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, Associate 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

Abstract **ii**

Contents **iv**

List of Tables **v**

List of Figures **vi**

Acknowledgments **vii**

1 Introduction: Policy Ideology and Congressional Primaries **1**

1.1 Policy Preferences and the Strategic Positioning Dilemma 3

1.2 Does the Strategic Positioning Dilemma Describe Primary Representation? . 10

1.3 Project Outline and Contributions 28

2 Hierarchical IRT Model for District-Party Ideology **32**

2.1 Spatial Models and Ideal Points 33

2.2 Item Response Theory 39

2.3 Modeling Party-Public Ideology in Congressional Districts 46

2.4 Dynamic Model 50

2.5 Bayesian Estimation and Computation 54

2.6 Testing the Model with Simulated Data 67

2.7 Ideal Point Estimates for District-Party Publics 67

2.8 Testing the Model with Simulated Data 67

2.9 Data Sources 67

2.10 Model Results 67

3 Bayesian Causal Models for Political Science **69**

3.1 Overview of Key Concepts 70

3.2 Bayesian Causal Modeling 82

3.3 Making Sense of Priors in Causal Inference 93

3.4 Pragmatic Bayesian Modeling 105
3.5 Bayesian Opportunities 109
3.6 Other Frontiers of Bayesian Causal Inference 121

4 Constituent Ideology and Candidate Positioning: Bayesian Structural De-Mediation Model 123
4.1 Constituency Preferences and Candidate Positioning 124
4.2 How Partisan Preferences Affect Candidates 126

5 District-Party Ideology and Primary Election Outcomes: Bayesian Double Machine Learning 127
5.1 Machine Learning and Causal Inference 128
5.2 Bias and Variance in Causal Identification 130
5.3 Empirical Setup 130
5.4 Findings 130

Group IRT Model 131

Colophon 132

References 134

List of Tables

List of Figures

1.1 Two districts with identically located median voters 18
1.2 Non-identifiability of partisan group preferences from district vote shares 20
1.3 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. 25

2.1 An Actor and two policy outcomes (Left and Right) represented as locations in ideological/policy space 33
2.2 A representation of quadratic utility loss over policy choices 36
2.3 Examples of item characteristic curves under different item parameter assumptions 43
2.4 The region of the probit model's latent index that maps to response probabilities between 1 and 99 percent. 62
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). 65
2.6 Posteriors 68

3.1 A causal graph where the effect of A on Y is confounded by U. 74
3.2 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. 97
3.3 A spectrum of attitudes toward priors. 97

3.4 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. . . . 108
3.5 OLS estimates of the predicted probability that a candidate wins the general election. Local average treatment effect is estimated at the threshold. Treatment effect is defined by parameter estimates whose confidence sets contain values with no possibility of being true. 114
3.6 Histogram of posterior samples for the treatment effect. Using flat priors for each win probability intercept, a substantial share of the treatment distribution consists of parameters that cannot possibly occur. 115
3.7 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. 119
3.8 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. 121

Acknowledgments

Many people supported this me through this project. Thanks to…

- My family, who would do anything in their power for me
- My committee: Barry Burden, Ken Mayer, Ellie Powell, Alex Tahk, and Mike Wagner.
- Members in the department who provided institutional support: the Election Research Center, Scott Straus, Deb McFarlane and others in the department administration, the "Beasts of Burden" roundtable group, and members of the Fall 2016 prospectus workshop course.
- Faculty and students at other universities who shared data or provided advice: Robert Boatright, Devin Caughey, David Doherty, Andrew Hall, Seth Hill, Georgia Kernell, Shiro Kuriwaku, Rachel Porter, Andrew Reeves, Michael Ting, and Sarah Treul.
- Software developers and users, whose tools and supporting community helped bring the project to life in my computer: Mara Averick, Jenny Bryan, Andrew Heiss, Matthew Kay, Mike Kearney, TJ Mahr, Thomas Lin Pederson, and Kara Woo.
- My closest friends in the program: Micah Dillard, Jordan Hsu, Rachel Jacobs, Anna Meier, Anna Oltman, Rachel Schwartz
- The Bike Gang: Josh Cruz and Erin Nelson
- My friends in other cities who always cheered me on: Marissa Finazzo, Zach Gray, Spencer Melgren, Jake Moore, Daniel Putman, Ian Servantes, India Watts
- Cafes and their staff, especially Black Locust Cafe and Johnson Public House.
- People who provided feedback in workshops: Devin, Evan, Rochelle, Blake, David C, Marcy,

— 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 alternatives, 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), 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 attentive to politics and hold stronger, non-centrist issue preferences compared to the average general-election voter.[1] As a result, competition in the primary stage may present a more acute electoral threat to partisan candidates than general elections do—a "strategic-positioning dilemma" that leads candidates to take ideological issue stances in favor of convergent stances that target the median voter (Aldrich 1983a; Brady, Han, and Pope 2007; Burden 2004; Hill 2015).

The strategic positioning dilemma (SPD) is a central theoretical feature of this project, and tests of the SPD are key empirical contributions in the following chapters. The sections

[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).

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 liberalism in a measurement modeling context, which we return to in Chapter 2. Policy liberalism affects voters' issue beliefs, and while issue beliefs can be measured using a survey, policy liberalism itself cannot be. Instead, policy liberalism exists in a latent space, and survey items on specific issues reveal only limited information about voters' locations in the latent space. This is different from constructing ideology as a simple average of policy beliefs using additive issue scales. Although the additive scaling approach is common, it implies an assumption that all items about all issues are equally informative about ideology. Modern measurement approaches doubt 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 Jr 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 similarity.

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 and the Republican Parties. I call these two groups of constituents *district-party groups*. All 435 congressional district contains 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 that decomposes the district electorate in this way 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 among Democrats differs from policy ideology among Republicans even, and how these 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

Burden, AnSnSt team, Bonica, Rogowski and Langella,

Henderson, Barberá,

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).

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 elect conservative candidates.

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 questionable theoretical underpinning. First, the SPD depends on lofty assumptions about the sophistication of voter behavior in primaries. 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 candidates; voters can be broadly confident about the issue disagreements that distinguish Republican candidates from Democratic candidates (Hill 2015). Primary elections, however, occur most of the time between candidates in the same party,[2] which denies voters' the informational shortcut of a candidate's party affiliation. Primary elections often occur during months when voters are paying less attention to politics, and the press typically cover primary campaigns less closely than general election campaigns.

In these low-information environments, voters may decide to cast their ballot for various non-policy reasons. They may 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. 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.

Indeed, the notion that "voters matter" is a theoretical orientation toward primaries that not all scholars of electoral representation share to the same degree. 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).

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

[2] 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.

more conservative on many issues than they actually are, which could affect how accurately candidates position themselves in relation to constituent views. It is possible, however, that candidates have a *relative* perception of voter accuracy: they can tell if constituency *A* is more conservative than constituency *B*, even if they can't identify the *absolute* degree of conservatism in either constituency. A candidate's ability to respond to their constituents may itself be downstream from other notions of candidate "quality." For instance, better-financed candidates may have better information on their constituents' views from pollsters or other campaign consultants. Candidates may also judge that their policy position is less important than other symbolic communications touting their successes as an incumbent or their "outsider" status, which leads voters to have less policy-relevant information about the candidates. The efficacy of these non-policy appeals may be correlated with the actual ideological positions of the candidates—for instance, if incumbents are more skilled at perceiving district-party ideology, positioning themselves in relation to the partisan base, and communicating symbolically powerful messages.

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 general premises of primary representation and competition on which it rests, the results are ambiguous and often contradictory. This section reviews existing research on electoral representation and accountability in Congress to provide a background for the substantive innovations in this project.

"Representation" means many different things in the study of elections. Many theories invoke dichotomies of representational styles, such as "delegate" versus "trustee" models of representation, or "descriptive" versus "substantive" representation [Burke (2012; Canon 1999; Phillips 1995; Pitkin 1967). Much of the empirical work in the study of Congressional elections and Congressional primaries has operated within a delegational framework, asking whether electoral candidates campaign by appealing to local policy attitudes, whether elected legislators according to their constituents' policy preferences, and whether constituents vote for the candidate who most closely represents their preferences (e.g. Miller and Stokes 1963; Campbell et al. 1960; Erikson and Wright 1980; Poole and Rosenthal 1997).

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).[3] 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

[3] 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. The incentives for moderation are more theoretically important than the whether we observe perfect candidate convergence at the median voter.

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 (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. 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 (Ansolabehere, Snyder, and Stewart 2001).

The Downsian logic is a plausible explanation for the electoral benefits of moderation, but what explains the non-moderate stances? Political scientists have explored several theories whose underlying mechanisms are distinct from the SPD notion of competing constituencies. Even without primaries, 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 Jr 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, who are then more likely to turn out to vote, staff campaigns, and engage in other forms of pro-party activism than moderate voters are (Aldrich 1983b).

- parties gate-keep the ballot (direct primary?)
- turnout among partisans
- party activists (campaign labor)
-

Why non-moderate stances?

- xyz
- primaries…

Qualitative evidence suggests that candidates take careful positions on issues of local concern even if these delicate tactics don't influence high-level relationships between local preferences and legislator behavior (Fenno 1978), but systematic tests of localized, particularistic position-taking are mixed. - faithfulness to local opinion on specific issues (Canes-Wrone, Minozzi, and Reveley 2011) - not much convergence even where you "most expect it" Fowler and Hall (2016)].

1.2.3 Vote shares do not identify policy ideology

The SPD's empirical support is also shaky. This project is animated by a critique that existing research has not conceptualized and operationalized constituency preferences in a way that suitably tests the theoretical ideas at play. The SPD pits two constituencies within the 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 the party's 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 applied studies, instead operationalizing the preferences of all three constituencies—the general constituency and two partisan primary constituencies—using the same single and inappropriate measure: the district-level presidential vote. The presidential vote share is not suitable for the study of primary representation for the simple reason that votes are not equivalent to policy preferences or political ideology. (I refer to "ideology" in its meaning as a summary measure of an individual's issue preferences—i.e. an "ideal point"—rather than its meaning as an abstract political value set from which policy positions can be derived.) Republican voters in a district may be ideological moderates or ideological conservatives, but the fact that they all vote Republican does not inform us on their relative ideological tastes. Similarly, a district's vote outcome captures how its constituents vote on average, but because partisans tend to vote foremost for their preferred party, this average may be more strongly affected by number of voters in each party rather than their ideological preferences. The district presidential vote fails to represent primary constituent preferences both as a concept and as a feasible proxy measure, and the U.S. representation literature's reliance on this measure for decades potentially compromises its understandings of accountability under democratized party nominations. Because we have not studied primary constituency preferences, we cannot characterize their impact in electoral representation.

Suppose a distribution of voter ideal points. There are two candidates who take positions (given real values), with a midpoint between the two of them. Formally, we can show that it is impossible to make inferences about candidate preferences using only the vote distribution.

We start with very strict assumptions about the composition of the electorate, showing that not even the structure provided by these assumptions can we make inferences about the underlying distribution of voters. Suppose we have a district containing an equal number of Republican and Democratic voters, whose preferences are distributed Normal around some unknown median, which is equal to the mean. We can take as given that the median

Republican voter is at least as conservative as the median Democratic voter. Two candidates stand to be elected president, Republican and Democrat. The same assumption is made about the ordering of candidate positions as the ordering of partisan voter medians: the Democratic candidate is no more conservative than the Republican. Voters deterministically vote for the nearest candidate to them in ideological space, which means that we can perfectly predict how a voter with a known ideal point will vote if we also know the midpoint between candidates. All voters to the "left" of the midpoint vote for the Democratic candidate, so the cumulative distribution function (CDF) of the Normal distribution

If the voter's ideal point is to the "right" of the midpoint, the voter is closer to the Republican candidate, so they vote Republican. The voter supports the Democrat if they are left of the candidate midpoint. This implies that the vote share for the Democratic candidate is equal to the cumulative distribution function (CDF) of the voter ideal point distribution evaluated at the candidate midpoint, and the vote share for the Republican is 1 minus the CDF at the candidate midpoint. Furthermore we can restrict the location and scale of the ideological distribution so that the candidate midpoint is 0.

We have assumed normality of the vote distribution, strict ordering of partisan voter medians and partisan candidate positions. What can we infer about the distribution of voters from the vote share? Very little. At minimum, the vote share only tells us the median

We can infer the median of the voter preference distribution *only* if the vote share is exactly 50-50. We cannot, based only on this information, infer the dispersion of candidate preferences around the mean, since we can stretch the underlying distribution of voter preferences without shifting the aggregate distribution left or right. As a result, a vote share of 50-50 does not tell us anything about the relative moderation or extremism about the underlying voter distribution. Figure 1.1.

If we move the vote share off of 50-50, we cannot determine either the median or the dispersion of the underlying voters.

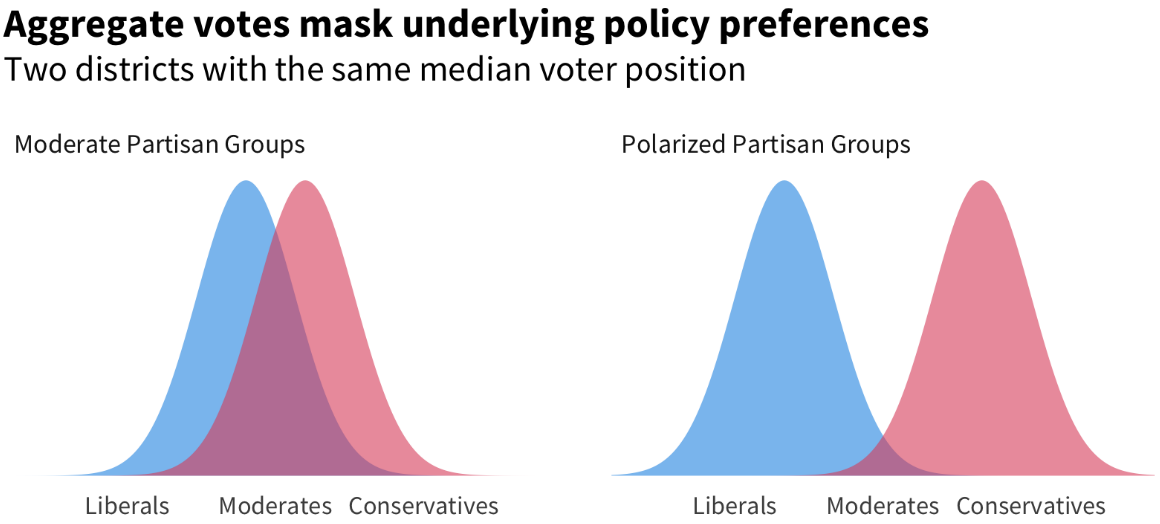


Figure 1.1: Two districts with identically located median voters

1. The vote share is the CDF of the underlying preference distribution. Assuming a symmetric distribution, we can infer the mean of the distribution from the vote share *only* if the vote share is 50-50. Even if we could identify the mean, we would not be able to infer the variance, of the distribution. If the variance of the distribution is owed to the At any other non-50-50 vote share, there are infinite possible medians, since inference about the median depends on fixing the variance.
2. If we fix the candidate positions and consider multiple districts, what does variation in vote share imply about How does variation in the vote share identify variation in Even if a simplified world, we can't do anything with this.

Insert simulations of non-identifiability…

Figure 1.2 demonstrates the problem in a simplified theoretical model. The simple model considers how constituents in a district vote for president based on their policy-ideological preferences. The two panels on the left demonstrate the basic mechanics of the scenario. 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 distribution of ideal points for each partisan base. Democratic-identifying constituents hold policy preferences that are more progressive than Republican constituents on average: the median Republican and Democrat are respectively located at 1 and -1, Each distribution has a standard deviation of 1, so there is some overlap where 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.[4] 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, and vote for the Republican candidate if they have ideal points greater than 0. The aggregate election result, therefore, is equal to the cumulative distribution function of the composite 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 1.2 show how changes in the underlying partisan distributions affect the aggregate distribution of preferences and, as a result, the presidential vote share in the district. The modifications to the partisan distributions are chosen to show the problems with using district vote share as a proxy for policy ideology in the

[4] 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 the proportion of the total distribution contributed by party p, and $\sum_p w_p = 1$. In this first example, both partisan constituencies are equally weighted at $w_p = \frac{1}{2}$. If parties had different population sizes within the same district, w_p would take values in proportion to those population sizes.

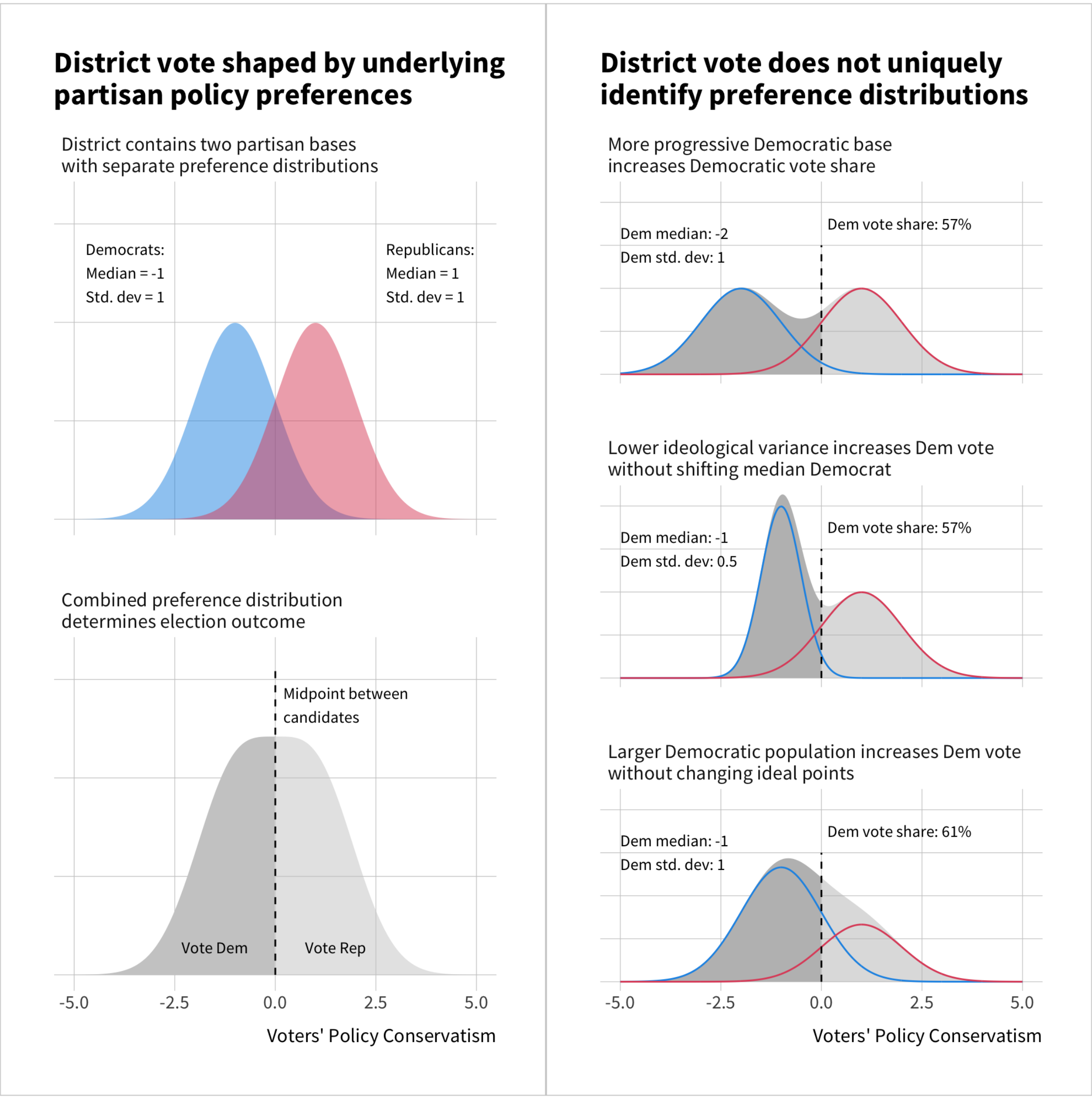


Figure 1.2: Demonstrating how district vote shares from a single election are insufficient to identify underlying policy-ideological features of the district. The left side shows how the policy preference distributions for two parties in a district (top panel) combine to form an aggregate preference distribution for the district as a whole (bottom panel).

voting population. In each panel, I intervene on only one feature of the Democratic ideal point distribution, leaving the Republican distribution untouched (median of 1, standard deviation of 1). Intervening only on one component of 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.

The top-right panel shows how the composite distribution of voters in the district changes by shifting 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 holding ideal points left of the candidate midpoint, increasing the Democratic vote share in the district from 50% to 57%. The top-left panel shows how the district vote changes by shrinking the scale of the Democratic ideal point distribution from 1 to 0.5. Lower ideal point variance within the Democratic base has the exact same effect on the vote as shifting the location, increasing 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 *more concentrated* in particular areas of ideological space (scale). The final panel in the figure shows how we can obtain different votes across distributions without changing the underlying ideological distribution in either party whatsoever, instead changing only the relative 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 panel: median of -1 and standard deviation of 1. The only difference is that the district contains an unequal balance of partisan voters, three Democratic constituents to every one Republican constituent. This results in an increased Democratic vote from 50% to 61%. To review, observing a greater than 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 in almost no cases does this mean that the median partisan ideal point has changed. Since the Republican distribution is unaffected 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, this inference would be wrong in two out of three cases: since the median Democratic ideal point is affected only by location-shifting the distribution to the left, not by lowering the Democratic ideal point variance or increasing the size of the Democratic constituency.

It is worth repeating that the scenario laid out in Figure 1.2 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.

Georgia Kernell (2009) demonstrates the difficulty of inferring district-level preferences (median voter locations) from observed votes. First, she shows that the ordering of district medians cannot be inferred from the vote shares of one election. Suppose we wish to place districts 1 and 2 on an ideological dimension. We observe that the Republican share of the two party vote in each district is p_1 and p_2. Assuming that voter preferences are normally distributed around the median voter (equivalent to the mean voter), then vote shares can be understood as the result of the normal CDF by comparing the candidate's ideal point to the distribution of voter preferences.

$$p_i = \Phi\left(\frac{c - \mu_i}{\sigma_i}\right) \tag{1.1}$$

By inverting the normal CDF, she shows that the difference in medians μ_1 and μ_2 is not proportional to vote shares in each district, but to the z-score of the vote in each district.

$$\mu_2 - \mu_1 \propto Z\left(p_2\right) - Z\left(p_1\right), \tag{1.2}$$

The nonlinear relationship between μ_i and p_i suggests that district preferences are unidentifiable without using multiple elections resulting from the same fixed set of voter preferences μ_i.

Following a similar setup, I demonstrate that the task of inferring the positions of multiple parties is even more difficult. First, rather than assuming that voter preferences in a district are normally distributed with one mean and one dispersion term, we assume that voter preferences are mixture of two distributions (indexed 1 and 2). Each party votes for a Republican candidate akin to the normal CDF as before, but the vote share p for the Republican reflects the size of each partisan constituency in the district as well. Suppose that the size and ideal point of party unaffiliated is represented by a normally distributed error term ε:

$$p = \Phi\left(\frac{c - \mu_1}{\sigma_1}\right)\pi_1 + \Phi\left(\frac{c - \mu_2}{\sigma_2}\right)\pi_2 + \varepsilon, \tag{1.3}$$

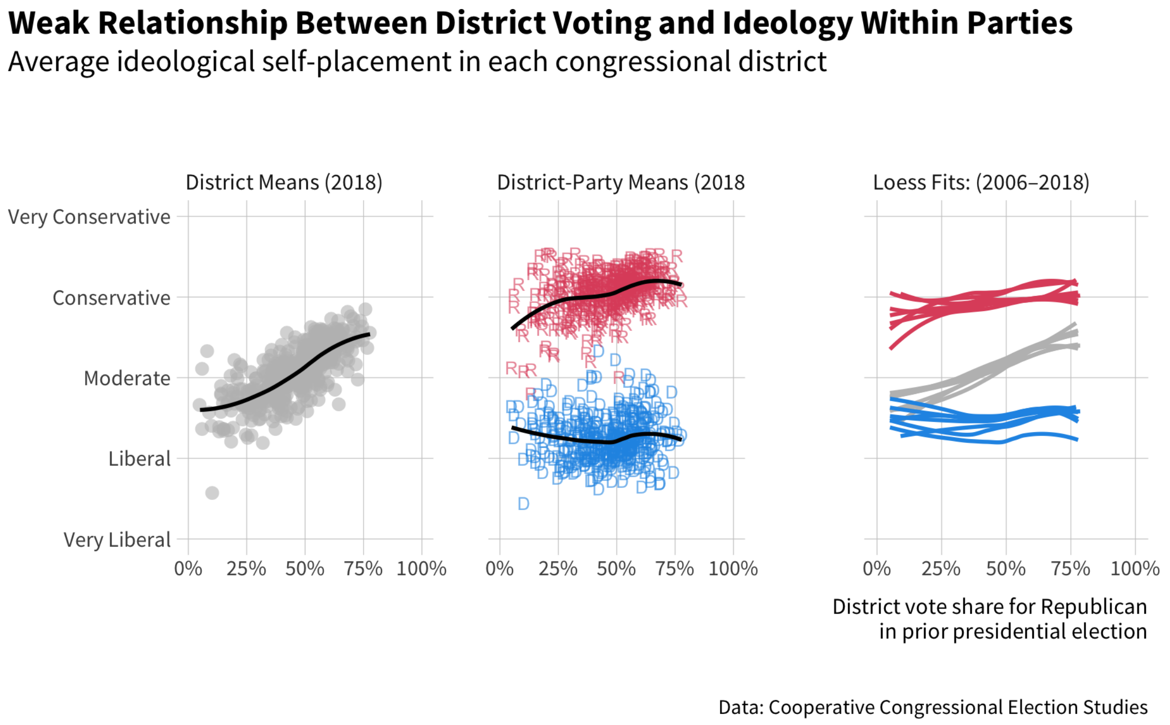
where p_i reflects the proportion of the electorate that identifies with party i. Isolating μ_1 and μ_2 in this setup is rather difficult, and the choice of simplifying assumptions is limited. We could assume that the variance of preferences in each party is equal, $\sigma_1 = \sigma_2 = \sigma$, and manipulate the equation somewhat…

$$Z\left(p\right)\sigma = \left(c - \mu_1\right)\pi_1 + \left(c - \mu_2\right)\pi_2, \tag{1.4}$$

but the limitations are still significant. We could not make a simplifying assumption that $\pi_1 = \pi_2$ for but a handful of districts. We might estimate π_i from survey data

The conceptual difference between district vote shares and aggregate ideology appears in real data as well, as shown in Figure 1.3. 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 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 conservative, indicated by a positively sloped loess fit line. The middle pattern 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 1.3 is that we should doubt the assumption that aggregate voting in the 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 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 to proxy within-party ideology may simply be measuring the *size* of a partisan group in a district instead of its ideological makeup.



Weak Relationship Between District Voting and Ideology Within Parties

Average ideological self-placement in each congressional district

District Means (2018)

District-Party Means (2018

Loess Fits: (2006–2018)

Data: Cooperative Congressional Election Studies

Figure 1.3: 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.

It must be stressed that these obstacles are not merely methodological, since the theoretical consequences are enormous. The literature's dependence on the presidential vote as a proxy for district preferences and its omission of primary candidate ideology have prevented scholars from testing basic theoretical propositions in the study of primary politics. Put simply, without serviceable measures of local partisan preferences or primary candidate positions, 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 nominations as well. To study how politicians weigh the opinions of their party's base of support against other voters in their districts, we must be able to measure the preferences of a politician's local partisan constituency.[5] As scholars explore whose voices truly matter in shaping party platforms and policy (Bartels 2009; Cohen et al. 2009; Gilens 2012; Grossman and Hopkins 2016; Lax and Phillips 2012), these researchers will benefit enormously from an approach to estimate the subnational preferences of partisan voters. This dissertation will demonstrate the added value of conceptualizing voter preferences in this way.[6]

I fill this gap in the representation literature by developing a new measurement model for local partisan preferences and applying these estimates to important questions of primary representation. The model builds on the work of Tausanovitch and Warshaw (2013) and Caughey and Warshaw (2015), estimating a latent index of policy liberalism within partisan groups in Congressional districts. For each district, the model estimates the mean ideal point of Democratic and Republican constituents using survey data. With these new estimates, I directly test the hypotheses that other studies have fallen short of testing. Is it truly the case that the primary constituency is a standout cause of candidate positioning? Do primary voters nominate partisan candidates for their ideological "fit"? And do institutional configurations such as gerrymandering distort these relationships in ways commonly suggested by scholars and political observers alike?

This remainder of this chapter reviews leading theoretical approaches to primary representation, highlights the importance of partisan constituent preferences, and uses formal models to demonstrate how existing approaches fail to measure the important theoretical constructs.

[5] Clinton (2006) provides suggestive evidence on the importance of this distinction. Using a cross-sectional survey, he finds that the roll-call voting behavior of Democratic members of Congress is more responsive to the district's median constituent, while Republican members respond more strongly to the average Republican constituent.

[6] New work by Caughey, Dunham, and Warshaw (2018) operationalizes the average ideal points of partisans within states. I offer important extensions on their modeling approach and apply the estimates to very different substantive questions.

The chapter ends by previewing the remaining chapters of the dissertation about the new measurement model (Chapter 2) and analyses that apply the new estimates to important areas of primary representation: the strategic positioning of primary candidates (4) and the vote choices of primary constituents (5).

1.2.4 How have past studies incorporated voter preferences?

District vote

- A.S.S. district vote?
- Burden district vote?

Some weird fuckin residualization thing

Subconstituency preferences

- Fenno 1978 (Homestyle) for why this matters
- Arnold 1990
- Kingdon 1973
- scholars fail to advance a general theory, which makes models difficult to specify? (Poole and Romer 1993) cited in Bishin 2000
- constituency matters if you drop member ideal points in here either: (Markus 1974; Shapiro et al. 1990; Wright 1989); doesn't matter if you include member ideal points (Medoff, Dennis, and Bishin 1995); cited in Bishin 2000
- Bishin 2000 is a good example, but measurement isn't great.
- Clinton: within-party

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

Substantive value of the model.

Chapter 4 studies how district-party ideology affects candidate positioning in primary elections.

Chapter 5 studies how district-party ideology affects candidate selection in primary elections.

Orientation: Which games are we playing:

- let's test the SPD
- we do not root for or against the SPD.
  - the SPD makes sense for candidate behavior, but there are also reasons to doubt the competences of candidates.
  - the SPD is hard for voters, but it's also possible that voters "accidentally" choose the correct candidate because of spurious or confounding factors such as fundraising or elite network machinations.
- "Spatial" Models of Policy Preferences
- Representatives as Delegates vs. Trustees

1.3.3 Causal inference with structural models

What is the problem?

- Ignorable assignment of aggregate ideology
- Ignorable assignment of candidate ideology

Using SCMs.

DAGs as heuristic devices for causal inference are not new to political science in general (Gerring…), but the formalism of causal estimands and the strict derivation of causal assumptions from structural/graphical models is new.

This is new to this LIT as well, which has not been great about explicit causal assumptions and empirical designs, with some notable exceptions.

- hall 1 and 2
- who else….literally idk.
- hall diff-in-diff stuff?
- hall convergence

We actually don't have much of this here.

- Few examples.

What do I propose?

- Many observational studies using "selection on observables" assumptions.
- We may not be able to break out of this
- But we can improve on the (1) declaration of the causal query, (2) examination of which assumptions are required for which queries, and (3) be more cautious about what we can and can't say using the data and research designs that we have.
- Pragmatic orientation: Taboo against explicit causal inference, (Grosz, Rohrer, Thoemmes), Whatever the Hernán (2018) article says
- The point of causal inference isn't to shame observational studies away from making causal claims. The point is to clarify which designs and assumptions enable certain causal claims and which do not.

Machine learning

Specifically:

- modeling approach for controlling mediators: sequential-g (ch 4)
- observation causal inference with machine learning models for covariate adjustment: neural network with neyman orthogonalization (ch 5)

1.3.4 Bayesian causal modeling

Where do we have to confront things

- the treatment assignments themselves are imprecisely measured. We have a probability distribution of *observed* treatment.

- We view NHST with skepticism
- We aim to normalize a Bayesian interpretation of causal effects
  - there are many possible scenarios that could explain the data that we see
  - the posterior distribution represents the plausibility of scenarios, after seeing the data.

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 (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).

— 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.

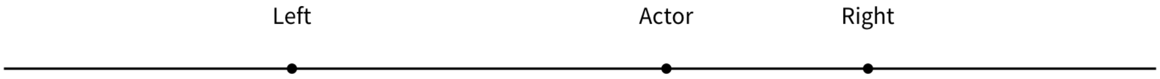


Figure 2.1: An Actor and two policy outcomes (Left and Right) represented as locations in ideological/policy space

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 $i \in \{1, \dots, n\}$ making an arbitrary number of policy choices (bills, survey items, etc.) indexed $j \in \{1, \dots, J\}$. 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.

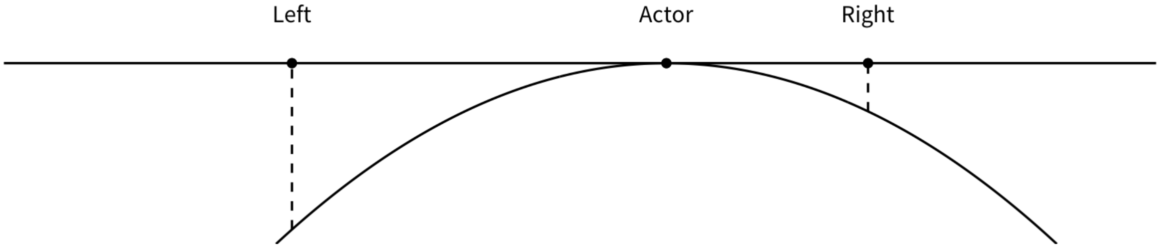


Figure 2.2: A representation of quadratic utility loss over policy choices

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 $\mathrm{E}\left(u_{ij}^{\mathsf{L}}\right) = \mathrm{E}\left(u_{ij}^{\mathsf{R}}\right)$ and that $\mathrm{Var}\left(u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right) = 1$. 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 $\mathrm{Var}\left(u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right) = \frac{s_j^2\pi^2}{3}$, 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]

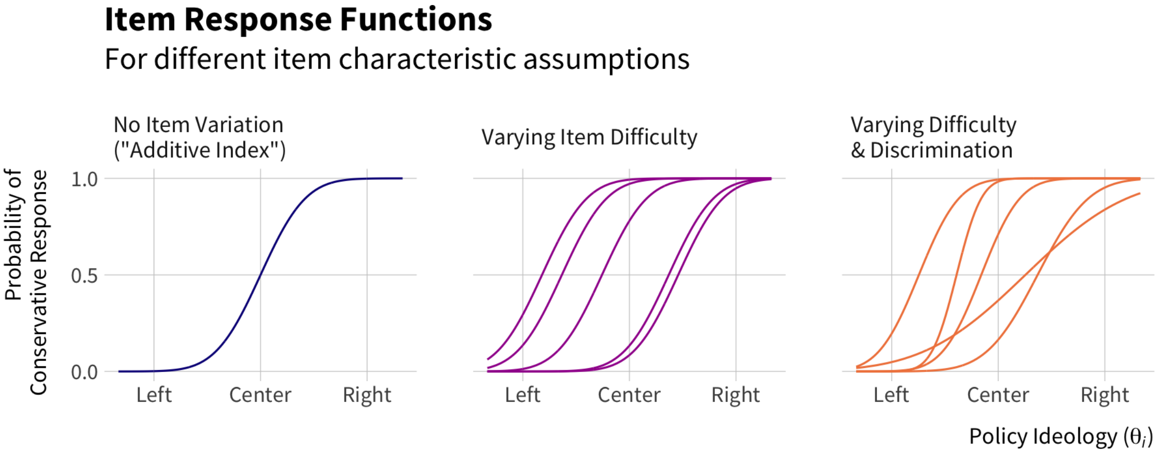


Figure 2.3: Examples of item characteristic curves under different item parameter assumptions

[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: $\dfrac{\partial \iota_j \left(\theta_i - \kappa_j\right)}{\partial \kappa_j} = -\iota_j$, 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 $\dfrac{\partial \iota_j \left(\theta_i - \kappa_j\right)}{\partial \iota_j} = \left(\theta_i - \kappa_j\right)$. 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

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

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 $\alpha_{sp}^{\mathtt{state}}$ 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 by Caughey and Warshaw (2015), which estimates the same set of geographic effects for all

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

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

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

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 sd(·) and mean(·) 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

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

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 $\lfloor s^*_{gj} \rfloor$ 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.

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

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

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.

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

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,

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

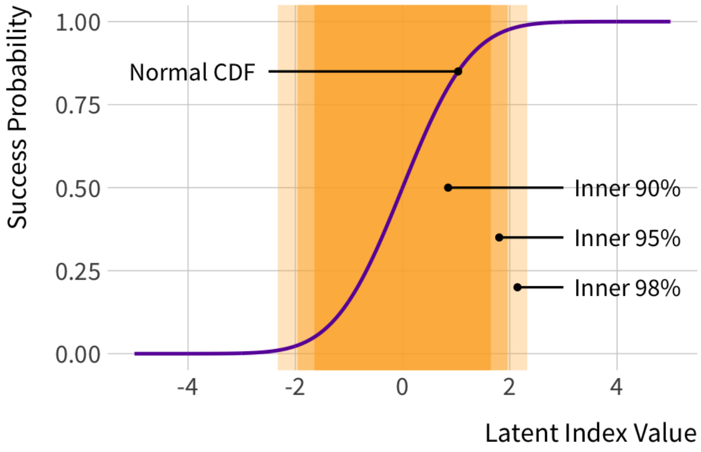


Figure 2.4: The region of the probit model's latent index that maps to response probabilities between 1 and 99 percent.

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

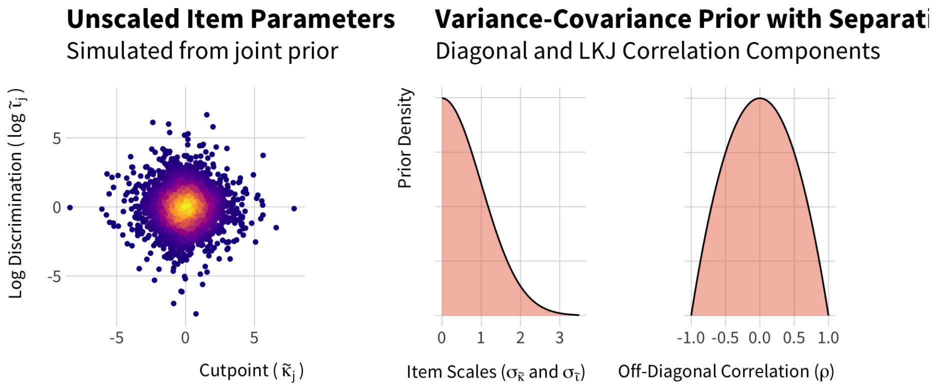


Figure 2.5: Components of the joint hierarchical prior for the unscaled item parameters. Left panel shows prior values for unscaled item parameters from the joint prior. Remaining panels show priors for decomposed covariance matrix components: including the standard deviation that form the matrix diagonal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

all covariates are at their means. I then give this grand mean a $\text{Normal}\left(0,1\right)$ 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 $\text{Normal}\left(0,0.5\right)$ 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 $\text{Normal}\left(0,1\right)$ 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 $\text{LogNormal}\left(0, 1\right)$ 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

2.8 Testing the Model with Simulated Data

2.9 Data Sources

Describe data

2.10 Model Results

The model was estimated using a remote server at the University of Wisconsin–Madison.[22] I generated posterior samples using MCMC on 5 Markov chains. Each chain was run for 2,000 iterations, divided into 1,000 warmup iterations to tune Stan's adaptive HMC algorithm and 1,000 post-warmup iterations saved for analysis.[23] Following the advice of Link and Eaton (2011), I stored every post-warmup sample with no thinning of chains, resulting in a total of 5,000 samples per parameter across all chains.[24]

Here we can reference Figure 2.6.

[22] A Linux server ("Linstat") maintained by Social Science Computing Cooperative.

[23] The algorithm was initialized with an `adalt_delta` parameter of 0.9 and a `max_treedepth` of 15.

[24] The chains mix well and exhibit little autocorrelation, which is owed to the fact that Hamiltonian Monte Carlo algorithms are much more efficient at proposing transitions and thus exploring a parameter space.

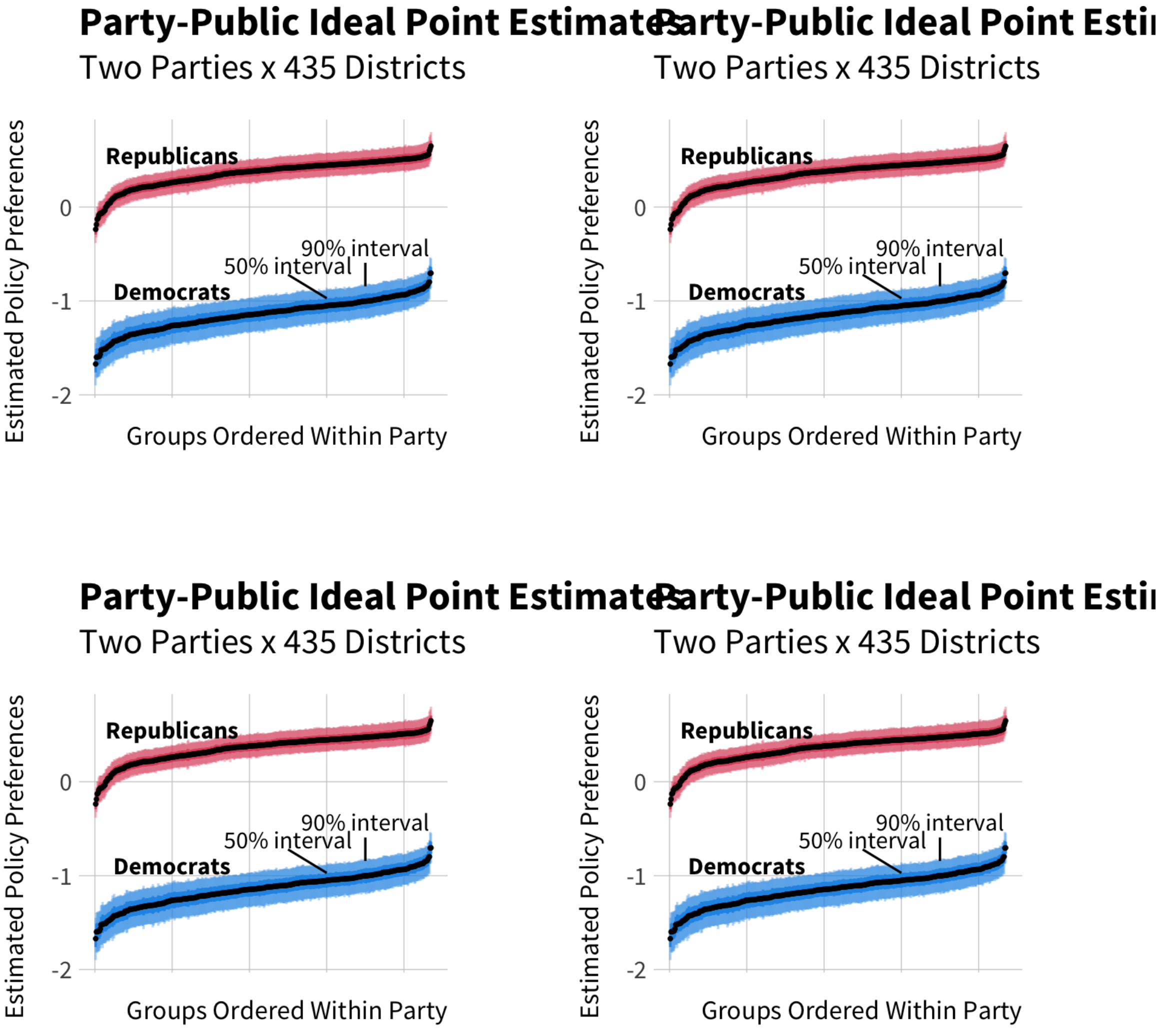


Figure 2.6: Posteriors

— 3 —

Bayesian Causal Models for Political Science

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

First, this project views causal inference as a problem of posterior predictive inference. Causal models are tools that enable inferences about missing data: the data that would be observed if a key independent variable could be set to a different value. The unobserved data are "unobserved potential outcomes" in the Rubin causal framework or "counterfactual observations" in the Pearl framework. Regardless of the notational/semantic conventions employed, "Bayesian causal inference" is a modeling framework for structuring the inferences about the probability distribution of unobserved potential outcomes, having observed one set of potential outcomes. In other words, which counterfactual data are plausible, given the observed data and a model relating the data to causal queries of interest? In this sense, the Bayesian approach to causal inference confronts our uncertainties about the "left-hand side" of our model.

Second, the Bayesian approach confronts our uncertainties about the "right-hand side" of the model as well. Causal estimands (to use the Rubin terminology) are comparisons of potential outcomes are two hypothetical values of a treatment. In many cases, these estimands are comparisons of a unit's outcome under an observed treatment against an unobserved treatment. For this project, the treatment of interest—ideology in district-party-publics—is not directly observed. It is instead only observed up to a probability distribution from a measurement model. (To use potential outcomes notation, instead of observing $Y(\pi)$, we observe $Y(p(\pi))$). Bayesian probabilistic models provide machinery to describe causal queries when we have only a probabilistic understanding of the observed data as well as the unobserved data.

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 the observed data. Because Bayesian modeling remains largely foreign to political science, I confront several issues surrounding prior distributions for causal models. I discuss how priors are inescapable for many causal claims, how priors can provide important structure to improve causal inferences, and practical advice for constructing and evaluating priors. Finally, I provide examples of the Bayesian causal framework at work, highlighting how priors add value to causal inference at different levels of abstraction.

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. The rigorous accounting of the causal model and identifying assumptions distinguishes causal inference from informal, verbal discussions of causal effects.

The current movement in causal inference spans several fields, most notably in statistics, biostatistics, epidemiology, computer science, and economics. The dominant modeling approach to causal inference in political science is rooted in a notation of *potential outcomes* (Rubin 1974, 2005). This "Rubin model" formalizes the concept of a causal effect by first defining a space of hypothetical 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] For 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. If $Y_i(A = 1)$ differs from $Y_i(A = 0)$, then the treatment has a nonzero causal effect on unit i, denoted τ_i.

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

Defining the causal model in terms of unit-level effects provides an exact, minimally sufficient definition of a causal effect: A affects Y treatment has a nonzero effect for any unit.

By establishing this baseline model, the researcher can derive exact theoretical statements about causal effects by manipulating the equations that define the model. A causal model may describe more complex causal effects, such as whether Y_i it is observed at all, whether the treatment effect is indirectly mediated by another variable, whether the effect depends on other baseline characterics of units, and so on.

[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.

The entire space of potential outcomes is a hypothetical device. Although a causal model defines potential outcomes for every unit under every treatment assignment, it is not possible to observe all of these potential outcomes, since a unit can receive only one treatment, and thus can only take a single outcome value. This implies that the individual causal effect (τ_i), while a valid feature of the hypothetical causal model, is never actually observed be for any unit. This "fundamental problem of causal inference" is the core philosophical problem in causal inference; because the researcher never observes a unit under more than one treatment status, she can never make any causal claims based on the observed data alone (Holland 1986). Causal claims are possible only by imposing assumptions on the data that allow the researcher to predict what the data would look like if units received treatments that they did not actually receive. These "identification assumptions" specify the conditions under which data from one treatment group can be used to make generalizations about data from other treatment groups (Keele 2015).

A related feature of causal models is that the causal effect τ_i is defined at the unit level, so it can be different for each unit. This low-level causal heterogeneity is often an important feature of research designs, since the identifiable causal quantities will depend on which assumptions the researcher is willing to make about the structure of that heterogeneity. The body of work that we call "causal inference" focuses on (1) which the causal quantities are described by a causal model, (2) the assumptions required to identify those quantities in observed data and the research designs that invoke them, and (3) estimation methods for those causal quantities.

A separate but increasingly popular approach to causal modeling focuses on *structural causal models* or "SCMs" (Pearl 1995, 2009) that describe networks of causally related variables. Some treatment A causally affects the outcome Y if the treatment is a component of the function that assigns values to the outcome: $Y_i := f(A_i, \ldots)$. The notion of *assignment* is important to distinguish from *equality*, since it distinguishes a directional causal effect from a bidirectional correlation. This is a central feature of the SCM approach, which is premised on the difference between *observing* relationships as they appear in the world and *obtaining* relationships by intervening on the world. We can observe a relationship between outcome Y and treatment A by observing that the distribution of Y given A, $p(Y \mid A = a)$, changes as A takes different values, but we would observe a different relationship by intervening on the world to set the value of A ourselves.

$$\begin{aligned} &\text{Observational Distribution: } p(Y \mid A = a) \\ &\text{Interventional Distribution: } p(Y \mid do\,(A = a)) \end{aligned} \tag{3.2}$$

Like the potential outcomes model, A has a causal effect on Y if it changes at least one unit's Y value, since it only takes one unit to change the probability distribution $p(Y \mid do\,(A = a))$. The SCM framework comes to life by implementing *do*-calculus (Pearl 1995), rules for conditioning on variables in the causal system to nonparametrically identify causal effects. The SCM approach is important for this dissertation because its use of probability distributions to convey causal effects lends itself to Bayesian modeling more readily than the potential outcomes notation.

The most recognizable feature of the SCM framework is that every structural model can be visualized as a graphical model, where nodes represent variables and directed edges represent causal relationships. Many researchers refer to these graphs as "directed acyclic graphs," but I refer to them simply as "causal graphs."[2] These graphs are valuable because they reveal which variables must be adjusted for in order to satisfy certain identification criteria in *do*-calculus. Most notably, the *back-door criterion* states that a causal effect is identified by

[2] This is an effort to reduce jargon. DAGs can be used to represent many systems that have nothing to do with causal inference, since "directed" and "acyclic" refer only to the mathematical properties of the graph. Formally, any graph is a DAG if two nodes can only be first-order connected by just one directed edge, and no path through the graph begins and ends with the same node. This entails that nodes cannot be immediately connected by multiple edges, undirected edges, or bidirectional edges. Furthermore, the graph contains no "loops": edges that connect a node to itself.

"blocking any back-door paths" between treatment A and outcome Y—paths that connect A to Y that begin with an arrow flowing into A and that pass through other variables. For instance, Figure 3.1 shows a causal system where the effect of A on Y is not identified because we can connect A to Y through the back-door path $A \rightarrow U \rightarrow Y$. This means that U confounds the relationship between A and Y, but we can identify the relationship by conditioning on U and "blocking" the back-door path. Graphs are valuable for this project because I use them to convey important identification assumptions in later chapters.

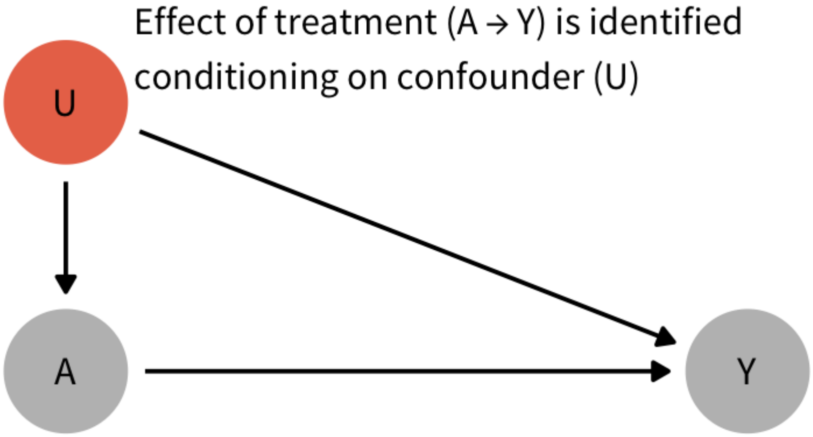

Effect of treatment (A → Y) is identified
conditioning on confounder (U)
U
A
Y

Figure 3.1: A causal graph where the effect of A on Y is confounded by U.

As with the potential outcomes framework, *do*-calculus (and causal graphs in turn) describe minimally sufficient conditions for *nonparametric* causal identification. There is no guarantee that linear regression models, or any parametric models, adequately control for confounders and isolate identifying variation in the treatment. For this reason, it can be helpful to lay out a hierarchy of modeling concerns for any causal inference problem.

1. Causal modeling: the definition of potential outcomes or structural equations that define the space of causal effects, and the declaration of relevant causal estimands.
2. Identification assumptions: the assumptions required to identify causal estimands using only the observed data. These assumptions are usually posed as statements about

expectations of Y and when we can ignore the fact that counterfactuals are unobserved. For instance, conditioning on a confounder U identifies the conditional causal effect of A because it satisfies an assumption that A is independent of the potential outcomes after conditioning on U.

$$Y_i(A = 1),\ Y_i(A = 0) \perp\!\!\!\perp A_i \mid U_i = u \tag{3.3}$$

Conditional independence between treatment and potential outcomes lets the researcher make inferences about potential outcomes, which we can't always observe because they are hypothetical entities in the causal model, having only observed the real-world data. For instance, if potential outcomes are conditionally independent of treatment status A, then we can conclude that the Y actually observed at $A = a$ is equal to the Y that we *would have observed* if we could set $A = a$ ourselves, within strata $U = u$.

$$\begin{aligned} \mathbb{E}\left[Y_i \mid A_i = a, U_i = u\right] &= \mathbb{E}\left[Y_i(A = a) \mid A_i = a, U_i = u\right] \\ &= \mathbb{E}\left[Y_i(A = a)\right] \mid U_i = u \end{aligned} \tag{3.4}$$

This conditional independence assumption (CIA) is the same assumption invoked by the back-door criterion. By conditioning on U, we subtract any variation in A that can be explained by U, so the remaining relationship between A and Y represents the relationship we would observe by setting $do\left(A = a\right)$ within stratum $U = u$.

3. Statistical assumptions: What estimators are we using to estimate $\mathbb{E}\left[Y_i \mid A_i, U_i\right]$, and do they introduce additional assumptions that interfere with causal inference? Do they suffer additional biases due to misspecification, functional form assumptions, that jeopardize key identification assumptions?

I lay out this hierarchy for two reasons. First, it clarifies why researchers use certain research designs or statistical approaches to overcome certain problems with their data.

Statistical assumptions, we will see, can undermine identification assumptions, 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). If researchers cannot design away these difficult assumptions, other methods are available to adjust for confounders without as many strict assumptions about the functional form of the causal model as are commonly invoked in parametric regression models. Causal inference is not synonymous with the new "agnostic statistics" (Aronow and Miller 2019), but it is animated by a similar motivation to implement statistical methods that relies on fewer fragile assumptions. For causal inference problems, these methods include matching, propensity score weighting, and even machine learning methods for estimating flexible conditional expectation functions that are robust to nonlinearities and non-additivities in the data generating process (Aronow and Samii 2016; Green and Kern 2012; Hill 2011; Samii, Paler, and Daly 2016; Sekhon 2009). This dissertation will use machine learning methods, in particular Bayesian neural networks (BNNs), to estimate flexible expectation functions without relying on strict assumptions about unknown functional forms.

Second, this three-party hierarchy of causal clarifies where my contributions around Bayesian causal estimation will be focused. As I discuss below, a Bayesian view changes the interpretation of the causal model (level 1) by invoking a *probability distribution* over the space of potential outcomes. This probability distribution allows the researcher to say which counterfactual data are more plausible than others, which is a desirable property in a causal modell that conventional inference methods nonetheless do not readily enable. A Bayesian approach to causal inference also has the potential to extend the meaning of identification assumptions (level 2) by construing them also as probabilistic features of a model rather than fixed features. Finally, and most practically, the Bayesian framework offers a greater variety of statistical tools for estimating causal effects (level 3).

3.1.2 Bayesian Inference

Bayesian reasoning is a contentious and misunderstood topic in empirical political science, so it is important to establish some essential tenets to the approach before melding it with causal modeling. Bayesian analysis is the application of conditional probability for statistical inference. Its mechanical underpinnings are uncontroversial, essential building blocks of probability theory: how the probability of an event changes by conditioning on other known information. The controversy surrounding Bayesian methods in political science is better understood as a disagreement over which modeling constructs we choose to describe using probabilities.

Whereas many statistical methods begin with a model of data given fixed parameters, Bayesian inference is premised on a joint modeling for all components in a system. For example, suppose that we are interested in jointly describing age and vote choice in a population: We could write the joint distribution of these two variables as $p(Age, Vote)$. This distribution can be equivalently expressed by factoring it in two ways:

$$p(Vote \mid Age)p(Age) = p(Age \mid Vote)p(Vote) \tag{3.5}$$

Supposing that we observed an individual's vote choice, how would this information affect our information about their likely age? In order to condition on known information, probability theory says that we divide the joint probability by the probability of the conditioning event.

$$\begin{aligned} \frac{p(Vote \mid Age)p(Age)}{p(Vote)} &= \frac{p(Age \mid Vote)p(Vote)}{p(Vote)} \\ \frac{p(Vote \mid Age)}{p(Age)} &= p(Age \mid Vote) \end{aligned} \tag{3.6}$$

This simple operation reveals a statement that is conventionally understood as Bayes' theorem: the conditional probability of A given B, expressed in terms of B given A. The machinery underpinning Bayesian inference is nothing but a by-the-book application of probability theory. It provides, through no magic whatsoever, a formal method for rationally updating a joint probability distribution by conditioning on known information.

This machinery can be applied to a more general modeling scenario containing data $\mathbf{y}$ and parameters $\boldsymbol{\pi}$. The joint model for the data and the parameters takes the form

$$p(\mathbf{y}, \boldsymbol{\pi}) = p(\mathbf{y} \cap \boldsymbol{\pi}) = p(\mathbf{y} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}), \tag{3.7}$$

where $p(\mathbf{y} \mid \boldsymbol{\pi})$ represents the probability distribution of the data, conditioning on parameters, and $p(\boldsymbol{\pi})$ represents a *prior distribution* for parameters, marginal of (or unconditioned on) the data. The prior distribution is commonly described as the researcher's "prior information" (or, controversially, prior "beliefs") about the parameter, an important but misunderstood component of Bayesian modeling.

The joint model provides machinery for learning about parameters by conditioning the parameters on the data,

$$p(\boldsymbol{\pi} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi})}{p(\mathbf{y})} \tag{3.8}$$

also known as obtaining the *posterior distribution* of the parameters.

These basic concepts of conditional probability and Bayesian updating are not foreign to political science, but it will be important to establish an interpretation of Bayesian thinking and Bayesian modeling that is productive for causal inference. This project invokes what McElreath (2017b) calls an "inside view" of Bayesian statistics—Bayesian statistics on its own terms. This implies a rejection of an "outside view" that construes Bayesian statistics as a penalized variation of maximum likelihood. Under the outside view, heavily influenced by the prominence of non-Bayesian likelihood approaches to statistical inference, data follow probability distributions, while parameters are fixed constants. Bayesian estimation uses the original likelihood model and attaches a penalty in the form of a prior distribution for model parameters. Prior distributions, in turn, represent the subjective beliefs of the researcher. Bayesian modeling can be reduced to a biased estimation procedure that downweights data in favor of arbitrary subjectivity. This view of Bayesian statistics is admittedly confusing, and if we take it at face value, it is no wonder that causal inference in political science has largely avoided Bayesian tools other than for computational convenience.[3].

The inside view, as mentioned above, construes a Bayesian model as a joint model for all variables in a system. The use of the word "variables" to encompass both data and parameters is crucial. The technology of a Bayesian model does not regard data and parameters as distinct from one another. They follow the same rules, just as age and vote choice followed the same rules in the above example. Data and parameters are both instantiations of uncertain processes, with the only semantic difference between the two being that observed variables are called "data" while unobserved variables are called "parameters" (McElreath 2017b).[4] Prior distributions and likelihood functions are the same thing: probability distributions that quantify uncertainty about a variable. If I were to observe a new data point from a model, I would be unable to predict its value exactly, but some values are more probable than others, given the parameters that condition the data. The same premise holds for parameters: if I could observe a parameter, I would have been unable to anticipate it exactly, but I could bet that some parameters are more likely than others, given the data that I have already seen. The joint model for all variables encapsulates the probabilistic relationships between data and parameters. Starting with the prior model, $p(\mathbf{y}, \boldsymbol{\pi})$, we can condition the model on chosen parameters to obtain a rationalizable distribution of data, $p(\mathbf{y} \mid \boldsymbol{\pi})$. Or we can condition the

[3] I return to the distinction between "Bayes for the sake of MCMC" and "Bayes for the sake of Bayes" in a later discussion of prior distributions.

[4] The semantic conventions are often sloppier in practice than many researchers would like to think. Many analyses use data that summarize lower-level processes, such as per-capita income in U.S or the percentage of women who vote for the Democratic presidential candidate, which behave like random variables in that their values could differ under repeated sampling. The semantic distinction between data and parameters 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.

model on data to obtain a rationalizable distribution of parameters, $p(\boldsymbol{\pi} \mid \mathbf{y})$.[5]

From the inside view, Bayesian updating proceeds by considering a variety of possible scenarios that create data and evaluating which scenarios are consistent with the data. The joint prior model, $p(\mathbf{y} \mid \boldsymbol{\pi})p(\boldsymbol{\pi})$ describes an overly broad set of possible configurations of the world. This configurations are the combination 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 in light of data actually observed. The plausibility, or posterior probability, of a parameter value is greater if the observed data are more likely to occur under that parameter value versus another value: larger $p(\mathbf{y} \mid \boldsymbol{\pi})$. Conditioning on the data downweights the configurations of the world that are inconsistent with the observed data, resulting in a distribution of scenarios that reflect more plausible configurations of the world (McElreath 2020, chap. 2). This is an important distinction from non-Bayesian statistical inference, since there can be no formal notion of "plausible parameters" without a posterior distribution, which necessitates a prior distribution. For causal inference, this means there can be no formal notion of "plausible counterfactuals" without a probability distribution over the counterfactuals, which necessitates 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.

The inside view of Bayesian modeling, and the philosophical unity that it brings to statistical machinery and inference, is possible even if using uninformative prior distributions that are indifferent to possible parameters *a priori*. This is how Bayesian methods tend to appear in political science to date, with noninformative priors that exist primarily to facilitate Bayesian computation for difficult estimation problems. The infamy of Bayesian methods,

[5] McElreath (2017a) calls these maneuvers "running the model forward" (data given parameters) or "running the model backward" (parameters given data). It can be helpful to think of the model as a machine with many interlocking parts, where if you crank the machine, the pieces that you choose to constrain will affect which pieces do and do not move.

however, is owed to the ability of the researcher to specify "informative" priors that concentrate probability density on model configurations that are thought to be more plausible even before data are analyzed. There are many modeling scenarios where this concentration of probability delivers results that are almost unthinkable without prior structure: multilevel models that allocate variance to different layers of hierarchy, highly parameterized models with correlated parameters such as spline regression, and sparse regressions where regularizing priors are used to shrink coefficients and preserve degrees of freedom to overcome the "curse of dimensionality" (e.g. Gelman et al. 2013). At the same time, many researchers are skeptical of Bayesian methods because supplying a model with non-data information can be spun as data falsification (García-Pérez 2019). As I elaborate in Section 3.4, the mere non-flatness of a prior is often less consequential than the prior's *structure*. For instance, whether a prior has variance of 10 or 100 often matters less than whether the prior specifies a hierarchical structure for borrowing information from other groups in the data. It is no coincidence the Bayesian contributions in this dissertation and in other areas of applied statistics are concerned with different prior structures more than the precision of otherwise equal prior structures.

3.1.3 Shared Goals, Different Tactics

Fux with GGK:

- measure = truth + bias + variance
- causal inference:
    - reduce bias "by design"
    - shrink prior on bias
- bayesian estimation:
    - shrink prior on variance using prior info

  - or shrink prior on measure, where truth = measure - bias - variance

Not *inherently* the same as "agnostic" inference though we fux with both.

3.2 Bayesian Causal Modeling

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 updating our information about counterfactual potential outcomes in light of the observed data. The headings under 3.2 introduce this modeling framework at a high level. I provide a probabilistic interpretation and notation for potential outcomes models (3.2.1), a connection between causal parameters and model parameters (**??**), and a broad justification for the Bayesian interpretation of causal effects (3.2.3).

3.2.1 Probabilistic Model for Potential Outcomes

As with other causal models, we begin at the unit level. Unit i receives a treatment $A_i = a$, with potential outcomes $Y_i\left(A_i = a\right)$. Suppose a binary treatment case where A_i can take values 0 or 1, so the unit-level causal effect is $\tau_i - Y_i\left(1\right) - Y_i\left(0\right)$.

The unit-level causal effect τ_i is unidentified, but it is possible to estimate to estimate population-level causal quantities by invoking identification assumptions. For instance, the conditional average treatment effect at $U = u$, $\bar{\tau}(U = u) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid U_i = u\right]$, can be estimated from observed data assuming consistency, non-interference, conditional ignorability, and positivity,

$$\begin{aligned}\bar{\tau}(U = u) &= \mathbb{E}\left[Y(A = 1) - Y(A = 0) \mid U = u\right] \\ &= \mathbb{E}\left[Y(A = 1) \mid U = u\right] - \mathbb{E}\left[Y(A = 0) \mid U = u\right] \\ &= \mathbb{E}\left[Y \mid A = 1, U = u\right] - \mathbb{E}\left[Y \mid A = 0, U = u\right]\end{aligned} \tag{3.9}$$

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 stated in terms of potential outcomes, which makes them unobservable. Observable estimands are the equivalent quantities that could be recovered from observable data under the identification assumptions. Other literature refers to observable estimands as "nonparametric estimators" (Keele 2015), but I steer clear of this language because the gap between observable estimands and estimators is an essential distinction for understanding the Bayesian causal approach.

The transition to a Bayesian probabilistic model begins with an acknowledgment that no estimator for the observable estimand, $\mathbb{E}\left[Y \mid A = a, U = u\right]$, is exact. The assumptions identify causal effects only in an infinite data regime, where the observable estimand can be 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 $f(Y \mid A, U, \boldsymbol{\pi})$ be a prediction function for the mean of Y that depends on treatment A, confounders U, and parameters $\boldsymbol{\pi}$. This model's expectation is assumed to be equal to the observable estimand.

$$\mathbb{E}\left[Y \mid A = a, U = u\right] = f(Y \mid A = a, U = u, \boldsymbol{\pi}) \tag{3.10}$$

This assumption 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).[6] The conditional average treatment effect $\bar{\tau}(U = u)$ is obtained by differencing these model predictions over the treatment.

$$\begin{aligned}\bar{\tau}(U = u) &= \mathbb{E}\left[Y \mid A = 1, U = u\right] - \mathbb{E}\left[Y \mid A = 0, U = u\right] \\ &= f(Y, A = 1, U = u, \boldsymbol{\pi}) - f(Y, A = 0, U = u, \boldsymbol{\pi})\end{aligned} \tag{3.11}$$

These expectations are infinite-data results, but in any finite sample, our expectations of Y are uncertain because we estimate $\boldsymbol{\pi}$ with uncertainty. The Bayesian framework provides a natural ability to incorporate uncertainty into the causal estimate by characterizing our probabilistic information about the treatment effect. Supposing that we begin with a joint model for data and parameters: $p\left(Y, \boldsymbol{\pi}\right) = p\left(Y \mid f(Y, A, U, \boldsymbol{\pi}), \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi}\right)$. The data are distributed conditional on their expected value, modeled as $f(Y, A, U, \boldsymbol{\pi})$, as well as the parameters $\boldsymbol{\pi}$. The parameters also have a prior distribution $p\left(\boldsymbol{\pi}\right)$, or a distribution marginal of the data. This model is sufficient to characterize the probability distribution for the conditional average treatment effect as defined in Equation (3.11),

$$p(\bar{\tau}(U = u)) = \int p\left[f(Y, A = 1, U = u, \boldsymbol{\pi}) - f(Y, A = 0, U = u, \boldsymbol{\pi})\right] p\left(\boldsymbol{\pi}\right) \,\mathrm{d}\boldsymbol{\pi} \tag{3.12}$$

which is the probability distribution over differences in model predictions for $A = 1$ versus $A = 0$. Because the data are conditionally independent of $f()$ given $\boldsymbol{\pi}$, we need not incorporate the distribution of the data directly. This expected difference in model predictions is marginalized over the uncertain model parameters, so uncertainty about the causal effect reflects uncertainty over parameters $\boldsymbol{\pi}$. The expression in represents a prior distribution for $\bar{\tau}(U = u)$, since it is a joint distribution of the CATE and the prior distribution $p\left(\boldsymbol{\pi}\right)$. Conditioning on the data gives us the posterior distribution for the CATE, $p\left(\bar{\tau}(U = u) \mid Y\right)$, where we would integrate over the posterior distribution of the parameters instead of the prior distribution.

[6] Although researchers are focusing more attention on estimation methods that focus on these statistical assumptions themselves, either by model ensembles/averages or "robust" models for propensity and response.

Probabilistic expressions about plausible and implausible causal effects are the key feature of a Bayesian causal model. Building probabilistic expressions by conditioning on data is a natural way to think about causal inference. Having seen the data, what is the likely range of causal effects? For this reason, Bayesian causal models are a natural way to think about causal inference. Although this sounds banal, any similar language invoked in non-Bayesian analyses is done without any formal basis, since updating the probability of causal effects from data requires some prior for the causal effects. Non-Bayesian methods conduct inference by judging that the *data* were unlikely to arise if the true effect were some fixed value (usually zero). The only way this becomes a statement about which parameters are likely or unlikely is by invoking decision rules that are external to the model to compress evidence into a binary conclusion about parameters.

Causal models, at their core, are theoretical models for counterfactual data. Because Bayesian models are joint generative models for parameters and data, the model contains all machinery required to characterize the probability of counterfactual potential outcomes directly. To see this in action, remember that we can "run the model forward" to create a predictive distribution for Y given treatment, covariates, and parameters. Denote these simulated observations as Y' to distinguish them from the data Y that are actually observed. 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 from our prior parameters (Gelman et al. 2013),

$$p\left(Y' \mid A = a, U = u\right) = \int p\left(Y' \mid A = a, U = u\boldsymbol{\pi}\right) p\left(\boldsymbol{\pi}\right) \mathrm{d}\boldsymbol{\pi} \tag{3.13}$$

again fixing the treatment status and covariate profile. We update this distribution by conditioning on new data, delivering a "posterior predictive distribution"—the distribution of data that we expect from the posterior parameters.

$$p\left(Y' \mid Y, A = a, U = u\right) = \int p\left(Y' \mid A = a, U = u, \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi} \mid Y\right) \mathrm{d}\boldsymbol{\pi} \tag{3.14}$$

These predictive distributions are the natural machinery of Bayesian generative models: the joint distribution and the laws of probability dictate which distributions can be obtained by conditioning and marginalizing over different variables in the model.[7] The causal identification assumptions enable the model to generate counterfactual data by setting the treatment A to some other value a'. Denote these counterfactual observations $\tilde{Y}$, which I will subscript i to show that this model implies a probability distribution for individual data points as well as aggregate treatment effects.

$$p\left(\tilde{Y}_i \mid Y, A_i = a', U_i = u\right) = \int p\left(\tilde{Y}_i \mid A_i = a', U_i = u, \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi} \mid Y\right) \mathrm{d}\boldsymbol{\pi} \qquad (3.15)$$

Stated more simply: if causal models are hypothetical models that *define* potential outcomes, Bayesian causal models assign probability distributions to potential outcomes. These probability distributions are defined in the prior and after conditioning on observed data, and they can be defined all the way to the unit level as long as the Bayesian model contains a probability model for the unit-level data. When viewed this way, a Bayesian causal model is nothing more remarkable than a Bayesian model for missing data, where the missing data are unobserved potential outcomes.[8]

Rubin (1978) Rubin (2005) Baldi and Shahbaba (2019)

Rubin (1981) Meager (2019) Green et al. (2016)

Imbens and Rubin (1997) Horiuchi, Imai, and Taniguchi (2007)

Green and Kern (2012) Guess and Coppock (2018)

Ratkovic and Tingley (2017) Carlson (2020)

[7] Simulations of this sort are possible under any likelihood-based model that posits a generative probability distribution for the data. Bayesian posterior predictive distributions are wider than the likelihood-only distributions, since the Bayesian predictions marginalize over the parameters instead of merely conditioning on the parameter values that maximize the likelihood function.

[8] If we can obtain average causal effect estimates with group-level statistics without the need for the raw data, we can avoid modeling unit-level data altogether. Naturally, models of this type may not define probability distributions for counterfactual observations without assumptions for the distribution of units around their means. In cases such as binary outcome data, however, means in each group are sufficient statistics for the raw data in each group, so the unit level model is implied by the group-level model. See Section 3.5.4 for explanation and examples.

Ornstein and Duck-Mayr (2020) Branson et al. (2019) Chib and Jacobi (2016)

Lattimore and Rohde (2019)

This is possible by having a joint model for the data and parameters…

With the joint model, we can write a probabilistic expression for unobserved potential outcomes directly. This is possible because, under the generative model, our model for counterfactual data is nothing more than a missing data model. (Rubin) We are simply churning out new data points, tweaking the treatment status.

- Direct counterfactual estimation:
    - Ratkovic and Tingley (2017):
        - directly predicts counterfactuals using a massive basis function set, screening, and then sparse modeling on the reduced basis set
        - treatment effects are estimated as individual functions of covariates, approximated as instantaneous derivatives
        - average effects are just averaging over the relevant sample
    - David Carlson IMC for synthetic control

Guidelines:

- we have a probability distribution for an unknown potential outcome
- this gives us a probability distribution for a unit causal effect
- and a potential outcome for a mean
- This is *missing data problem* (Rubin 1978). Assumptions let us model E[Y | A] for all A.

Causal model: $y_i(1)$ and $y_i(0)$

Observed data: $y_i = z_i y_i(1) + (1 - z_i) y_i(0)$

Unobserved data: $\tilde{y}_i$

Inference about $\tilde{y}_i$: $p(\tilde{y}_i \mid \mathbf{y}) \propto p(\tilde{y}_i \mid \pi, \mathbf{y}) p(\pi)$

This fits an ML approach:

- ML approach to causal inference recasts the problem as a prediction problem for the treatment assignment.
- Treatment effects are then integrated over the uncertainty in the propensity
- This propagates uncertainty using the exact model machinery, rather than a post-hoc computational workaround like bootstrapping
  - Bootstrapping does have an appealing feature that it isn't making a parametric assumption about errors in the model. It simply uses the "sampling uncertainty" intuition to build intuitive bounds on effect uncertainty.
  - However in many real-world cases the data we have cannot be resampled, so the framework for inference under bootstrapping doesn't make theoretical sense without appealing to a "superpopulation" philosophy.
  - Bayes naturally deals with this by assigning a probability distribution to the unmodeled features, which mechanistically follows a similar assumption as a parametric assumption about errors in any typical regression case: the prior isn't really extra.
  - If this assumption isn't something you want, it is possible to generalize the model by specifying an overdispersion parameter and letting the data inform what which model estimates can be rationalized against plausible overdispersion parameters. For example, Kruske's BEST approach to difference-in-means testing.
- In a Bayesian framework, we have posterior More of a "priors to facilitate posteriors" approach, not a prior information view (since there is some formal data peeking involved)

Works without repeated sampling.

Probabilistic treatment status?

3.2.2 Bayesian Inference in the Hierarchy of Causal Inference

What was I trying to say here?

A causal parameter is a feature of the *causal model*.

- The causal model is the model if you knew everything (level 1).
- Individual causal effect is only a transformation of *data*. if I knew all potential outcomes, I could calculate this with only data.
- Estimands are different forms of causal effects, individual or aggregate.
- Assumptions let us state the conditions under which an individual or aggregate parameter can be estimated from observed data only (level 2).

Estimands are framed in terms of individual effects or aggregate expectations.

- The estimator (3) of an expectation or a counterfactual is thus a different thing from the hypothetical estimand.
- Some causal models imply nonparametric estimators, so no Bayesian anything will be required. However, estimates themselves are in-sample quantities, and any generalization about the in-sample effect can be augmented with priors.
- Bayes is an *estimation* framework that doesn't change the causal model itself. It is a different way of estimating the quantities in an in-sample causal estimand.
- 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 $E[Y]$

Bayes is - Bayesian inference *begins* as technology for thinking about statistical assumptions. - It eventually becomes technology for experimenting with identification assumptions.

- 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.

Do battle with the "implied nonparametric estimator" framework.

- Causal models are manipulated to express causal *estimands*.
- the "nonparametric estimators" are usually just averages of treatment and control group outcomes, under ignorability assumptions.
- Parametric models for estimands are usually a byproduct of thinking within an MLE/regression framework. Many causal inf techniques don't do that. Instead they try to simply predict E[y(1) - y(0)] and don't care about the interpretability of the other RHS terms.
- Thinking outside the "typical econometric" framework it's actually kinda easy to see how one flavor of fitting Bayesian models for causal effects. If the technology for prediction is Bayesian, you just get the prediction and the posterior distribution. You then deal with the bias/variance properties of $\hat{y}$ or $\tilde{y}$ (unobserved counterfactuals) but at least your focus is on the predictions rather than coefficients (which, from this view, who cares about those?)

- This actually fits quite nicely with the machine learning approach to causal inference

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 don't like (Navarro, "Devil and Deep Blue Sea"). Me debating with myself: how to think about Bayesian model selection vs "doubly robust" estimation ideas… Maybe some hybrid view in the "quasi-experimental" approach to causal model selection. We estimate two models from the same data—one with a treatment parameter and one where we impose no treatment effect—and compare models using some likelihoodist or Bayes factor measure of evidence.

What is THIS project going to do?

- Pragmatic view of priors?
- Are we doing more flexible covariate adjustment?
- Maybe we should decide this AFTER we experiment with Ch 4 and 5 data/methods

3.2.3 Inferential Goals

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 don't 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 don't.

Proper frequentist analysis is violated as soon as you look at the data.

Are we in a repeated sampling framework at all?

Frequency properties are still possible for Bayesian estimators, but we view frequency properties as a byproduct of something more essential (MSE).

3.3 Making Sense of Priors in Causal Inference

3.3.1 Priors as Data Falsification

Data falsification versus unavoidable choice: imagine a study with a posterior distribution $p(\mu \mid \mathbf{y})$ that is proportional to the likelihood times the prior.

$$p(\mu \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mu) p(\mu) \tag{3.16}$$

Rewrite the right side in product notation for n observations of y_i for units indexed by i, letting $l(y_i) = p(y_i \mid \mu)$

$$p(\mu \mid \mathbf{y}) \propto \prod_{i=1}^{n} l(y_i) p(\mu) \tag{3.17}$$

Suppose that we express this proportionality on the log scale, where the log posterior is proportional to the log likelihood plus the log posterior.

$$\begin{aligned} \log p(\mu \mid \mathbf{y}) &\propto \sum_{i=1}^{n} \log l(y_i) + \sum_{i=1}^{n} \log p(\mu) \\ &\propto \log l(y_1) + \log p(\mu) + \log l(y_2) + \log p(\mu) + \ldots + \log l(y_n) + \log p(\mu) \end{aligned} \tag{3.18}$$

The setup in (3.18) highlights a few appealing intuitions. First, it shows how each observation "adds information" to the log posterior distribution. Data that are more likely to be observed given the parameter (larger $l(y_i)$ values) increase the posterior probability that parameter. We also see that the prior probability "adds information" to the posterior in the same way data add information, captured by the addition of each $p(\mu)$ term. Parameters that are more probable in the prior are more probable in the posterior.

The proportionality (3.18) also reveals how the posterior "learns" from flat priors. A flat priors implies that prior probability $p(\mu)$ is constant for all potential values of μ. Because (3.18) is a proportionality, this lets us disregard $p(\mu)$ entirely by factoring it out of the proportionality, leaving us with an expression that the log posterior is proportional to the likelihood of the data only if the prior is flat.

$$\log p(\mu \mid \mathbf{y}) \propto \log l(y_1) + \log l(y_2) + \ldots + \log l(y_n) \tag{3.19}$$

If $p(\mu)$ is not flat, however, and $p(\mu)$ varies across values of μ, we can no longer ignore $p(\mu)$ in (3.18) shows that $p(\mu)$ varies across values μ. Not only does this prevent us from dropping $p(\mu)$ from the proportionality, but it also reveals how the prior "adds information" to the posterior by the same mechanism that observations do: adding to the log posterior distribution. This general expression where both data and priors contribute to the posterior distribution has led some researchers to argue the Bayesian inference with non-flat priors is analytically indistinguishable from data falsification (García-Pérez 2019). We can highlight this behavior by obscuring each $p(\mu)$ term with a square □.

$$\log p(\mu \mid \mathbf{y}) \propto \log l(y_1) + \square + \log l(y_2) + \square \ldots + \log l(y_n) + \square \tag{3.20}$$

Behind each square is *some contribution* to the log posterior. The fact that it adds information to the log posterior is unaffected by whether the hidden term is the probability of an additional observation $l(y_i)$ or the prior probability of a parameter value $p(\mu)$.

3.3.2 Priors and Model 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 unaffected.

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\left(\frac{x-\mu}{2\sigma}\right)^2} = \sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau(x-\mu)^2}{2}} \tag{3.21}$$

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.

[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.

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.

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \text{where } \varepsilon_i \sim \text{Normal}\left(0, \sigma\right) \tag{3.22}$$

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.

$$y_i \sim \text{Normal}\left(\alpha + \beta x_i, \sigma\right) \tag{3.23}$$

Algebraically, these two models are identical. The difference is only a matter of which component has the distributional assumption. In Equation (3.22), the distribution is assigned to ε_i, so y_i is a random variable only by way of ε_i. In Equation (3.23), 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 4 and standard deviation 2.

$$\text{Centered Parameterization:} \quad v \sim \text{Normal}(4, 2) \tag{3.24}$$

$$\text{Non-Centered Parameterization:} \quad v = 4 + 2z, \qquad \text{where } z \sim \text{Normal}(0, 1) \tag{3.25}$$

To demonstrate that these parameterizations are equivalent, I simulate 500 simulations from each parameterization and plot their empirical cumulative distribution functions alongside

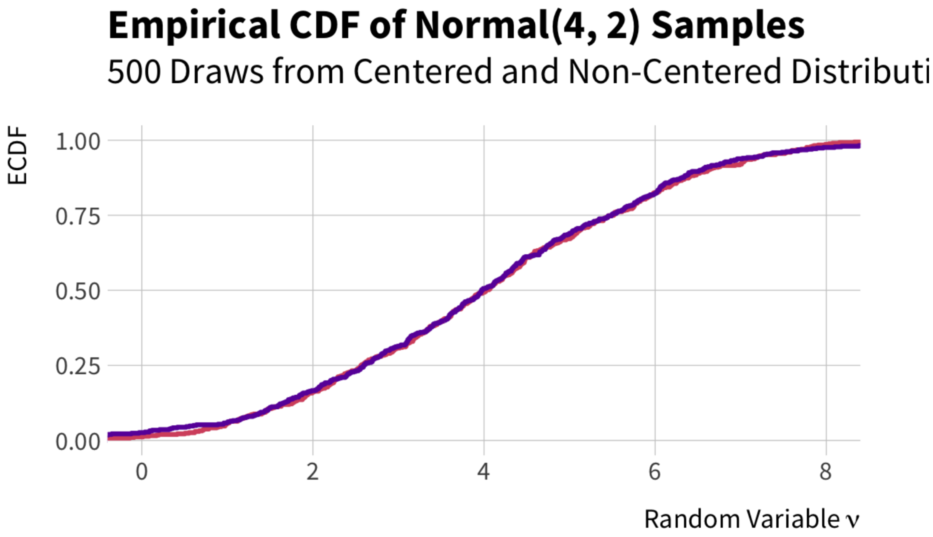


Figure 3.2: Demonstration of centered and non-centered parameterizations for a Normal distribution. The non-centered parameterization is statistically equivalent, but the location and scale are factored out of the distribution.

each other in Figure 3.2. Because the distributions are the same, the empirical CDFs are identical except for random sampling error.

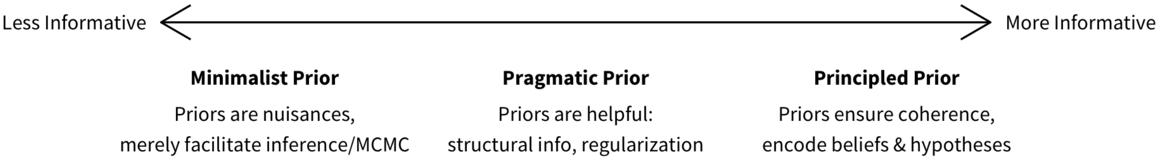


Figure 3.3: A spectrum of attitudes toward priors.

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 don’t 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

3.3.3 Flatness is a Relative, Not an Absolute, Property of Priors

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 $\overrightarrow{\alpha}$.

$$p\left(\overrightarrow{\alpha} \mid y\right) \propto p(y \mid \overrightarrow{\alpha})p(\overrightarrow{\alpha}) \tag{3.26}$$

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

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.

3.3.4 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'm 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 don't 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 don't 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 don't observe.

3.4 Pragmatic Bayesian Modeling

3.4.1 Orientations Toward Prior Distributions

Various roles that priors take on

- merely facilitate posterior inference
- structural information
- weak information
- regularization / stabilization

- 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

3.4.2 Modeling Cultures in Political Science: Complexity and Agnosticism

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

3.4.3 Principled Approaches to Model Parameterization

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,

$$y_i = \alpha + \beta z_i + \varepsilon_i \tag{3.27}$$

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.

$$y_i = z_i \mu_1 + (1 - z_i) \mu_0 + \varepsilon_i \tag{3.28}$$

Although the treatment effect β from Equation (3.27) is equivalent to the difference in means $\mu_1 - \mu_0$ from Equation (3.28), 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 3.4. 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.

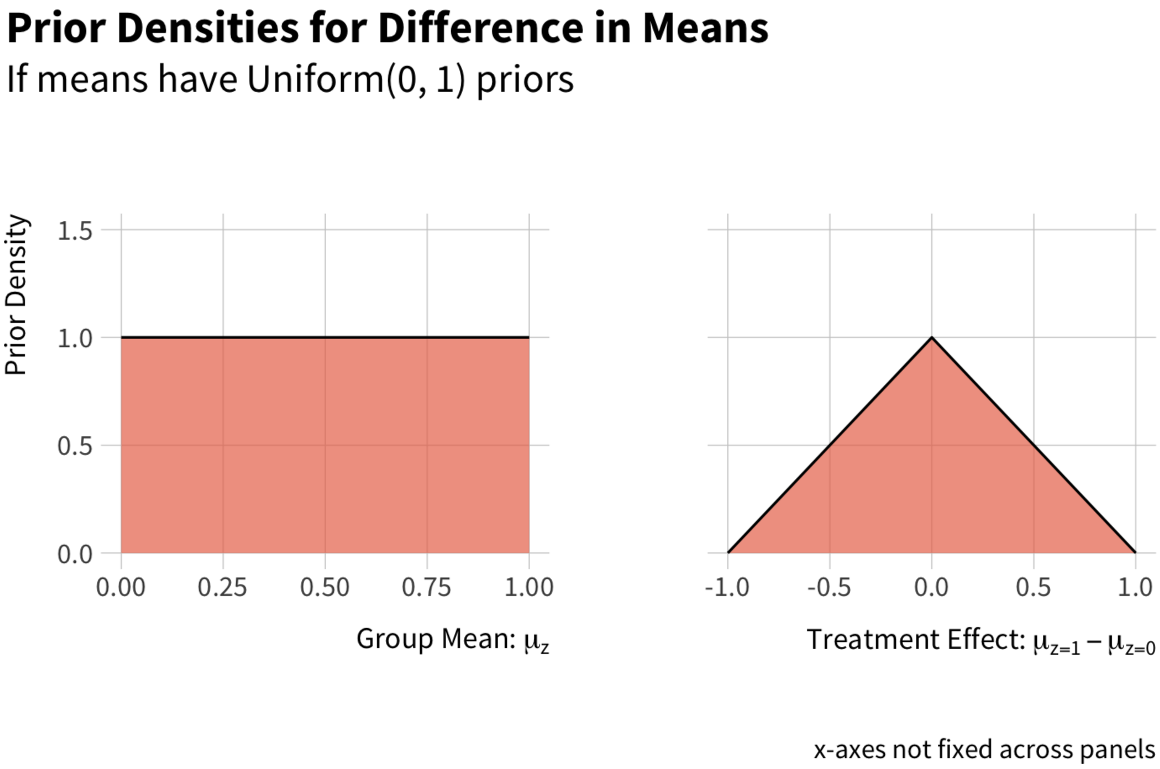


Figure 3.4: Model parameterization affects prior distributions for quantities of interest that are functions of multiple parameters or transformed parameters. The left panel shows that the difference between two means does not have a flat prior if the two means are given flat priors. Note that the x-axes are not fixed across panels.

- equivalent parameterizations

3.4.4 Structural Priors and Weak Information

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

p doesn't care about your n.

3.4.5 Understanding Log Prior *Shape*

This is low-key pretty big

3.4.6 Regularization-Induced Confounding

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

3.4.7 Priors for Imperfect Identifiability

"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)

3.5 Bayesian Opportunities

3.5.1 Full Posterior Uncertainty

Multi-stage models

- uncertain measures *are priors*
- propensity models (Zigler, Plummer/cutting)
- structural models
- to bootstrap or not to bootstrap?

Matching

Predictive/ML models

Multiple comparisons and regularization

3.5.2 Priors for Key Assumptions

Relatedly: priors over models

3.5.3 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).

$$\lim_{x \downarrow x_0} \mathbb{E}\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow x_0} \mathbb{E}\left[Y_i \mid X_i = x\right] = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = x_0\right] \tag{3.29}$$

Equation (3.29) 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 (3.29); 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,

$$\begin{aligned} y_{dpt} = \beta_0 + \beta_1(\text{Extremist Wins Primary})_{dpt} + \beta_2(\text{Extremist Primary Margin})_{dpt} \\ + \beta_3(\text{Extremist Wins Primary} \times \text{Extremist Primary Margin})_{dpt} + \varepsilon_{dpt} \end{aligned} \tag{3.30}$$

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*),

$$\begin{aligned} y_{dpt} &= \alpha_{w[dpt]} + \beta_{w[dpt]}(\text{Extremist Primary Margin})_{dpt} + \varepsilon_{dpt} \\ \varepsilon_{dpt} &\sim \text{Normal}(0, \sigma) \end{aligned} \tag{3.31}$$

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

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.

I plot the OLS win probability estimates around the discontinuity in Figure 3.5. At the discontinuity, we estimate that extremism decreases a candidate's win probability by 0.53 percentage points, which is the same effect found by Hall (2015). 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.

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

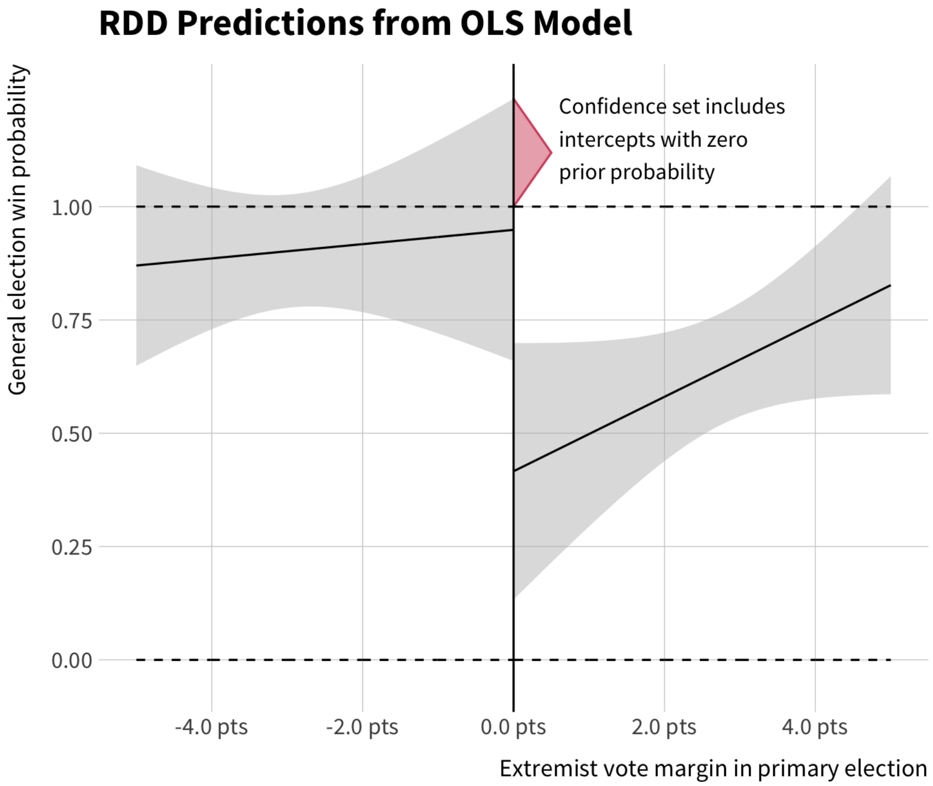


Figure 3.5: OLS estimates of the predicted probability that a candidate wins the general election. Local average treatment effect is estimated at the threshold. Treatment effect is defined by parameter estimates whose confidence sets contain values with no possibility of being true.

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

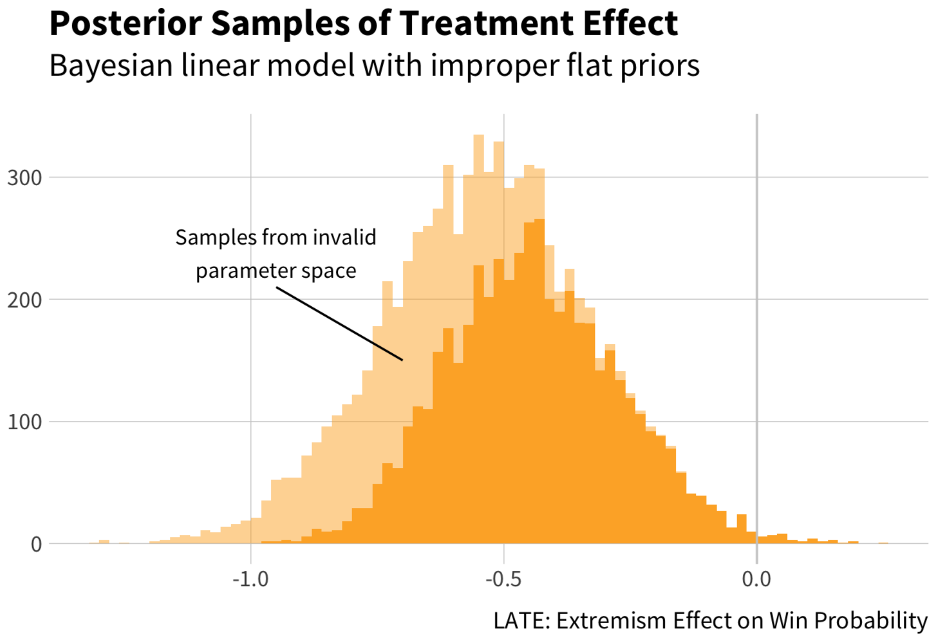


Figure 3.6: Histogram of posterior samples for the treatment effect. Using flat priors for each win probability intercept, a substantial share of the treatment distribution consists of parameters that cannot possibly occur.

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.

$$\alpha_{w=0}, \alpha_{w=1} \sim \text{Uniform}\left(0, 1\right) \tag{3.32}$$

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 3.4, which is vague despite not being flat.

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

$$\begin{aligned} y_{dpt} &\sim \text{Normal}\left(\alpha_{w[dpt]} + \beta_{w[dpt]}(\text{Extremist Primary Margin})_{dpt}, \sigma\right) \\ \beta_{\text{w}} &\sim \text{Normal}\left(0, 10\right) \\ \sigma &\sim \text{Uniform}(0, 10) \end{aligned} \tag{3.33}$$

The Normal model for the outcome data in the first line is equivalent to the Normal error term defined in (3.31). 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,

$$y_{dpt} \sim \text{Bernoulli}\left(\pi_{dpt}\right) \tag{3.34}$$

$$\text{logit}\left(\pi_{dpt}\right) = \alpha^*_{w[dpt]} + \beta^*_{w[dpt]}(\text{Extremist Primary Margin})_{dpt} \tag{3.35}$$

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.

$$\begin{aligned} \text{logit}\left(\pi_{dpt}\right) &= \alpha^*_{w[dpt]}, \text{ at Extremist Primary Margin}_{dpt} = 0 \\ \text{which implies } \pi_{dpt} &= \text{logit}^{-1}\left(\alpha_{w[dpt]}\right) \end{aligned} \tag{3.36}$$

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,

$$\alpha_w^* \sim \text{Logistic}\left(0, 1\right) \tag{3.37}$$

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 Logistic $\left(0, 1\right)$ as an uninformative prior for a key parameter. But at discussed in Section 3.3.3, 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 3.7 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 $\left(0, 1\right)$ 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 $\left(0, 10\right)$ 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,

$$\beta_w \sim \text{Normal}\left(0, 10\right) \tag{3.38}$$

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.

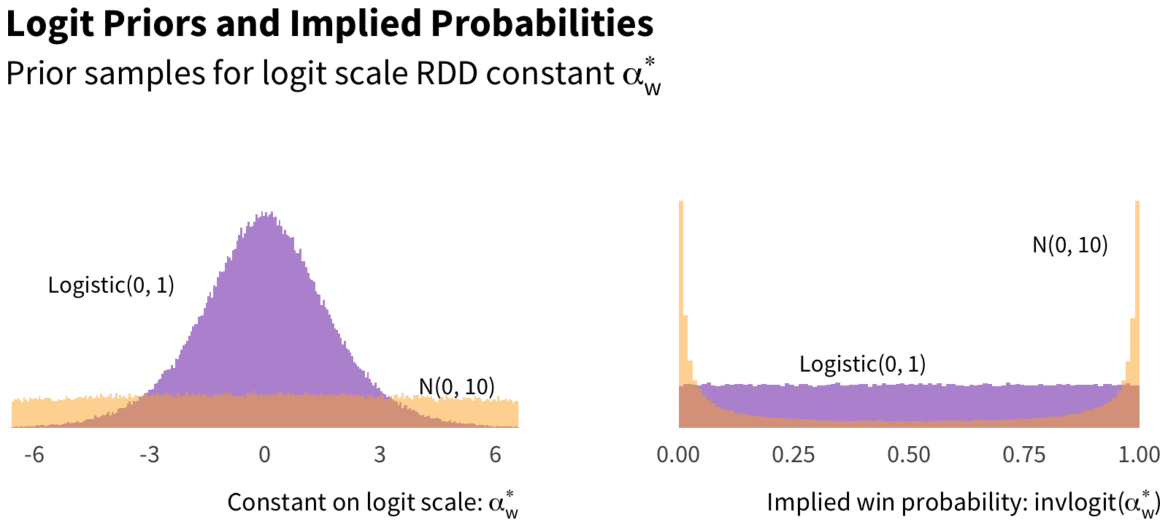


Figure 3.7: Scale invariance of logit model priors. Standard logistic prior on logit scale becomes a flat prior on the probability scale. "Diffuse" priors on logit scale imply priors on probability scale that bias toward extreme probabilities.

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 before seeing any data. What effect do these minor interventions have? Figure 3.8 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 3.8 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 3.6, 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-

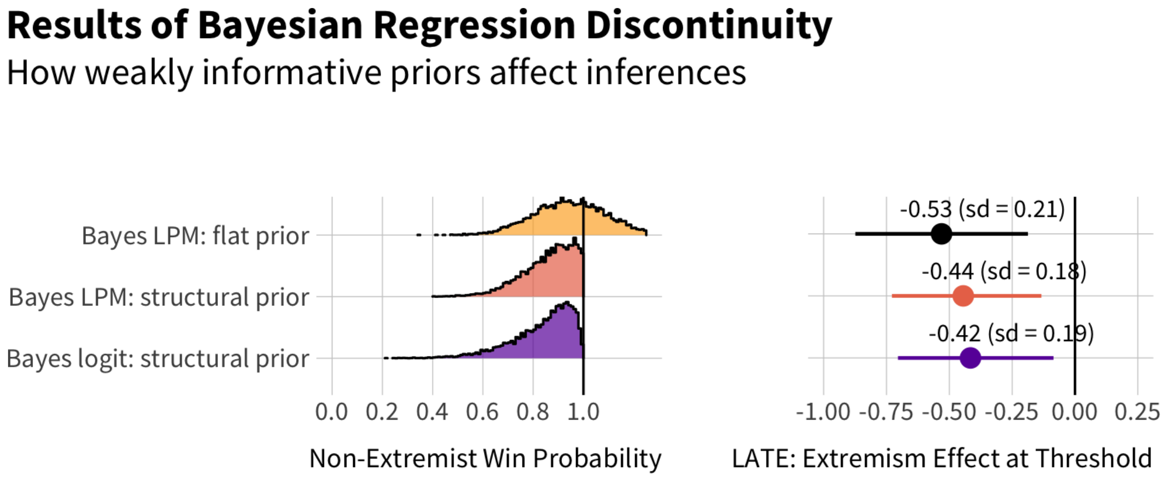


Figure 3.8: Comparison of posterior samples from Bayesian models with improper flat priors and weakly informative priors. Weakly informative models have flat priors over valid win probabilities at the discontinuity. Priors reduce effect magnitudes and variances by ruling out impossible win probabilities.

limited experimental samples, simple interventions like this have the potential to substantially improve the precision of research findings in contexts where researchers don't realize how much information they are leaving out of their analyses.

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

3.6 Other Frontiers of Bayesian Causal Inference

3.6.1 Beyond Estimation: Inferences About Models and Hypotheses

Inherit material from earlier section

3.6.2 Priors are the Basis for all Generalization

No-Free-Lunch theorems

3.6.3 Agnostic Causal Inference

Conventional:

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)p(\theta)}{p(\mathbf{y})} \tag{3.39}$$

$$p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta)p(\theta) \tag{3.40}$$

Implied:

$$p(\theta \mid \mathbf{y}, \mathcal{H}) = \frac{p(\mathbf{y} \mid \theta, \mathcal{H})p(\theta \mid \mathcal{H})}{p(\mathbf{y} \mid \mathcal{H})} \tag{3.41}$$

— 4 —

Constituent Ideology and Candidate Positioning: Bayesian Structural 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 candidates who run to represent that district-party.

It is important to distinguish the influence of the district-party public from the influence of district *partisanship*. Does Senator Susan Collins (R-ME) have a reputation as a moderate Republican senator because the balance of a close numerical balance of Republican and Democratic voters in Maine? Or is it because her Republican constituency in Maine is relatively moderate compared to the Republican constituencies in other states represented by more conservative Senators? 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 district construct from aggregate voting patterns. 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, because district-level voting behavior is certainly affected by district-party ideology, and it likely mediates the effect of district-party ideology on candidate positioning. Estimating the direct effect of district-party ideology by simply controlling for district voting in a regression is likely to introduce collider bias by conditioning on a post-treatment variable (Greenland, Pearl, and Robins 1999).

This chapter investigates the effect of district-party ideology on primary candidate positions using a sequential-*g* model, a multistage approach that estimates the direct effect of district-party ideology while holding fixed district-level voting, a likely mediator of the total effect (Acharya, Blackwell, and Sen 2016). I use causal graphs to illustrate the modeling problem and discuss the assumptions required for identifying the targeted effect, which I adapt to a multilevel context where treatment and mediators exist at different levels of the data hierarchy. Finally, I describe and implement a method for propagating measurement uncertainty through the sequential-*g* method, since the key independent variable is an uncertain estimate from a measurement model.

4.1 Constituency Preferences and Candidate Positioning

Lit review:

- I have lit review but it isn't well organized right now. I'll summarize some main points:
- classic models of electoral competition view candidate moderation as an electoral benefit. You can make non-moderate positions make sense if you incorporate campaign

volunteers, primaries, etc.

- Researchers find evidence that district-level aggregate voting (e.g. the presidential vote) is related to candidate positioning.
- Why do candidates take non-moderate stances in reality? Studies that look into the polarizing effects of primaries find little evidence that primaries matter for candidate positioning, but many of these studies are focused on Congressional incumbents.
- Studies that include non-incumbent candidates show that many of the empirical implications of the "strategic positioning dilemma" don't receive a lot of evidence. For instance, studies of primary "openness" consistently show basically zero or even reversed relationships to candidate ideology as you would expect from theory.

The second obstacle preventing a more complete study of primary representation is the failure to incorporate primary candidates' ideal points into the analysis. Although ideal point estimates derived from roll-call votes such as Nominate are a popular tool for measuring politicians' ideological locations (Poole 2005; Poole and Rosenthal 1997), only incumbents cast roll call votes, so these measures are unavailable for non-incumbent candidates.[1] Further, when non-incumbents enter the picture, researchers tend to focus on the positioning of general election candidates rather than primary candidates (Ansolabehere, Snyder, and Stewart 2001; Burden 2004; Canes-Wrone, Brady, and Cogan 2002). Some studies have argued that primary competition leads incumbent legislators to take non-median positions, but these studies do not observe primary candidate positions directly, instead observing the presence or threat of challengers (Brady, Han, and Pope 2007; Burden 2004). Recent advances in ideal point modeling using campaign contributions are a promising path forward (Bonica 2013, 2014; Hall and Snyder 2015), but they are not designed for the careful study of

[1] Studies of candidate positioning that go beyond incumbents sometimes use survey data from challenger candidates (Ansolabehere, Snyder, and Stewart 2001; Burden 2004), but the surveys only interview general election candidates. Furthermore, the rarity of these surveys limits the generalizability of their findings over time.

primary competition and thus contain many “post-treatment” measurement artifacts.

What to do with the Hopkins/Sniderman theory

- does party’s issue emphasis mean parties should collapse, or that variation should track
- Grossman-Hopkins; parties care about issues so they try to position themselves on it
    - read their stuff to see how they quality this
    - do they think districts collapse?
- weighting issues: left groups see fine variation in left members, but republicans are all bunched at zero (Brunell? Linda Fowler?)
- or Sniderman conflicted theory: Democrats are conflicted about appeasing groups, so variation matters. Republicans are conflicted
- THINK about abortion as a toy case
    - if weighting… if not weighting…

4.2 How Partisan Preferences Affect Candidates

4.2.1 Naive regression

- what is the relationship
- let’s control for the thing
- what does this imply?

— 5 —

District-Party Ideology and Primary Election Outcomes: Bayesian Double Machine Learning

What is the research question

- choice as f(ideology)
- stratify on primary type?
- stratify in open seats?

What's the plan?

- conditional logit vs. Hainmueller et al. OLS choice?
- interact candidate and group ideal point: how does p(win) change?
- concern: is the CF data shit?
    - estimate victory among incumbents running for reelection, what's p(win)?
    - is there not enough data?

covariates:

- what is fixed within choice set?

  - ideal points
  - party of incumbent
  - cycle

5.1 Machine Learning and Causal Inference

ML is basically just different methods for $E[Y|X]$

5.1.1 Why ML and Causal Inference

ID assumptions are all about $E[Y|X]$.

- models are post-identification
- if we have to model, we like robustness and flexibility
- we are specification and functional form dependences

Problems:

- don't know the covariates
- even if we did, don't know the model

Solution: ML

- regularization and other covariate selection methods (not necessarily the purview of "machine learning")
- many models are set up for interactions, nonlinearities
- The causal models themselves are derived in terms of expectations, not in terms of regression functions.

5.1.2 how does ML work?

workflow:

- model tuning, train/test split
- penalize overfitting
- regularization and shrinkage priors

5.1.3 Bayesian causal ML

Notation:

- model predictions are inherently posterior predictions

  Natural analogs:

- ridge
- lasso
- extensions

  Benefits:

- uncertainty in ML predictions has been a problem
- Bayes gives a straightforward, unified framework for characterizing uncertainty in model parameters (feeding forward into model predictions)
- handles non-identifiability in certain models pretty naturally!

5.2 Bias and Variance in Causal Identification

5.2.1 Regularization and Shrinkage Priors

5.2.2 Neyman Orthogonalization

We can regularize in the first stage, then we try to be more agnostic in the second stage?

5.3 Empirical Setup

5.3.1 Data

5.3.2 Neural Network Choice Model

Conditional choice likelihood

Linear Index is Neyman-Parameterized

Neural net in the selection stage

5.3.3 Merging Model Subposteriors

5.3.4 Modeling Testing Workflow

Model validation with LOO

Searched Parameters

LOO results

5.4 Findings

Group IRT Model

Colophon

This document was produced using R, R Markdown, LaTeX. The document was built using the bookdown package for R. The document template is a variation on TJ Mahr's buckydown template, which was itself an adaptation of documents designed by and for students at the Universities of Washington and Wisconsin. The PDF is typeset using pdfTeX. The body text is *MinionPro-LF* in 12pt size.

The data and source code for this dissertation have been organized into an online Git repository on Bitbucket. A hard copy of the thesis can be found in the University of Wisconsin library system.

- Git repository: https://bitbucket.org/mikedecrescenzo/dissertation
- huskydown: https://github.com/benmarwick/huskydown
- bookdown: https://github.com/rstudio/bookdown

This version of the thesis was generated on 2020-07-22 10:34:52. The repository is currently at this commit:

```
## Commit:  c629d884d9360ecc054b7b59da018148695fd7f5
## Author:  Michael DeCrescenzo <mgdecrescenzo@gmail.com>
## When:    2020-07-21 01:11:02 GMT
```

```
##
##      more ch 1 writing, other org
##
## 10 files changed, 509 insertions, 123 deletions
## 10_arg.Rmd                                   | -24 + 98  in 9 hunks
## 20_model.Rmd                                 | - 3 +  1  in 1 hunk
## 23_model-posteriors.Rmd                      | - 1 +  1  in 1 hunk
## 30_causality.Rmd                             | - 9 + 14  in 5 hunks
## 40_positioning.Rmd                           | -23 + 61  in 4 hunks
## 50_voting.Rmd                                | -10 + 96  in 2 hunks
## assets-bookdown/thesis-bib.bib               | - 1 + 89  in 2 hunks
## code/05-voting/52_model-test.R               | - 0 +  4  in 1 hunk
## notes/agenda.rmd                             | -52 +145  in 3 hunks
## present/apw-2020/MGD-thesis-APW-excerpt.pdf | - 0 +  0  in 0 hunk (binary file)
```

References

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.

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. 1983a. "A downsian spatial model with party activism." *American Political Science Review* 77(04): 974–990.

Aldrich, John H. 1983b. "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.

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, Jonathan Rodden, and James M Snyder Jr. 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.

Aronow, Peter M, and Cyrus Samii. 2016. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60(1): 250–267.

Baldi, Pierre, and Babak Shahbaba. 2019. "Bayesian causality." *The American Statistician*: 1–9.

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.

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, 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.

Boatright, Robert G. 2013. *Getting primaried: The changing politics of congressional primary challenges*. University of Michigan Press.

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.

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.

Branson, Zach et al. 2019. “A nonparametric bayesian methodology for regression discontinuity designs.” *Journal of Statistical Planning and Inference*.

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.

Burden, Barry C. 2004. "Candidate positioning in u.s. Congressional elections." *British Journal of Political Science* 34(02): 211–227.

Burke, Edmund. 2012. "Speech to the electors of bristol at the conclusion of the poll. 1774." *Acesso em* 28.

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.

Canon, David T. 1999. *Race, redistricting, and representation: The unintended consequences of black majority districts*. University of Chicago Press.

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.

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.”

Chib, Siddhartha, and Liana Jacobi. 2016. “Bayesian fuzzy regression discontinuity analysis and returns to compulsory schooling.” *Journal of Applied Econometrics* 31(6): 1026–1047.

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, and Mathew D McCubbins. 2005. *Setting the agenda: Responsible party government in the us house of representatives*. Cambridge University Press.

Downs, Anthony. 1957. *An economic theory of democracy*. New York: Harper; Row.

Enns, Peter K, and Julianna Koch. 2013. "Public opinion in the us states: 1956 to 2010." *State Politics & Policy Quarterly* 13(3): 349–372.

Erikson, Robert S, and Gerald C Wright. 1980. "Policy representation of constituency interests." *Political Behavior* 2(1): 91–106.

Fenno, Richard F. 1978. *Home style: House members in their districts*. Pearson College Division.

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.

Fox, Jean-Paul. 2010. *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

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, 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.

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. 2012. *Affluence and influence: Economic inequality and political power in america*. Russel Sage Foundation; Princeton University Press.

Gill, Jeff. 2014. 20 *Bayesian methods: A social and behavioral sciences approach*. CRC press.

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.

Grossman, Matthew, and David A. Hopkins. 2016. Oxford University Press *Asymmetric politics: Ideological republicans and group interest democrats*.

Guess, Andrew, and Alexander Coppock. 2018. "Does counter-attitudinal information cause backlash? Results from three large survey experiments." *British Journal of Political Science*: 1–19.

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.

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.

Hirano, Shigeo et al. 2010. "Primary elections and party polarization." *Quarterly Journal of Political Science* 5: 169–191.

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.

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.

Keele, Luke. 2015. "The statistics of causal inference: A view from political methodology." *Political Analysis* 23(3): 313–335.

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."

Koger, Gregory, Seth Masket, and Hans Noel. 2009. "Partisan webs: Information exchange and party networks." *British Journal of Political Science*: 633–653.

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

Lattimore, Finnian, and David Rohde. 2019. "Replacing the do-calculus with bayes rule." *arXiv preprint arXiv:1906.07125*.

Lax, Jeffrey R., and Justin H. Phillips. 2012. "The democratic deficit in the states." *American Journal of Political Science* 56(1).

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).

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.

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.

Link, William A., and Mitchell J. Eaton. 2011. "On thinning of chains in MCMC." *Methods in Ecology and Evolution* 3(1): 112–115. https://doi.org/10.1111/j.2041-210x.2011.00131.x.

Londregan, John. 1999. "Estimating legislators' preferred points." *Political Analysis* 8(1): 35–56.

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.

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.

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.

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.

Miller, Warren E, and Donald E Stokes. 1963. "Constituency influence in congress." *American Political Science Review* 57(01): 45–56.

Neal, Radford M. 2012. "MCMC using hamiltonian dynamics." *arXiv preprint arXiv:1206.1901*.

Norrander, Barbara. 1989. "Ideological representativeness of presidential primary voters." *American Journal of Political Science*: 570–587.

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.

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.

Pitkin, Hanna Fenichel. 1967. *The concept of representation*. Univ of California Press.

Poole, Keith T. 2005. *Spatial models of parliamentary voting*. Cambridge University Press.

Poole, Keith T, and Howard Rosenthal. 1997. “Congress: A political-economic history of roll call voting.” *New York: Oxford University Press*.

Rahn, Wendy M. 1993. “The role of partisan stereotypes in information processing about political candidates.” *American Journal of Political Science*: 472–496.

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

Rogowski, Jon C. 2016. “Voter decision-making with polarized choices.” *British Journal of Political Science*: 1–22. https://doi.org/10.1017%2Fs0007123415000630.

Rogowski, Jon C, and Stephanie Langella. 2015. “Primary systems and candidate ideology: Evidence from federal and state legislative elections.” *American Politics Research* 43(5): 846–871.

Rubin, Donald B. 1978. “Bayesian inference for causal effects: The role of randomization.” *The Annals of statistics*: 34–58.

Rubin, Donald B. 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.

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.

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).

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 Jr, James M. 1994. "Safe seats, marginal seats, and party platforms: The logic of platform differentiation." *Economics & Politics* 6(3): 201–213.

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.

Tausanovitch, Chris, and Christopher Warshaw. 2013. "Measuring constituent policy preferences in congress, state legislatures, and cities." *The Journal of Politics* 75(02): 330–342.

Treier, Shawn, and D Sunshine Hillygus. 2009. "The nature of political ideology in the contemporary electorate." *Public Opinion Quarterly* 73(4): 679–703.

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. *Bayesian Analysis*.

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.