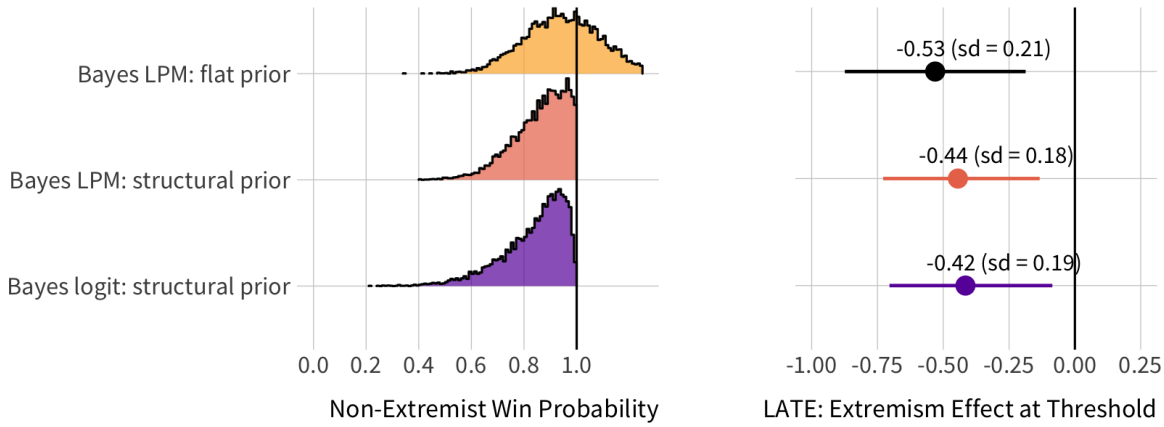


Figure 3.7: Comparison of posterior samples from Bayesian models with improper flat priors and weakly informative priors. Weakly informative models have flat priors over valid win probabilities at the discontinuity. Priors reduce effect magnitudes and variances by ruling out impossible win probabilities.

3.4.3 Models for Nonparametric Treatment Effects with Applications to Meta-Analysis

3.5 Other Frontiers of Bayesian Causal Inference

3.5.1 Beyond Estimation: Inferences About Models and Hypotheses

We can think bigger about Bayesian *inference* for a parameter as distinct from Bayesian *estimation* of the in-sample quantity. This lets us use a nonparametric data-driven estimator for the data, but the "inference" or "generalization" still has a prior. For instance a sample mean estimates a population mean without a likelihood model for the data, but inference about the population mean often follows a parametric assumption from the Central Limit Theorem that the sampling distribution from the mean is asymptotically normal (but doesn't have to, c.f. bootstrapping). Even if the point estimator we use for a mean is unbiased,

we can assimilate external information during the interpretation of the estimator (biasing the inference without biasing the point estimator). Restated: the posterior distribution is a weighted average of the raw point estimator and external information, rather than biasing the data-driven estimate directly.

Even bigger: Bayesian inference about *models* (Baldi and Shahbaba 2019). This is probably where I have to start my justification for this? The *entire point* of causal inference is to make inferences about counterfactuals given data (Rubin 1978?). Invoking Bayesian inference is really the only way to say what we *want* to say about causal effects: what are the plausible causal effects given the model/data. We do *not* care about the plausibility of data given the null (as a primary QOI). - Probably want to use Harrell-esque language? Draw on intuition from clinical research, or even industry. We want our best answer, not a philosophically indirect weird jumble. - This probably also plays into the Cox/Jeffreys/Jaynes stuff I have open on my computer.

This presumes an m-closed world(?right?), which maybe we do not like (Navarro, "Devil and Deep Blue Sea"). Me debating with myself: how to think about Bayesian model selection vs "doubly robust" estimation ideas…

Inherit material from earlier section (baldi paper)

Quasi-experiment paper (LR/BF of two models)

Conventional:

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)p(\theta)}{p(\mathbf{y})} \tag{3.34}$$

$$p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta)p(\theta) \tag{3.35}$$

Implied:

$$p(\theta \mid \mathbf{y}, \mathcal{H}) = \frac{p(\mathbf{y} \mid \theta, \mathcal{H})p(\theta \mid \mathcal{H})}{p(\mathbf{y} \mid \mathcal{H})} \tag{3.36}$$

3.5.2 Priors are the Basis for all Generalization

No-Free-Lunch theorems

3.5.3 Agnostic Causal Inference

Sidestepping priors

complexity of bayes vs. parsimony of causal inference NOT A RULE

Causal doesn't imply nonparametric, Bayesian doesn't imply complex

At any rate:

- simple case: sensitivity testing for noisy circumstances
- complex case: stabilizing highly parameterized problems
  - dynamic TCSC models, lots of parameters
  - that hierarchical conjoint thing
  - priors in high dimensions are scary: consider parameterizations and do simulations
- For nonparametric estimators, structural priors can be helpful for concentrating probability mass in sensible areas of space, since nonparametric estimators may be lower-powered than estimators that get a power boost from parametric assumptions.
- Parametric models, in term, are obvious areas where Bayesian estimates can go, but they are stacking more assumptions on top of the parametric assumptions.
- Semi-parametric models are an interesting middle ground, where we want to be flexible about the exact nature of the underlying relationships, but we want to impose some stabilization to prevent the model from behaving like crazy.

— 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

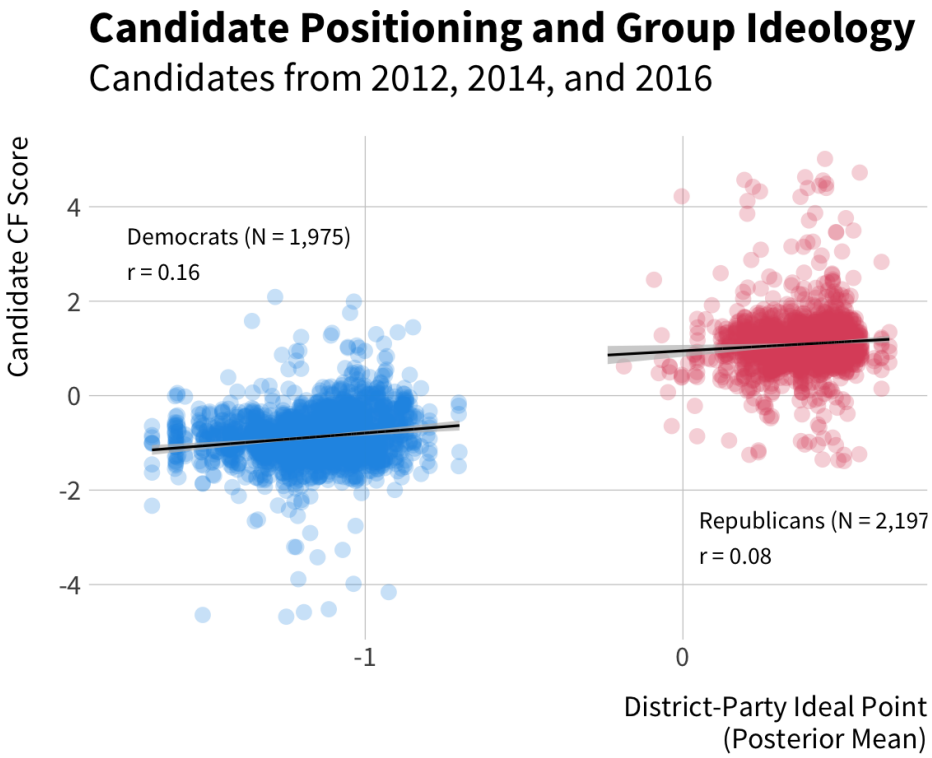


Figure 4.1: Topline relationship between district-party ideology and candidate positions. The horizontal axis plots the posterior mean for a group ideology, and the vertical axis is the dynamic CF score for primary candidates included in the DIME congressional candidate database.

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.

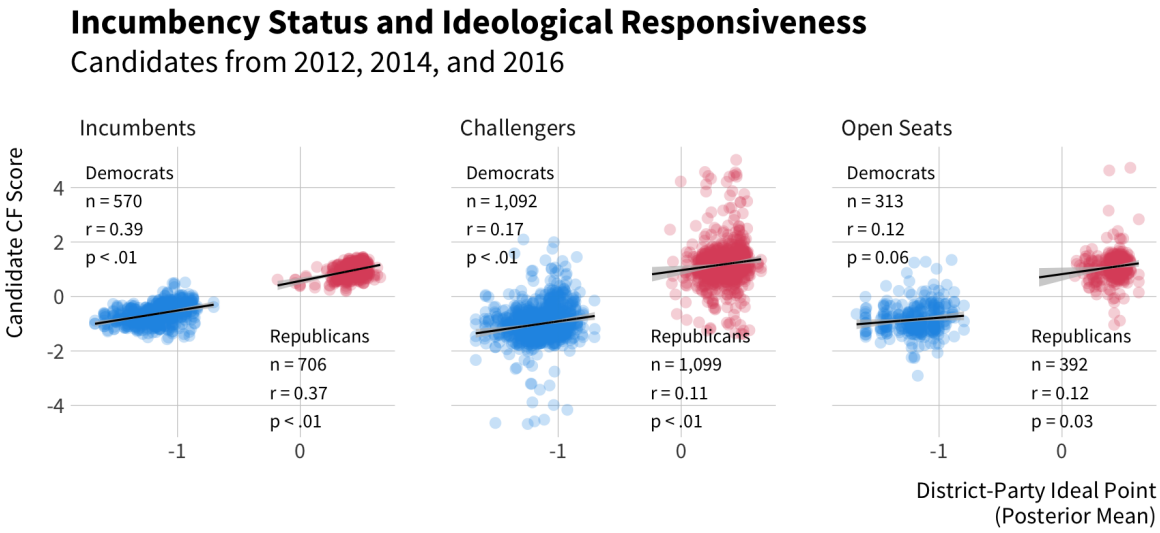


Figure 4.2: District-party ideology and candidate positioning across candidate incumbency status.

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.

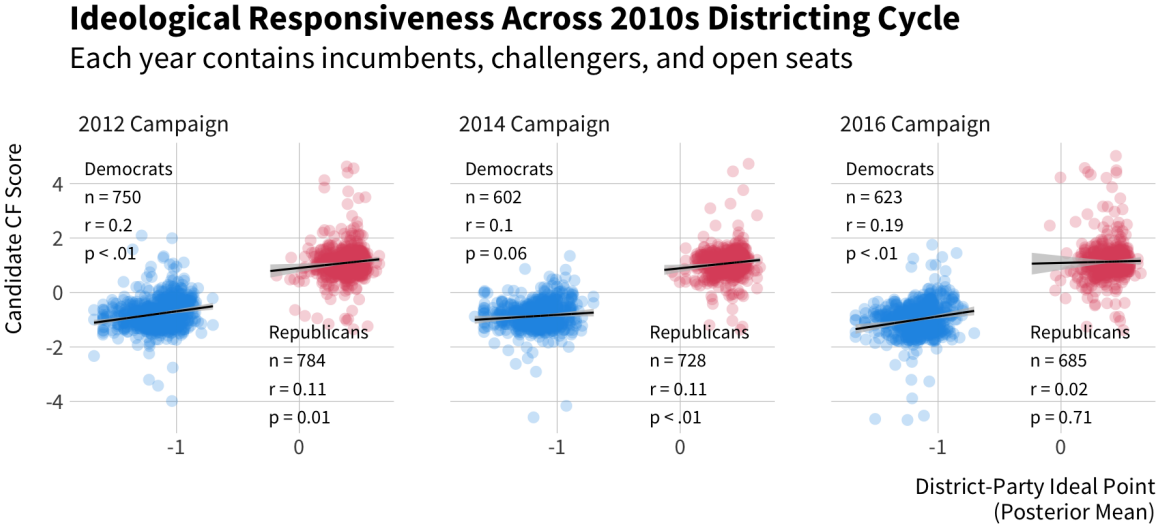


Figure 4.3: District-party ideology and candidate positioning across election cycles within the 2010s districting cycle.

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.

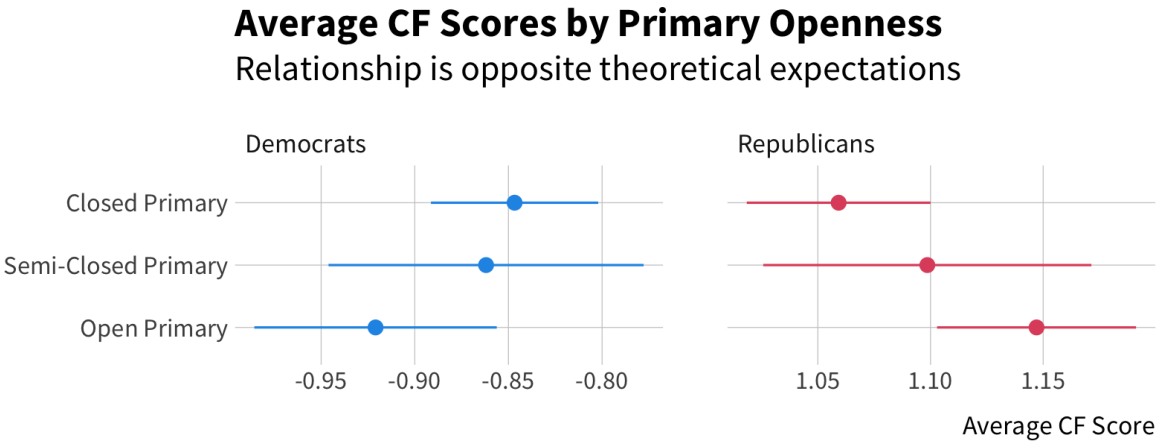


Figure 4.4: Average CF scores candidate positioning in states with closed, semi-closed, and open primary rules.

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.

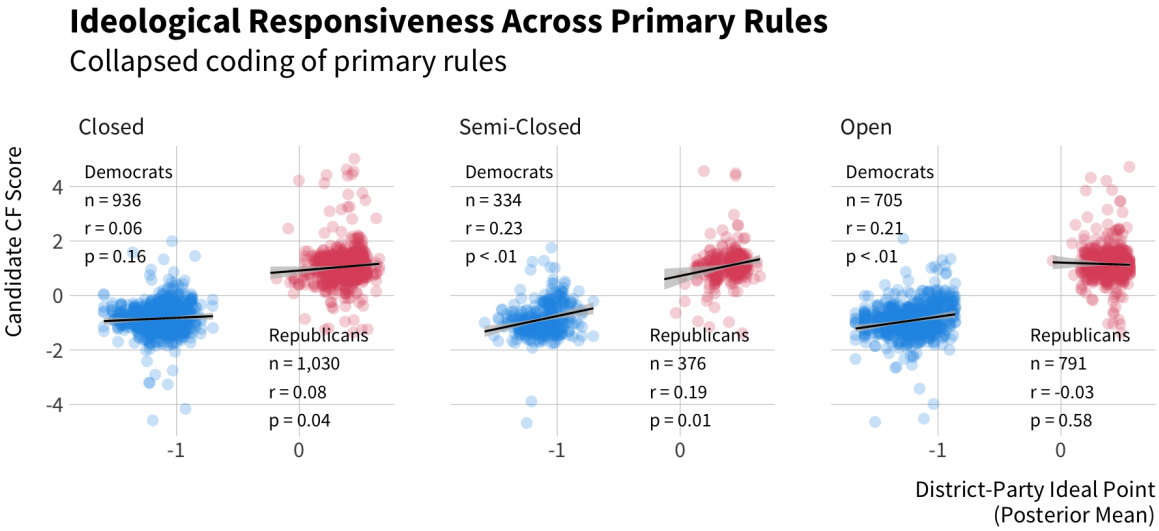


Figure 4.5: District-party ideology and candidate positioning in states with closed, semi-closed, and open primary rules.

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.

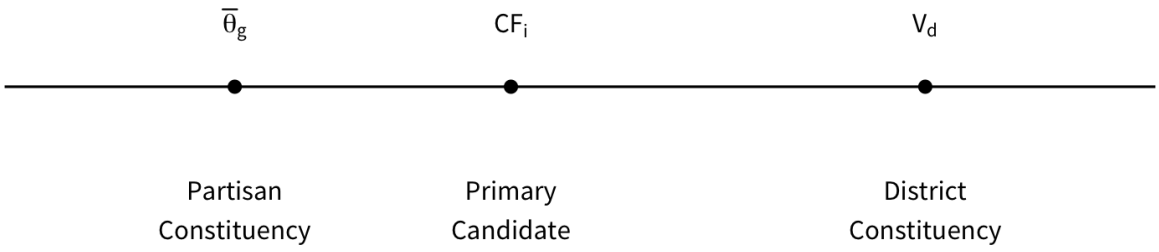


Figure 4.6: Spatial representation of key variables affecting the strategic positioning dilemma.

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

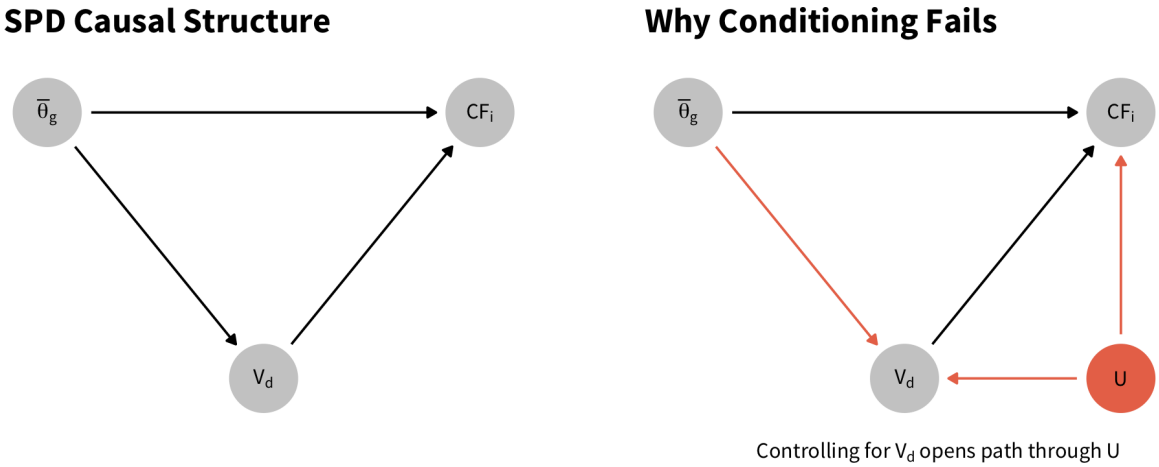


Figure 4.7: Left: A causal diagram for the strategic positioning dilemma. The district location V_d is a collider between district-party ideology $\bar{\theta}_g$ and the candidate position CF$_i$. Right: Conditioning on post-treatment variables can bias causal estimates by creating artificial associations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

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

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

[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).

even split between Republicans and Democrats in the district. We subtract the demediation 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.

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

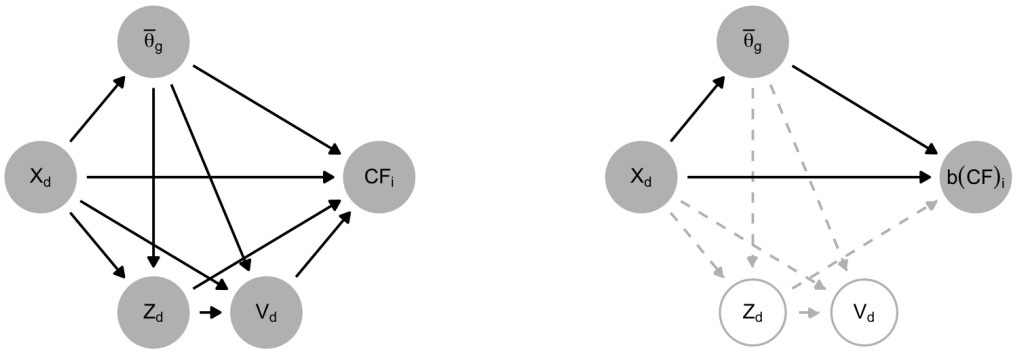
Finally, the researcher estimates the effect of the treatment on the demediated outcome, which is equivalent to the controlled direct effect on the original outcome.

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

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

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)

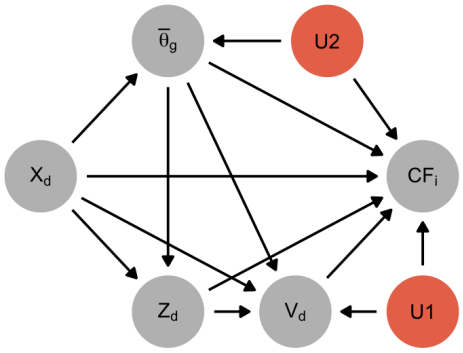


Figure 4.8: Causal graphs describing the modeling problem and sequential g estimation. The stage 1 graph identifies the effect of the past district voting (V_d) on candidate positioning (CF_i). The stage 2 treatment-outcome model subtracts the district vote effect from candidate positions and identifies 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 ($U2$).

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.

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 et al. 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]

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

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

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

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 MCMC samples to approximate this posterior distribution by picking any two values of $\bar{\theta}$ and $\bar{\theta}'$, extracting posterior samples for τ, and calculating the ACDE for each MCMC sample iteration. Posterior expectations for the ACDE can be calculated, up to Monte Carlo error, by averaging the ACDE values from a draw of MCMC samples. Uncertainty intervals for the ACDE can be estimated by noting which MCMC 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 MCMC 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. 2015b, 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]

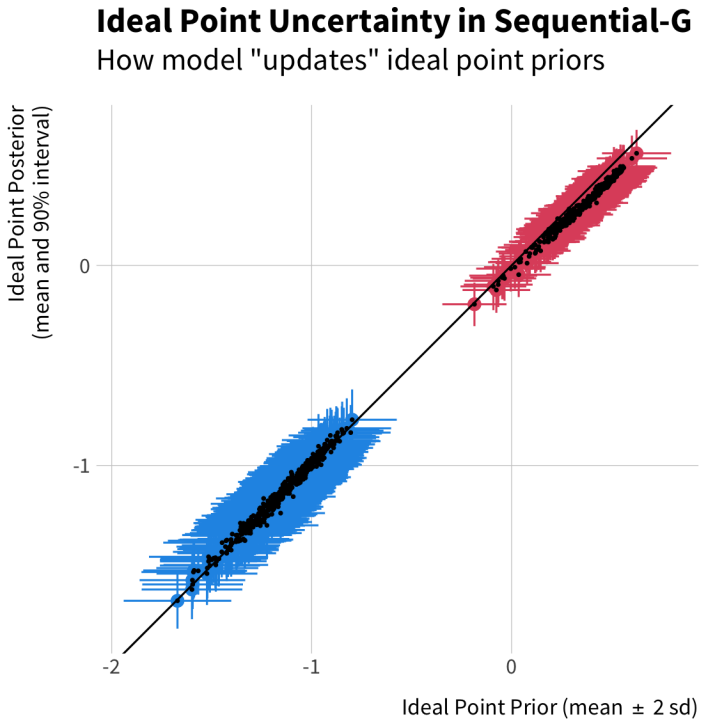


Figure 4.9: Marginal prior and posterior distributions for ideal points in sequential *g* model.

We now turn to the sequential *g* results. Figure 4.10 plots VB posterior means and 90%

[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.

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.

Sequential G Parameters
All Democratic Candidates

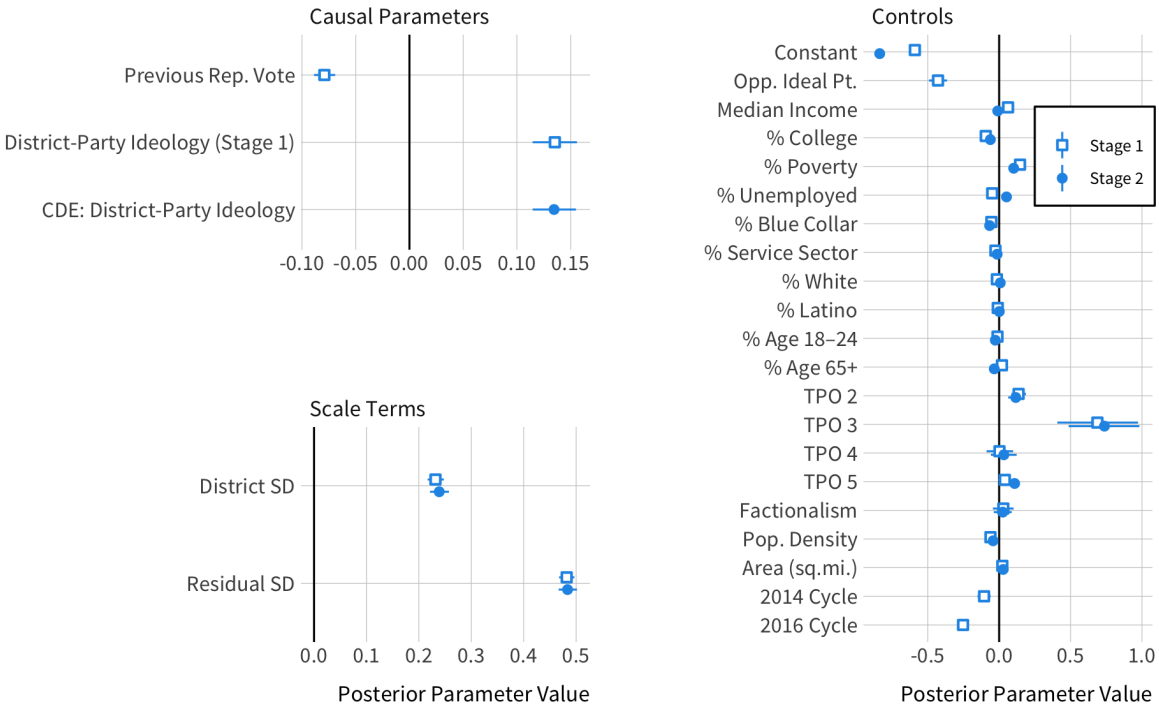


Figure 4.10: Sequential *g* results for Democratic candidates. Points and intervals are variational point estimates and 90 percent intervals from the approximate posterior distribution.

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 interpretation places greater emphasis on the standardized picture: the initial appearance

Sequential G Parameters
All Republican Candidates

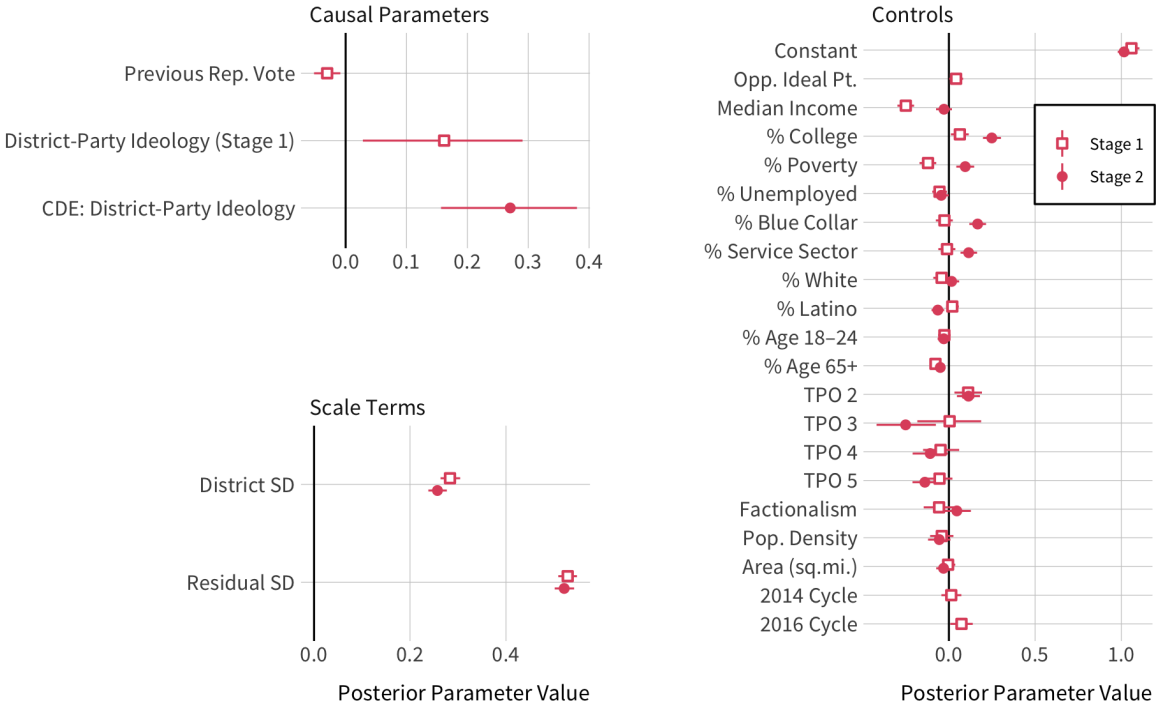


Figure 4.11: Sequential *g* results for Republican candidates. Points and intervals are variational point estimates and 90 percent intervals from the approximate posterior distribution.

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 incumbents are difficult to interpret, since the results show a negative effect of Republican

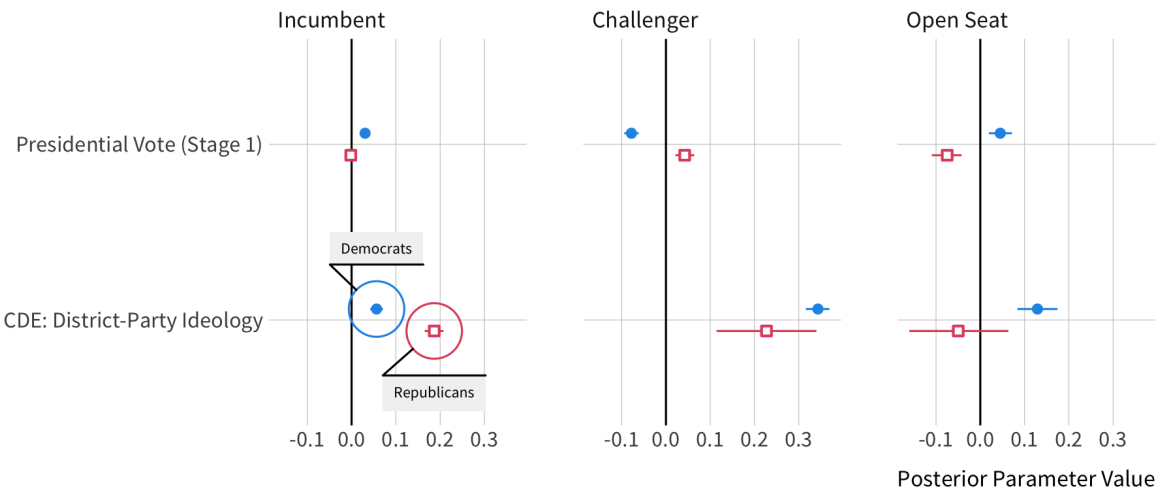


Figure 4.12: Sequential *g* results for incumbents, challengers, and open seat candidates. Points and intervals are variational point estimates and 90 percent intervals from the approximate posterior distribution.

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.

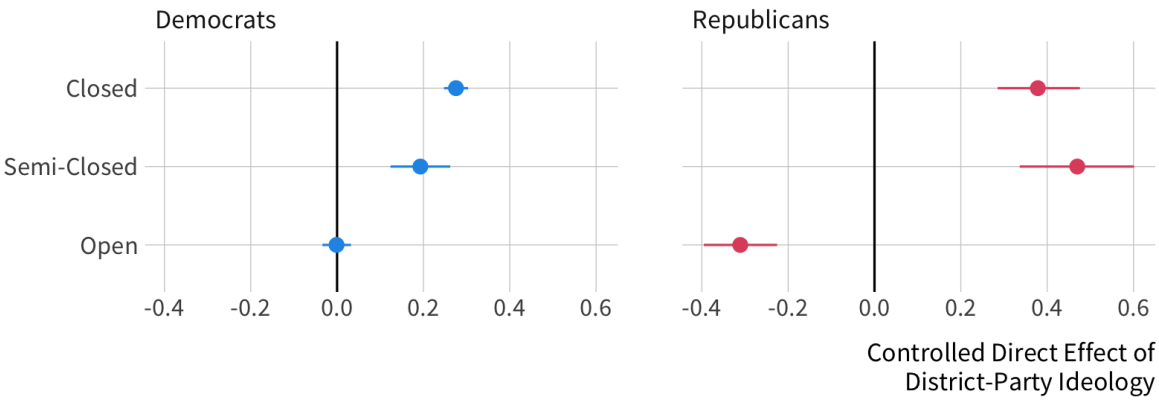


Figure 4.13: Sequential *g* results for candidates in states with closed, semi-closed, and open primary systems. Points and intervals are variational point estimates and 90 percent intervals from the approximate posterior distribution.

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

Intro:

- Now that we have district-party ideology, does it appear to matter for candidate choice?
- are more progressive/conservative districts more likely to prefer more progressive/conservative candidates?
- challenging modeling problem: the IV doesn't vary across candidates in the same primary, it only varies across primaries.
- This means that we require an approach to modeling candidate choice that incorporates the interaction between district-party ideology and candidate positioning

Overview:

- confront modeling challenges using a conditional logit approach: can only identify differences in candidate utility across candidates
- this affects which causal claims we can support, since the model can't directly identify how district-party ideology affects candidate utility in isolation

- discuss the relationship between feasible causal estimands and flexible conditional logit modeling approach for recovering how the effect of candidate positioning is heterogeneous across districts

We find:

- frame research question *as asked in the statistics*
- Democratic primaries exhibit ideological voting: there is an ideological "sweet spot" where candidates should position themselves to maximize their win probability
- However, the location of this sweet spot does not vary greatly across districts, meaning that variation in district-party ideology
- Among Republicans, there is much weaker ideological voting. Candidate value is not strongly related to their ideological positioning at all, leaving little room for effects to vary by districts.

5.1 Modeling Candidate Choice

What role does ideology play when constituencies nominate a primary candidate? 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 insight of these spatial voting models. A candidate is most appealing

to a constituency when the candidate position (represented on the left–right ideological continuum) matches the constituency's preferred ideological outcome. The candidate is less appealing, or provies less *value* or *utility*, as their ideological distance from the constituency ideal point grows. This utility loss accumulates regardless of whether the candidate is too progressive or too conservative.

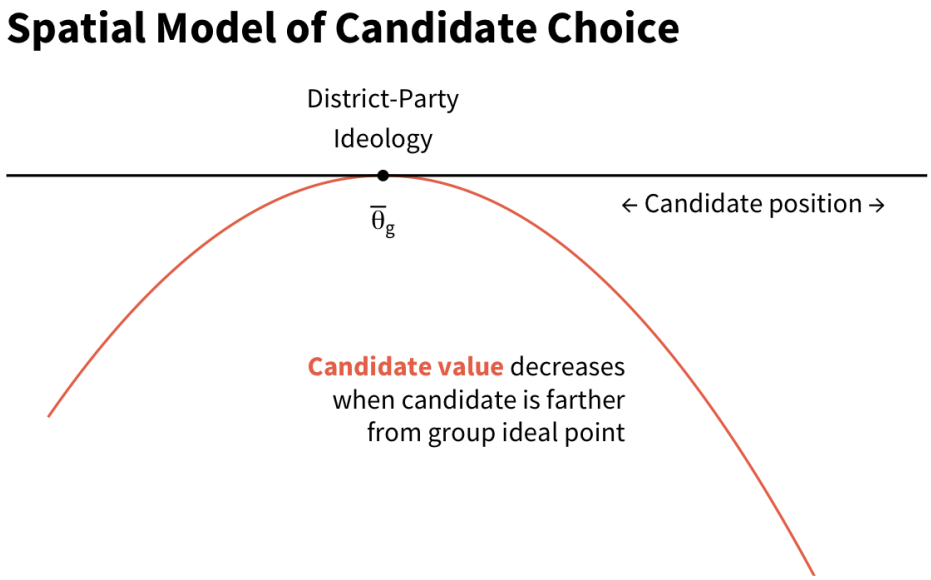


Figure 5.1: A spatial voting model's description of candidate value (utility) as a function of candidate position and district-party ideology. Candidate value is maximized at the group ideal point $\bar{\theta}_g$ and decreases in either direction. The example in this plot assumes quadratic utility loss.

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. The other key drawback in existing research is that many studies try to understand the effects of primaries using data from incumbent candidates or general election candidates, which can introduce selection bias by observing only primaries winners instead of all primary candidates including primary losers (e.g. Ansolabehere, Snyder, and Stewart 2001; Brady, Han, and Pope 2007; Burden 2004; Clinton 2006; 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, these studies cannot observe whether extreme candidates are actually preferred over less extreme candidates . Studies that avoid this problem exist but are less common (Porter and Treul 2020). The study design below remedies both of these problems by modeling candidate choice in primary elections using a conditional logit likelihood and using basis-splines to model nonlinear effects of candidate ideology.

As reviewed in Chapter 1, there are several reasons to doubt that House primary voting predominantly follows a spatial voting model. Fewer 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 aware of policy differences (Lelkes 2019), but it is costly for primary voters to learn about the contrasting policy views of primary candidates. Voters may also be strategic, preferring the candidate that they believe is most likely to win the general election even if that candidate isn't closest to their ideal point (Simas 2017). Candidate 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 bolster candidates' primary success. Candidate ideology may serve a selection function even before the primary election itself if moderate candidates are deterred from entering the primary in the first place (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 and others 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) selects a candidate i by comparing the utility they would receive from each candidate in the race r. Suppose that this utility μ_{ir} contains a systematic component u_{ir} and a random component e_{ir}.

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

The probability that i is chosen is defined as the probability that μ_{ir} is greatest among all alternatives. Because error term e_{ir} is not known, this probability is calculated as a function

of the systematic components that can be predicted, typically using the softmax function.

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

which follows from an 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 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, under the standard conditional logit approach, cannot *directly* affect which candidate is nominated by holding all else fixed. This tracks with the spatial model picture laid out in Figure 5.1: shifting the district-party ideal point $\bar{\theta}_g$ left or right does affect utility, but only because it changes the distance between $\bar{\theta}_g$ and the candidate location. In other words, 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). Building a statistical model that enables this interactivity is an important contribution of this research design.

[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.

[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).

This conditional logit model's identifiability constraint matters for causal inference as well, because it affects which causal quantities make sense in this context. Consider the potential outcome $\mu_{ir}\left(\bar{\theta}_{g[r]}, \text{CF}_{ir}\right)$, the systematic utility that is a function of district-party ideology and the candidate's ideological position. Imagine that we intervene on district-party ideology and measure the average utility effect[3] of setting $\bar{\theta}_g = \theta$ versus some other value θ', $\mathbb{E}\left[\mu_{ir}\left(\theta, \text{CF}_{ir}\right) - \mu_{ir}\left(\theta', \text{CF}_{ir}\right)\right]$. This effect does exist for individual candidates: making the district-party more liberal or conservative would affect the primary electorate's candidate utility by increasing or decreasing the ideological distance between the district-party and the candidate. But because multiple candidates can run in the same race, and the primary electorate has the same district-party ideology each time they consider a new candidate, the average causal effect of district-party ideology implicitly averaging all of the conditional effects for all candidates. This isn't unusual in the abstract: any average causal effect is an average over potential heterogeneities. But this is complicated by the fact that the conditional logit model does not naturally measure chooser-level effects (like district-party ideology), making it impractical to condition on other district-level characteristics to identify the average effect of district-party ideology on candidate choice.

It is much simpler, instead, to consider the average effect of candidate positioning on candidate utility. This average effect is also an implicit average over all interactions with district-party ideology, but conditioning on candidate features that confound the effect of CF scores is much simpler with conditional logit, so causal identification of choice-level effects is more analytically straightforward as well. The conditional average effect of CF scores on candidate utility would thus be $\mathbb{E}\left[\mu_{ir}\left(\theta, \text{CF}\right) - \mu_{ir}\left(\theta, \text{CF}'\right) \mid C_{ir} = c, r\right]$, for a comparison of two values CF and CF′, fixing the district-party ideology at θ and conditioning on other

[3] 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).

candidate-varying attributes $C_{ir} = c$ and the race r.[4]

Because this project is focused on the added value of observing district-party ideology, we go one step further to model effect heterogeneity over district-party ideology explicitly, rather than be content to average over it implicitly. I approach this from an effect modification perspective, meaning that heterogeneous effects do not reflect the causal effects of interventions on district-party ideology, but instead reflect causal effects of CF scores conditional on a given district-party ideology value. 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).

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

Figure 5.2 plots a causal graph of the system under consideration. The causal effect of candidate position CF_{ir} on candidate utility μ_{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 $\mathrm{CF}_{ir} \rightarrow \mu_{ir}$, represented with the path $\bar{\theta}_g \rightarrow \mathrm{CF}_{ir}$ and no direct path between $\bar{\theta}_g$ and μ_{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 is why the effect heterogeneity across districts has no causal interpretation.

[4]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.

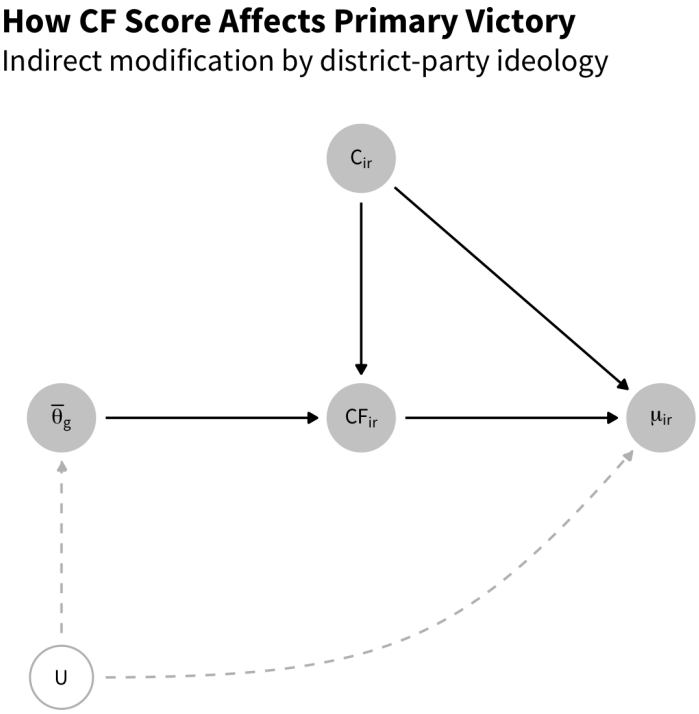


Figure 5.2: Causal Diagram of CF score effect on win probability. District-party ideology is an indirect modifier because has no direct effect on primary outcomes except through candidate proximity. Post-treatment candidate features P_i and unobservables U are uncontrolled but do not compromise identification.

5.1.2 Causal heterogeneity through continuous interaction

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 model of choice. 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 shared space that can

be interpreted like a "distance" metric, allowing candidate utility to increase or decrease as a function of the distance metric. The second tactic: The distance metric's 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 modeling 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.

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

The systematic utility u_{ir} is given a vague definition for the time being. I use $f()$ to represent a flexible function of candidate i's CF score and the district-party public ideology 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 point measures, the two measures are not constructed in the same ideal point space. The first step for constructing the spline is to create a function that effectively maps these two measures into a common

space. Let Δ_{ir} be a linear combination of CF_{ir} and $\bar{\theta}_g$,

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

The linear combination represents an assumption that CF score space and district-party ideology space are linear transformations of one another, similar to the way Aldrich–McKelvey scaling estimates linear mappings between ideology spaces (Aldrich and McKelvey 1977; Hare et al. 2015). The second line of (5.5) restricts the coefficients to have a norm of 1. This places an identifiability restriction on the location and scale of the Δ space. Furthermore it implies a 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.[5] The linear transformation gives Δ_{ir} the algebraic interpretation of a distance measure between the candidate's CF score and the district-party public ideology in the Δ space.

I then use Δ_{ir} to construct a set of basis functions, using a degree-3 polynomial basis with 30 knots across the range of Δ_{ir}.[6] 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}.

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

[5] Restricting the value of $\beta > 0$ would identify the rotation of the Δ space as well. Such a restriction would improve the interpretability of the Δ space, but it would hinder Bayesian estimation by introducing unnecessary discontinuities in the posterior distribution.

[6] 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 re-calculating Δ in each model, as CF scores and district-party ideology have different ranges for Republicans and Democrats.

The spline regression enables a continuous interaction effect between CF scores and district-party ideology. Because the basis functions are a nonlinear function district-party ideology, 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 interacting 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 districts with different district-party ideologies. Interacting two continuous variables through the spline function is much more flexible than a multiplicative interaction between CF scores and district-party ideology, which would fail to capture both the utility optimum predicted by spatial voting models as well as any other non-constant interactions. The intuition of the linear combination is also superior to the multiplicative interaction, since the notion of a common ideal point metric is a more faithful operationalization of the underlying spatial model. The multiplicative interaction has no comparable 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 effectively 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 multitude of possible Δ spaces. 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. As a result, the distribution of spline regressions is well identified from the data even if its component parameters—α, β, and spline weights ϕ_k—are not strongly identifiable on their own. Stated differently, this model sacrifices some interpretability in order to fit a more flexible regression function, but this trade is worth it in order to capture nonlinear patterns in spatial voting while avoiding specific utility functions or ideal point mappings.

5.1.3 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, and I drop all blanket and top-two primaries since those races are not limited to 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 primaries for 2012 and 2014, which led to lots of missing data. Missing primary outcomes in the DIME were supplemented with data from the PTP. Because candidate identifiers were not easily reconcilable using candidate identifiers,[7] I merge the databases using the probabilistic record-linkage algorithm developed by Enamorado, Fifield, and Imai (2018), linking candidates by name, state, district number, election cycle, and political party. In cases where the DIME and the PTP disagree about primary outcomes, 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

[7] Candidate IDs in the DIME are regenerated with each vintage of the database, creating over-time inconsistencies candidate identifiers. As a result, the DIME identifiers in PTP do not match the identifiers in more recent DIME vintages.

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 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).

5.1.4 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. Although 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,

$$\phi_k \sim \text{Normal}\,(0, \sigma) \tag{5.8}$$

where σ 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 σ.

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

The non-centered parameterization stretches a standard Normal distribution in order to create a Normal distribution with a scale of σ. This parameterization is valuable for Bayesian estimation because it de-correlates random variables in the posterior distribution, creating an easier posterior geometry for MCMC transitions. I give the scale factor σ a half-Cauchy prior,

$$\sigma \sim \text{Half-Cauchy}\left(0, 1\right) \tag{5.10}$$

which regularizes the scale value toward zero but has a flatter tail to allow strong signals from the data to depart from the prior. This Normal-Cauchy mixture is similar to the "horseshoe prior" and its variants (Carvalho, Polson, and Scott 2010; Piironen and Vehtari 2017; Piironen, Vehtari, and others 2017), which is a popular prior for estimating sparse coefficients with regularization.[8] The left-side panel in Figure **??** plots a histogram of simulated draws from this prior, which features a spike near zero and characteristically long Cauchy tails.[9]

The right panel 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 the spline function

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

[9] 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.

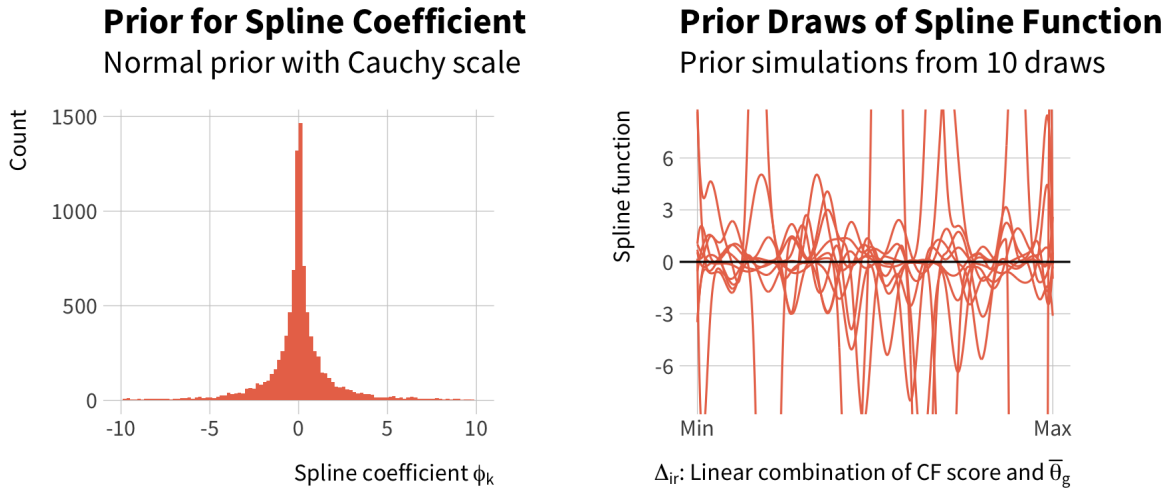


Figure 5.3: Prior draws of spline coefficient and spline function. Left: histogram of prior draws for an individual spline coefficient. Right: draws from the implied prior over spline functions.

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 Cauchy prior on σ has thicker tails that allow larger values to occur more frequently. The shape of the Cauchy tail retains enough flexibility to detect a spike in utility even if the center of the prior concentrates spline functions near zero.

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, 2020).

5.2 Findings

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

$$\text{CUE}\left(\text{CF}, \text{CF}', \theta, c\right) = \mathbb{E}\left[f\left(\text{CF}, \bar{\theta}_g\right) \mid \bar{\theta}_g = \theta, C_{ir} = c\right] \tag{5.12}$$

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

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

$$p\left(\text{CUE}\left(\text{CF}, \text{CF}', \theta, c\right)\right) = \int p\left(f\left(\text{CF}, \theta\right) - f\left(\text{CF}', \theta\right) \mid \theta, \alpha, \beta\right) \tag{5.14}$$

$$\times\, p\left(\alpha, \beta, \boldsymbol{\phi}\right)\, \mathrm{d}\alpha\, \mathrm{d}\beta\, \mathrm{d}\boldsymbol{\phi} \tag{5.15}$$

The conditional average effect on candidate utility can be gleaned by simply contrasting the function values in Figure **??** for two different CF score values.

The conditional logit is a linear model for latent candidate utility, but causal effects on candidate win probabilities can be derived from the model as well. 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 value. Second, a primary race can feature a variable number of candidates, so a utility shock can have a smaller effect on win probability if there are more candidates in the race. 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. The CWE is the expected difference in softmax functions of candidate utility between CF score values CF and CF'.

$$\text{CWE}\left(\text{CF}, \text{CF}', \theta, c, r\right) = \mathbb{E}\left[\text{softmax}\left(u_{ir}(CF) \mid \theta, c, r\right]\right. \tag{5.16}$$

$$\left. - \mathbb{E}\left[\text{softmax}\left(u_{ir}(CF') \mid \theta, c, r\right]\right.\right. \tag{5.17}$$

where $u_{ir}\left(CF_{ir}\right)$ represents a candidate utility. In turn, the probability distribution is

$$p\left(\text{CWE}\left(\text{CF}, \text{CF}', \theta, c, r\right)\right) = \int p\left(\text{softmax}\left(u_{ir}(CF) - \text{softmax}\left(u_{ir}(CF') \mid \theta, c, r\right)\right.\right. \tag{5.18}$$

$$\left. \times\, p\left(\alpha, \beta, \boldsymbol{\phi}\gamma\right)\, \mathrm{d}\alpha\, \mathrm{d}\beta\, \mathrm{d}\boldsymbol{\phi}\gamma\right. \tag{5.19}$$

Figure **??** visualizes causal effects on win probability. The figure contains four comparisons in four different types of races: races with two and three candidates with no incumbents, and races with two and three candidates where one of the competing candidates is an incumbent. The function plotted in each panel is the conditional win effect (vertical axis) of a candidate occupying a certain CF score (horizontal axis) compared to the average CF score in their party.

5.3 Discussion

5.3.1 Other nonparametric Bayes approaches for flexible causal inference

Matching under measurement error in the covariates: https://amstat.tandfonline.com/doi/full/10.1080/01621459.2015.1122601?casa_token=an-4oMYQN1AAAAAA%3AcmZp5oIxAb5Hs2kkR53JgHrrSXOd1HrLYDwQiuXOw_BaJVWcgG7ADtFcSQrIUQ2qRJTT999k4D-n

5.3.2 Causal inference with latent variables and measurement models

Group IRT Model

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:  20876c5f12cd9cffffb0cd2fb722cea7dcb8fed6
## Author:  Michael DeCrescenzo <mgdecrescenzo@gmail.com>
## When:    2020-09-13 00:41:42 GMT
##
##      fix post-estimation thetas, cauchy(0,1)
##
## 1 file changed, 45 insertions, 29 deletions
## code/05-voting/stan/choice-combo.stan | -29 +45  in 3 hunks
```

This version of the document was generated on 2020-09-12 21:19:25.

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.

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.

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*.

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, and others. 2006. "The case for objective bayesian analysis." *Bayesian analysis* 1(3): 385–402.

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.

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.

Borenstein, Michael et al. 2011. *Introduction to meta-analysis*. John Wiley & Sons.

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.

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."

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.

Chipman, Hugh A et al. 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.

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.

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, and others. 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.

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. 2004. "Parameterization and bayesian modeling." *Journal of the American Statistical Association* 99(466): 537–545.

Gelman, Andrew. 2017. "Theoretical statistics is the theory of applied statistics: How to think about what we do." https://statmodeling.stat.columbia.edu/2017/05/26/theoretical-statistics-theory-applied-statistics-think/.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, Andrew, and Gary King. 1990. "Estimating incumbency advantage without bias." *American Journal of Political Science*: 1142–1164.

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, 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.

Gerber, Alan S, Donald P Green, and Edward H Kaplan. 2004. "The illusion of learning from observational research." In *Problems and methods in the study of politics*, eds. Ian Shapiro, Rogers Smith, and Tarek Masoud. Cambridge University Press, p. 251–273.

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, Paul Pierson, and others. 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 et al. 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*.

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.

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.

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.

Kastellec, Jonathan P et al. 2015a. "Polarizing the electoral connection: Partisan representation in supreme court confirmation politics." *The journal of politics* 77(3): 787–804.

Kastellec, Jonathan P et al. 2015b. "Polarizing the electoral connection: Partisan representation in supreme court confirmation politics." *The journal of politics* 77(3): 787–804.

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.

Lancaster, Tony. 2000. "The incidental parameter problem since 1948." *Journal of econometrics* 95(2): 391–413.

La Raja, Raymond, and Brian Schaffner. 2015. *Campaign finance and political polarization: When purists prevail*. University of Michigan Press.

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*.

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.

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, and others. 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.

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.

Nyhan, Brendan, Christopher Skovron, and Rocío Titiunik. 2017. "Differential registration bias in voter file data: A sensitivity analysis approach." *American Journal of Political Science* 61(3): 744–760.

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.

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. 2017. "On the hyperprior choice for the global shrinkage parameter in the horseshoe prior." In *Artificial intelligence and statistics*, PMLR, p. 905–913.

Piironen, Juho, Aki Vehtari, and others. 2017. "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.

Ratkovic, Marc. 2019. "Rehabilitating the regression: Honest and valid causal inference through machine learning."

Ratkovic, Marc, and Dustin Tingley. 2017. "Causal inference through the method of direct estimation." *arXiv preprint arXiv:1703.05849*.

Ratkovic, Marc, Dustin Tingley, and others. 2017. "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. 1978a. "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, Donald B. 1978b. "The phenomenological bayesian perspective in sample surveys from finite populations: Foundations." *Imputation and the Editing of Faulty or Missing Survey Data*: 10–18.

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.

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.

Skovron, Christopher, and Rocio Titiunik. 2015. "A practical guide to regression discontinuity designs in political science." *American Journal of Political Science* 2015: 1–36.

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*.

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.

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.