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) ______________________________

Kenneth R. Mayer ______________________________

Eleanor Neff Powell ______________________________

Alexander M. Tahk ______________________________

5th member ______________________________

Abstract

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

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

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

Contents

List of Tables

List of Figures

— 1 —

Introduction: Primary Elections and Ideological Choice

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 identified a number of positive representational consequences to polarization. Compared to the parties of the early- and mid-1900s that political scientists believed were alarmingly undifferentiated (American Political Science Association 1950), in recent decades the Democratic and Republican Parties have taken increasingly divergent and oppositional stances across a wider variety of issues. Voters have "sorted" into partisan groups that occupy distinct ideological locations in American politics, leading to greater consistency in their issue beliefs, greater ideological abstraction in voter attitudes, and increased political participation (Abramowitz and Saunders 1998; Fiorina, Abrams, and Pope 2005; Layman and Carsey 2002; Levendusky 2009).

Even as polarization has strengthened many aspects of inter-party representation, it affects within-party representation in a number of troubling ways. Party labels provide starker informational and identity cues about candidates than in decades past, but for the typical voter, there is not much of a decision to be made. 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 has increased, there is a sense in which polarization exacerbates the notion of partisan electoral "capture." For most voters, their choices are locked in long before Election Day. Candidates from their preferred parties have been selected through a nomination process, and voters are more likely to abstain from voting when faced with an undesirable candidate than they are to vote against their parties (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 ideological scrutiny of candidates to cast low-cost votes for their own party, weakening the influence of policy preferences in electoral representation overall.

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 are relatively weak venue for democratic accountability, does the U.S. electoral system incorporate these 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, famously studied one-party rule in the American South, asking whether competition within the Southern Democratic 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 single-party dominance is a serious limitation. Within-party representation presents interesting questions that apply to far wider contexts. 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.

This dissertation is chiefly concerned with the policy preferences of partisan voters and their role in electoral representation through primary politics. 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 primary representation, 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 preferences on (1) the positioning of candidates and (2) the outcomes of 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 1983; Brady, Han, and Pope 2007; Burden 2004; Hill 2015).

I argue that if we were to construct ideal tests for this theory, the preferences of partisan primary voters play a key role. First, the theory predicts that candidates strategically position themselves to appeal to primary voters. That is, if we could construct an ideological summary of the partisan primary voters in a district, we ought to find that the level of voter conservatism affects the conservatism of the primary candidates' campaign stances. And second, if primary voters present a credible threat to primary candidates (specifically via their policy preferences), they should vote for the candidate who best represents their policy views. We should observe that as the degree of primary voter conservatism increases, the probability that a more conservative candidate is nominated increases as well. These two predictions are the core empirical implications of the "strategic positioning dilemma" theory of representation in primaries, and they each require researchers to know the preferences of partisan voters within an electoral district. Yet, the preferences of the partisan constituency are either absent or dangerously misconstrued in the existing literature, and we have been unable to answer these key questions as a result.

Stated differently, this dissertation asks if primaries "work" the way we think 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? Do candidates position themselves to win the favor of primary voters? Do primary voters select the candidate who best represents their issue beliefs? And further, do institutional factors that purportedly alter the

[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 moderate the strong partisan composition of primary voters (Hill 2015).

positioning incentives of candidates, such as redistricting, affect which candidates are nominated?

Not only is the evidence in support of this theory sparse or indirect,[2] its theoretical plausibility is not straightforward either. Primary elections present voters with uniquely high informational demands. Because most primary elections are held between candidates of the same party,[3] the party label does not provide the same heuristic cue for voters to impute the positions of the candidates. Although primary voters are generally informed about the typical policy positions associated with their preferred party (Hill 2015), the costs of learning about which candidate stands for which position will be higher than the information costs in a general election. Primary elections typically occur in odd 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. Candidates may determine that communicating their policy positions is less cost-effective than communicating symbolic positions by touting their incumbency or their "outsider" status. The efficacy of these appeals may be correlated with the actual ideological positions of the candidates (for example, if high-quality incumbents are more skilled at perceiving primary voter preferences,(On the perception of voter preferences by politicians, see Broockman and Skovron 2018.) positioning themselves in relation to primary voters, and at communicating symbolically powerful messages), but the information voters are acting on is itself non-spatial. In short, although the proposition that voters' policy preferences affect their primary candidate choices seems straightforward, the informational environment of the primary campaign presents voters with many incentives to vote using non-policy considerations. Indeed, the notion that "voters matter" is a theoretical orientation toward primaries that is not shared by all scholars of electoral representation. 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

[2] I elaborate on the existing evidence later in this chapter.

[3] There are a few exceptions to this general characterization of nomination elections. California, Louisiana, and Washington hold blanket primaries, sometimes called "jungle" primaries, where candidates from all parties compete on one ballot to be included in a runoff general election.

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).

There are two dominant obstacles that have kept the primary constituency from assuming its proper role in the the study of representation. First, and most importantly, the dominant conceptualization and operationalization of constituency preferences is unsuited the study of partisan primaries. Our working theory of primary representation, particularly as it relates to the "strategic positioning dilemma," pits two constituencies within the district against each other: the nominating constituency, which contains only nominating voters from one party, and the general election constituency, which is composed of citizens from both major parties as well as Independents. The former is theorized to prefer ideologically faithful candidates who adhere closely to the party's policy platform,[4] 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.

[4] Strategic voting complicates this characterization, which I discuss below.

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.[5] 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 primary competition and thus contain many "post-treatment" measurement artifacts.

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.[6] 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

[5] 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.

[6] 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.

conceptualizing voter preferences in this way.[7]

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. 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 (**??**) and the vote choices of primary constituents (**??**).

1.1 Limitations for inferring ideal points from votes

This section shows the limitations of inferring voter preferences from vote shares, using notation similar to Kernell (2009).

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

[7] 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.

(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

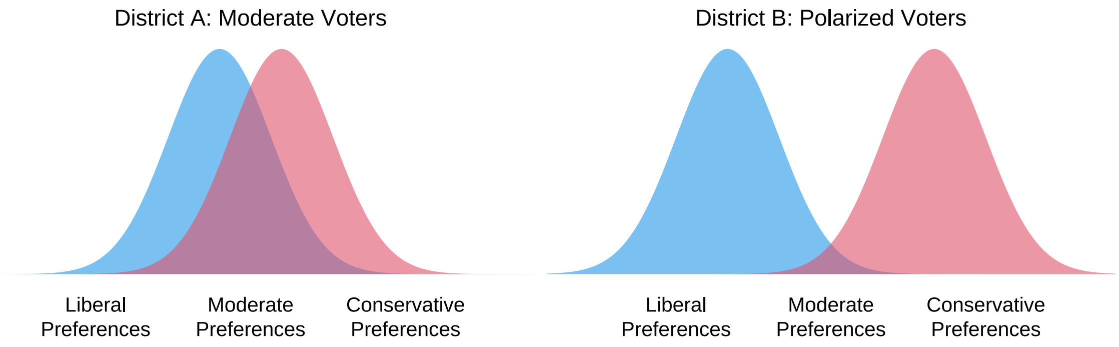

Figure 1.1: Two districts with identically located median voters

1.2 Outline for the dissertation

1.3 Theoretical Orientation

"Spatial" Models of Policy Preferences

Representatives as Delegates vs. Trustees

Proposed Model

Value of directly modeling each party:

- identify the actual quantities of interest
- do parties "move together" or not? This is only possible to infer by modeling each party's hierarchy separately.

1.4 Empirical Orientation

Causal Inference in the Study of Representation

What is the problem?

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

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.

Bayesian Causal Inference

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.

Pragmatic Orientation Toward Causal Inference

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.

— 2 —

Modeling the Partisan Constituency's Policy Preferences

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

This chapter proceeds in three major steps. First, I review the theoretical basis for ideal point models, which can be traced to spatial models of policy choice from classic formal theory work in American political science such as Downs (1957). I connect these formal models to statistical models of policy ideal points (in a style that follows Clinton, Jackman, and Rivers 2004) as well as their connection to Item Response Theory (IRT) models from psychometrics and education testing (e.g. Fox 2010).

Second, I specify and test the group-level model that I build and employ in my analysis of district-party publics. This discussion includes details that are relevant to Bayesian estimation, including identification restrictions on the latent policy space, specification of prior distributions, and model parameterizations that expedite estimation with Markov chain Monte Carlo (MCMC). I begin with a static model for one time period, and then I describe a dynamic model that smooths estimates across time using hierarchical priors for model parameters (Caughey and Warshaw 2015). I test both the static and the dynamic models by fitting them to simulated data and determining how well they recover known parameter values.

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

2.1 Spatial Models and Ideal Points

Ideal points are constructs from *spatial* models of political choice. These models exist under formal theory—they simplify scenarios in the political world into sets of actors whose behaviors obey utility functions that conform to mathematical assumptions. Spatial models invoke a concept of "policy space," where actor preferences and potential policy outcomes are represented as locations along a number line. A canonical example is a left-right continuum, where progressive or "liberal" policies occupy locations on the left side of the continuum, while conservative policies are on the right side. Actors are at least partially motivated by their policy preferences, so they strive to achieve policy outcomes that are closest to their own locations. Depending on the structure of the game, these actors often face constrained choices; they can't achieve their most-preferred policy, so they settle on something that is as close as they can get.

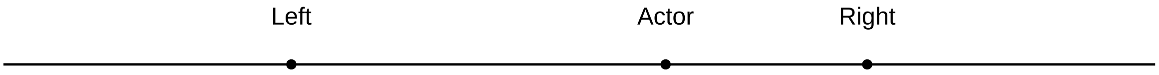

Figure 2.1: An Actor and two policy outcomes (Left and Right) represented as locations in ideological/policy space

Figure 2.1 plots a simple example of an Actor's choice over two policies in one-dimensional policy space. The "Left" outcome is a more progressive policy outcome than the "Right" outcome, indicated by their locations on the line. The Actor has a location herself, which corresponds to her most-desired policy outcome. There is no policy located exactly at the Actor's preferred location, but the Actor is closer to the Right policy than to the Left. Supposing that the Actor could make an error-free choice over which policy to implement, it appears she would prefer the Right outcome to the Left outcome.

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

Formal models of ideal points are distinct from statistical models of ideal points. Formal models are primarily theoretical exercises; they explore the incentives and likely actions of Actors in specific choice contexts, building theoretical intuitions that can be applied in the study of real-world politics with real data. Statistical models, on the other hand, explicitly or implicitly *assume* a formal model as given and estimate its parameters using data. Data could come from legislators casting voting on bills, judges ruling on case outcomes, survey respondents stating their policy preferences (as in this project), and other situations. Researchers are typically interested in parameter estimates for the Actors' ideal points, although sometimes the parameters about the policy alternatives are substantively interesting.

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

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

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

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

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

[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.

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.

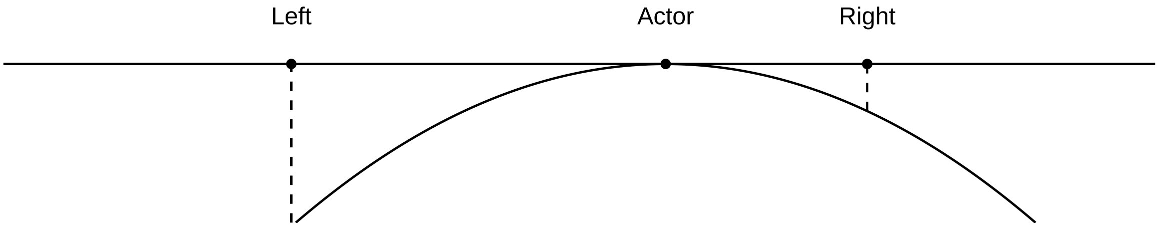

Figure 2.2: A representation of quadratic utility loss over policy choices

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.

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

[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.

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 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.

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

[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.

[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 $\mathrm{Var}\left(u_{ij}^{\mathsf{L}} - u_{ij}^{\mathsf{R}}\right) = \dfrac{s_j^2\pi^2}{3}$, where it is assumed that $s_j = 1$ for the standard Logistic model.

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.

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.

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:

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

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

[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.

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 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 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

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

[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.

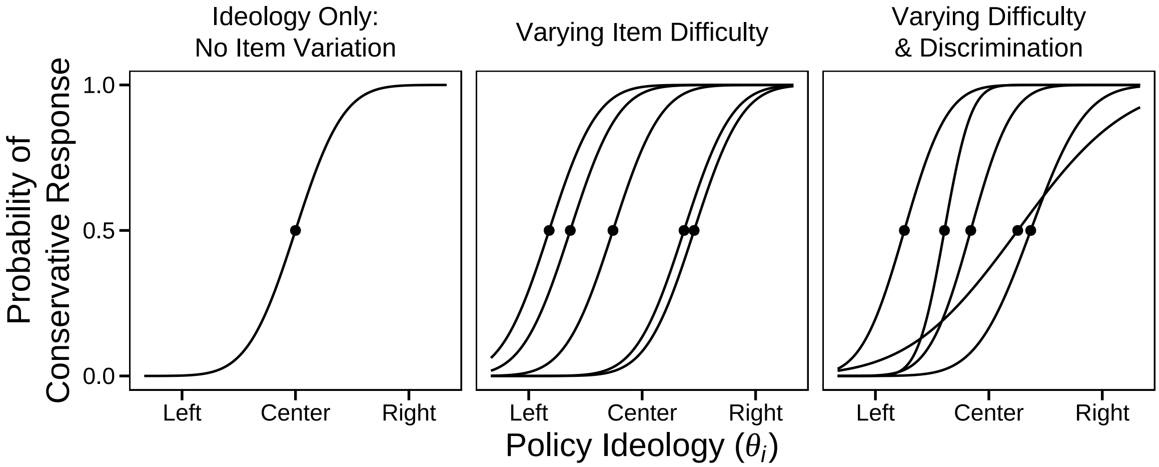

Figure 2.3: Examples of item characteristic curves under different item parameter assumptions

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.

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 $\dfrac{R_j + L_j}{2}$. 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.

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: $\dfrac{\partial \iota_j(\theta_i - \kappa_j)}{\partial \kappa_j} = -\iota_j$, 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 $\dfrac{\partial \iota_j(\theta_i - \kappa_j)}{\partial \iota_j} = (\theta_i - \kappa_j)$. 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) 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 $s_{gj} = \sum\limits_{i \in g}^{n_{gj}} y_{ij}$, 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), 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

[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.

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.

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

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

where $\mu_{p[g]}$ is a constant specific to party p,[13] $\mathbf{x}_d$ is a vector of congressional district-level covariates with party-specific coefficients β_p. State effects $\alpha_{sp}^{\texttt{state}}$ 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 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

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.

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

- 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 $\frac{0}{J} = 0$.

[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.

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 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 $\bar{\theta}_g$.

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 from groups that contain greater variation

[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.

in their survey design weights. It is defined as

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

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

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

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

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

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

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

It is likely that many values of n^*_{gj} and s^*_{gj} will be non-integers. Ordinarily this would be a problem for modeling a Binomial random variable, since a Binomial is an integer-valued count of successes given an integer number of trials and a success probability. For this reason, many statistical programs will return an error if a floating-point argument is bassed to the Binomial probability mass function. Whereas Caughey and Warshaw (2015) calculate $\lceil n^*_{gj} \rceil$ and $\lfloor s^*_{gj} \rfloor$ 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).

2.5 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.

Non-Centered Parameterization

Hierarchical models have posterior distributions whose curvatures present difficulties for sampling algorithms (Betancourt and Girolami 2015; Papaspiliopoulos, Roberts, and Sköld 2007). To improve the estimation in Stan, I program the hierarchical models using a "non-centered" parameterization rather than a "centered" parameterization. Whereas the centered parameterization considers $\bar{\theta}_g$ as a random draw from a hierarchical distribution (Equation (2.16) above), the non-centered parameterization defines $\bar{\theta}_g$ as a deterministic function of its conditional hypermean and a random variable.

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

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

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 the state and region effects.

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

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.

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, 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

[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.

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.

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

Optimized IRT Model

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

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), 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 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?)

[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."

  - 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.

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)

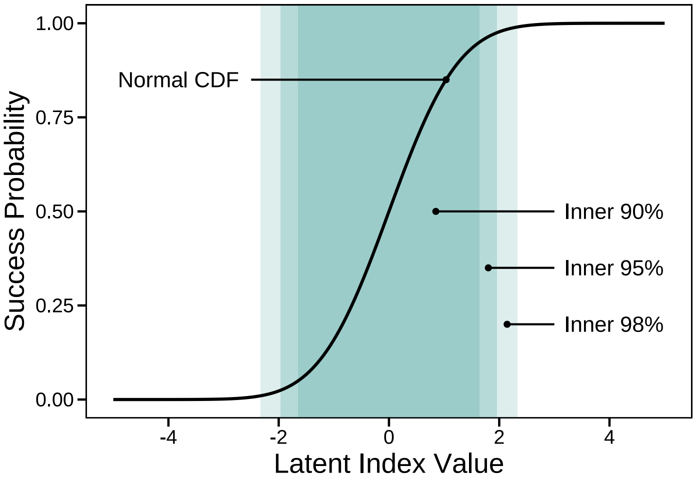

Figure 2.4: The region of the probit model's latent index that maps to response probabilities between 1 and 99 percent.

Item Parameters

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

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

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

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

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

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

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

I then specify priors for the scale terms, $\sigma_{\tilde{\kappa}}$ and $\sigma_{\tilde{\iota}}$, and the correlation matrix $\mathbf{S}$ separately. This approach is also known as a "separation strategy" for covariance matrix priors (Barnard, McCulloch,

[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.

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 $\mathbf{S}$ 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]

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

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 $\mathbf{S}$, 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 $\mathbf{\Sigma}$. 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 $\mathbf{S}$ 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

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

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.

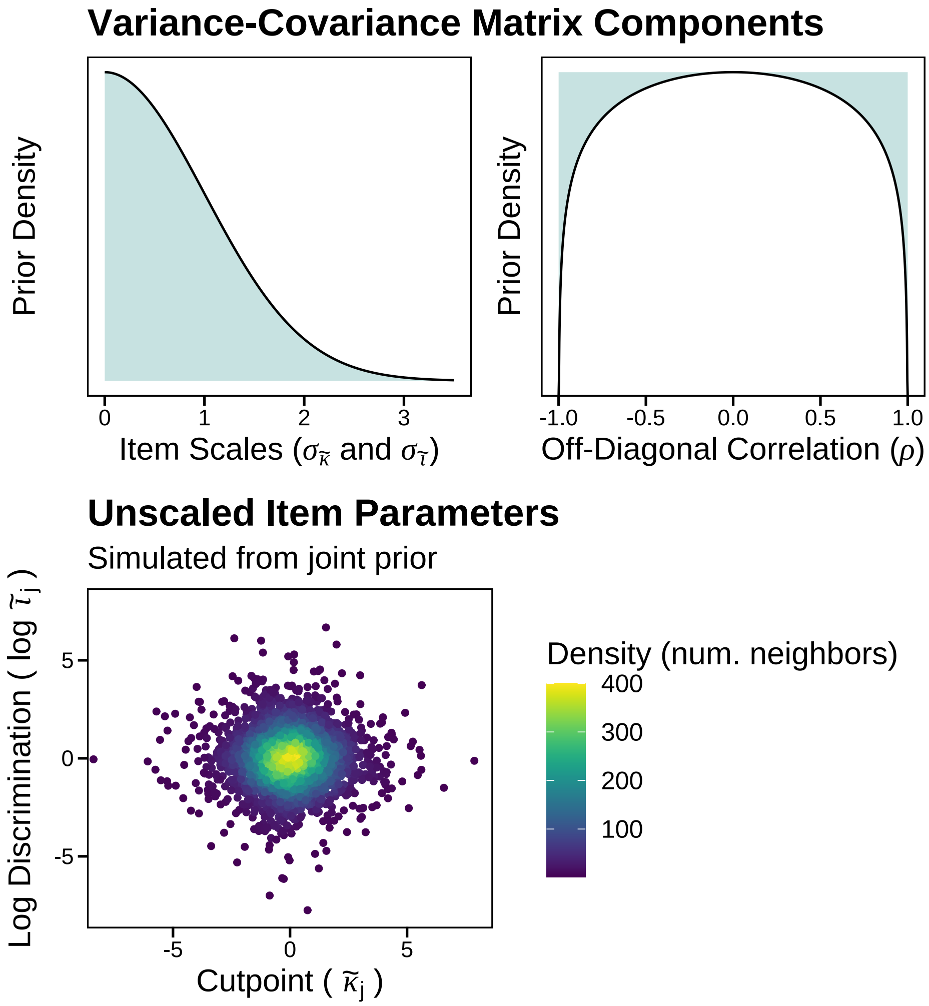

Figure 2.5: Components of the joint hierarchical prior for the unscaled item parameters. Top row shows prior densities for the parameters of the decomposed variance-covariance matrix, including the standard deviation terms (left) and the item correlation from the LKJ distribution (right). The bottom row shows prior values for unscaled items simulated from the joint prior.

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 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

Data

Posterior Analysis

— 3 —

A Bayesian Framework for Causal Inference in 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 $Y(\theta)$, we observe $Y(p(\theta))$) 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 provide a quick preview to the notation and terminology of two dominant approaches to causal inference in applied statistics: the "Rubin model" of potential outcomes and the "Pearl model" of *do*-calculus I then offer a Bayesian reinterpretation of these models, where unobserved potential outcomes/counterfactuals are characterized by probability distributions that are conditioned on the observed data This raises the daunting issue of prior specification for model parameters: 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 contribute to a causal study at different levels of abstraction.

3.1 Essential Concepts

Bayesian Inference

Conventional:

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)p(\theta)}{p(\mathbf{y})} \tag{3.1}$$

$$p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta)p(\theta) \tag{3.2}$$

Implied:

$$p(\theta \mid \mathbf{y}, \mathcal{H}) = \frac{p(\mathbf{y} \mid \theta, \mathcal{H})p(\theta \mid \mathcal{H})}{p(\mathbf{y} \mid \mathcal{H})} \tag{3.3}$$

Joint model of the system:

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

Better parameters with priors (reduce variance)

Causal Models

Intro Rubin and Pearl models.

Strictly speaking, causality is never observed. It is only inferred with the help of assumptions.

Better parameters "by design" (reduce bias)

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

Shared Goals, Different Tactics

3.2 Modeling Potential Outcomes with Posterior Distributions

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 $\tilde{y}_i$: $p(\tilde{y}_i \mid \mathbf{y}) \propto p(\tilde{y}_i \mid \theta, \mathbf{y}) p(\theta)$

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)

What "Parameters" Mean in Modeling and Estimation

What was I trying to say here?

The causal parameter is a feature of the *causal model*. The causal model is the model if you knew everything. When we impose assumptions for estimating, we can say under which conditions this parameter (or an average of the parameter in a finite sample) can be *estimated*. However, the estimator for the parameter is a different thing.

Bayesian estimation is a different *estimation* framework given the same causal model. This is easy to see when we think about MLE models as special cases of Bayesian models with all log-likelihoods equally weighted. We have parameters given by some "ideal" model, but the parameters are estimated with a pragmatic penalty term.

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 and confounders, integrating over confounders. "Statistical" assumptions define how we think we can build $E[Y]$

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

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

3.3 Exposing Hidden Assumptions

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.[1] 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.

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{2\sigma}\right)^2} = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(x-\mu)^2}{2}} \tag{3.4}$$

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

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

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \text{where } \varepsilon_i \sim \text{Normal}\left(0, \sigma\right) \tag{3.5}$$

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.

$$y_i \sim \text{Normal}\left(\alpha + \beta x_i, \sigma\right) \tag{3.6}$$

Algebraically, these two models are identical. The difference is only a matter of which component has the distributional assumption. In Equation (3.5), the distribution is assigned to ε_i, so y_i is a random

[1] 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.

variable only by way of ε_i. In Equation (3.6), we assign the distributional assumption directly to y_i, bringing the regression function into the mean rather than "factoring it out" of the distribution.

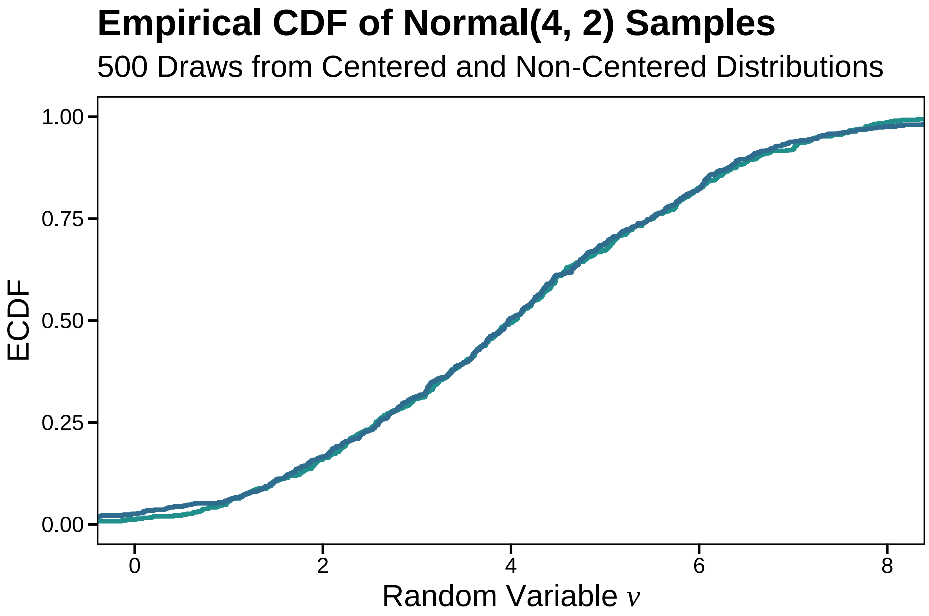

Figure 3.1: Demonstration of centered and non-centered parameterizations for a Normal distribution. The non-centered parameterization is statistically equivalent, but the location and scale are factored out of the distribution.

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.

$$\text{Centered Parameterization:} \quad v \sim \text{Normal}(4, 2) \tag{3.7}$$

$$\text{Non-Centered Parameterization:} \quad v = 4 + 2z, \qquad \text{where } z \sim \text{Normal}(0, 1) \tag{3.8}$$

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

Figure 3.1. Because the distributions are the same, the empirical CDFs are identical except for random sampling error.

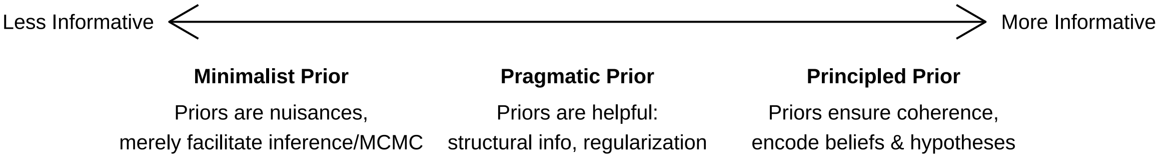

Figure 3.2: A spectrum of attitudes toward priors.

Problems of beliefs:

- No degree of belief.
- Parameterization makes this too challenging.
- Prior might change depending on what I ate for lunch.
- "Elicitation" of priors satisfying the wrong audience, or at the very least can be easily misused. We 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

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.

$$p(\mu \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mu) p(\mu) \tag{3.9}$$

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

$$p(\mu \mid \mathbf{y}) \propto \prod_{i=1}^{n} l(y_i) p(\mu) \tag{3.10}$$

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

$$\begin{aligned} \log p(\mu \mid \mathbf{y}) &\propto \sum_{i=1}^{n} \log l(y_i) + \sum_{i=1}^{n} \log p(\mu) \\ &\propto \log l(y_1) + \log p(\mu) + \log l(y_2) + \log p(\mu) + \ldots + \log l(y_n) + \log p(\mu) \end{aligned} \tag{3.11}$$

The setup in (3.11) 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 (3.11) 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 (3.11) 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.

$$\log p(\mu \mid \mathbf{y}) \propto \log l(y_1) + \log l(y_2) + \ldots + \log l(y_n) \tag{3.12}$$

If $p(\mu)$ is not flat, however, and $p(\mu)$ varies across values of μ, we can no longer ignore $p(\mu)$ in (3.11) 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 $\square$.

$$\log p(\mu \mid \mathbf{y}) \propto \log l(y_1) + \square + \log l(y_2) + \square \ldots + \log l(y_n) + \square \tag{3.13}$$

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 $p(\mu)$.

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 $\vec{\alpha}$.

$$p\left(\vec{\alpha} \mid y\right) \propto p(y \mid \vec{\alpha})p(\vec{\alpha}) \tag{3.14}$$

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

Nonlinear transformation of θ 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.

Priors on Models and "No-Free-Lunch" Theorems

Models (likelihoods) are priors

Identification assumptions are priors

Generalization to any population is a prior

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

3.4 Pragmatic Bayesian Modeling

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
- 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

Modeling Cultures in Political Science: Complexity and Agnosticism

Sidestepping priors

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

Causal don't mean nonparametric

Bayesian don't mean Complex

At any rate:

- stabilizing model-heavy designs (but priors in high dimensions are scary!)
  - that hierarchical conjoint thing
- sensitivity testing for noisy circumstances

Principled Approaches to Model Parameterization

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

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,

$$y_i = \alpha + \beta z_i + \varepsilon_i \tag{3.15}$$

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.

$$y_i = z_i \mu_1 + (1 - z_i)\, \mu_0 + \varepsilon_i \tag{3.16}$$

Although the treatment effect β from Equation (3.15) is equivalent to the difference in means $\mu_1 - \mu_0$ from Equation (3.16), the parameterization of the model affects the implied prior for the difference in means. If the researcher gives a flat prior to both $\mu_{\mathbf{z}}$ terms, the implied prior for the difference in means will not be flat. Instead, it will be triangular, as shown in Figure 3.3. The underlying mechanics of this problem are well-known in applied statistics—if we continue adding parameters, the Central Limit Theorem describes how the resulting distribution will converge to Normality—but it takes the explicit specification of priors to shine a light on the consequences of default prior choices in a particular case. In particular it shows how even flat priors, which are popularly regarded as "agnostic" priors because of their implicit connection to maximum likelihood estimators, do not necessarily imply flat priors about the researcher's key quantities of interest. Rather, flat priors can create a variety of unintended prior distributions that do not match the researcher's expectations. I return to this important idea in the discussion about setting priors for a probit model in Section 2.5.

- equivalent parameterizations ### Structural Priors and Weak Information

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

Understanding Log Prior *Shape*

This is low-key pretty big

Regularization-Induced Confounding

3.5 Bayesian Opportunities

Full Posterior Uncertainty

Multi-stage models

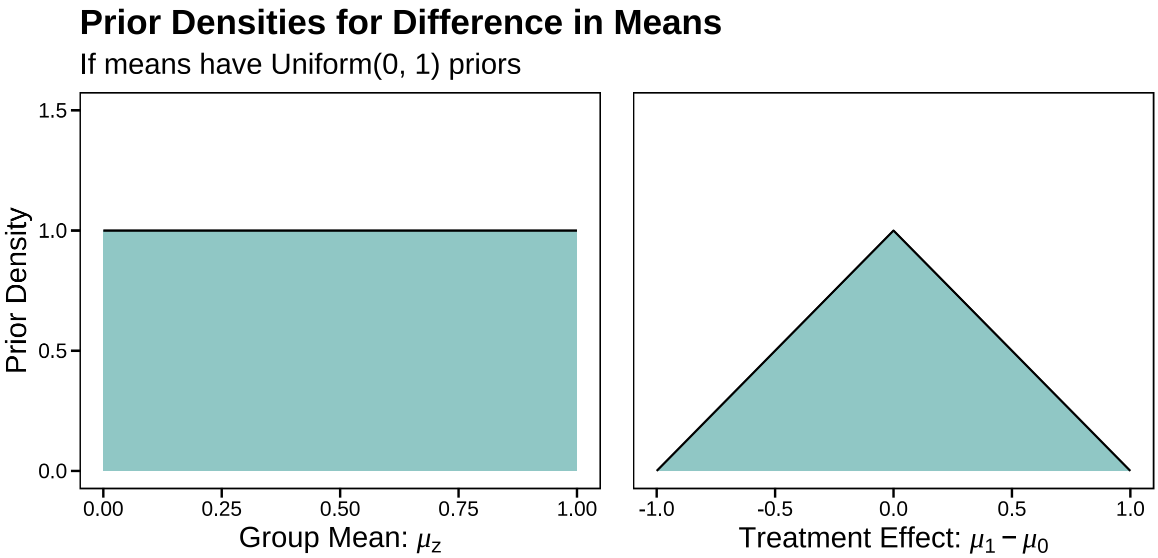

Figure 3.3: Model parameterization affects prior distributions for quantities of interest that are functions of multiple parameters or transformed parameters. The left panel shows that the difference between two means does not have a flat prior if the two means are given flat priors. Note that the x-axes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

- propensity models
- structural models
- to bootstrap or not to bootstrap?

Predictive models

Multiple comparisons and regularization

Identification Assumptions as Priors

Relatedly: priors over models

Regression Discontinuity with a Binary Outcome

Parameterization, structural information

Meta-Analysis of Nonparametric Treatment Effects

3.6 Other Frontiers of Bayesian Causal Inference

Beyond Estimation: Inferences About Models and Hypotheses

Inherit material from earlier section

Priors are the Basis for all Generalization

No-Free-Lunch theorems

Agnostic Causal Inference

Group IRT Model

References

Abramowitz, Alan I, and Kyle L Saunders. 1998. “Ideological realignment in the us electorate.” *The Journal of Politics* 60(03): 634–652.

Akinc, Deniz, and Martina Vandebroek. 2018. “Bayesian estimation of mixed logit models: Selecting an appropriate prior for the covariance matrix.” *Journal of choice modelling* 29: 133–151.

Aldrich, John H. 1983. “A downsian spatial model with party activism.” *American Political Science Review* 77(04): 974–990.

Aldrich, John H. 2011. *Why parties?: A second look*. University of Chicago Press.

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.

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, James M Snyder, and Charles Stewart. 2001. “Candidate positioning in u.s. House elections.” *American Journal of Political Science*: 136–159.

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.

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.

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.

Bürkner, Paul-Christian, and others. 2017. "Brms: An r package for bayesian multilevel models using stan." *Journal of Statistical Software* 80(1): 1–28.

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.

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.

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.

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

Fiorina, Morris P, Samuel J Abrams, and Jeremy C Pope. 2005. *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, Daniel Simpson, and Michael Betancourt. 2017. "The prior can often only be understood in the context of the likelihood." *Entropy* 19(10): 555.

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.

Green, Donald, Bradley Palmquist, and Eric Schickler. 2002. *Partisan hearts and minds*. New Haven, CT: Yale University Press.

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

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, 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.

Kernell, Georgia. 2009. “Giving order to districts: Estimating voter distributions with national election returns.” *Political Analysis* 17(3): 215–235.

Key, V.O. Jr. 1949. “Southern politics in state and nation.”

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

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

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).

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.

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

Masket, Seth. 2009. *No middle ground: How informal party organizations control nominations and polarize legislatures*. University of Michigan Press.

McGhee, Eric et al. 2014. "A primary cause of partisanship? Nomination systems and legislator ideology." *American Journal of Political Science* 58(2): 337–351.

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

Papaspiliopoulos, Omiros, Gareth O Roberts, and Martin Sköld. 2007. "A general framework for the parametrization of hierarchical models." *Statistical Science*: 59–73.

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).

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.

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.

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