Do Primaries Work?

Michael G. DeCrescenzo

DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy
Political Science
University of Wisconsin–Madison
2020

ORAL DEFENSE APPROVED TBD

BARRY C. BURDEN (CHAIR), POLITICAL SCIENCE
KENNETH R. MAYER, POLITICAL SCIENCE
ELEANOR NEFF POWELL, POLITICAL SCIENCE
ALEXANDER M. TAHK, POLITICAL SCIENCE
MICHAEL W. WAGNER, JOURNALISM AND MASS COMM.

for Tina:

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

Abstract

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

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

as candidate valence and campaign spending may be more powerful than in general elections. For these reasons, the proposition that primary elections advance the ideological interest of local partisan voters is theoretically contestable.

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

Contents

List of Tables

List of Figures

Acknowledgments

Many people supported this me through this project…

- My family
- My committee
- Department and related institutions: ERC, Straus, Deb, Beasts, Prospectus course fall 2016
- Faculty and students at other universities who shared data or provided advice: Devin Caughey, David Doherty, Seth Hill, Georgia Kernell, Shiro Kuriwaku, Jacob Montgomery, Rachel Porter, Andrew Reeves, Michael Ting, and Sarah Treul.
- People who provided feedback in workshops: Devin Judge-Lord, Evan Morier, Rochelle Snyder, Blake Reynolds, David Canon, Marcy Shieh
- Software developers: Garrick Aiden-Bouie (sp), Matthew Kay
- Friends in Madison: Shaan Amin, Hannah Chapman, Josh Cruz, Micah Dillard, Jordan Hsu, Rachel Jacobs, Hari Jost, Amy Kawleski, Anna Meier, Erin Nelson, Anna Oltman, Rachel Schwartz, Erin Zwick

1

Introduction: Policy Ideology and Congressional Primaries

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

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

than they did in the past (Abramowitz and Saunders 1998; Fiorina, Abrams, and Pope 2005a; Layman and Carsey 2002; Levendusky 2009).

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

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

When the choice before voters in the general election does not present realistic alternatives, political scientists naturally shift their focus to the nomination of partisan candidates. V.O. Key, for example, studied Democratic Party dominance in the American South, asking if competition within the party could provide a quality of representation similar to two-party competition (Key 1949). Although scholars are right to examine within-party competition, focusing on contexts of single-party dominance is a serious limitation. Even in races between viable candidates from both major parties, within-party competition plays a crucial role simply due to the fact that par-

tisan voters almost certainly cast a vote for their own party. Rank-and-file partisan constituents are all but captured. If they are to express their policy preferences through the act of voting, their voices may register as relatively weak because they present little electoral risk to their party in the general election. The nomination stage—the primary election in particular—remains an important venue for the representation of partisans' policy views, whether the general election is closely contested or not.

1.1 Policy Preferences and the Strategic Positioning Dilemma

This dissertation is chiefly concerned with the policy preferences of partisan voters and their role in electoral representation through Congressional primary elections. The study of American electoral politics has not ignored the representational function of primary elections (Aldrich 2011; Cohen et al. 2009; Geer 1988; Norrander 1989; Sides et al. 2018), but as I discuss below, the quantifiable impact of primary voters' policy preferences in government is a startlingly open question. Several existing studies have examined other aspects of representation through House primaries, such as the introduction of the direct primary (Ansolabehere et al. 2010), how candidates position themselves in response to the presence or threat of primary challenges (Brady, Han, and Pope 2007; Burden 2004; Hirano et al. 2010), and how primary nomination rules affect elite polarization (Hirano et al. 2010; McGhee et al. 2014; Rogowski and Langella 2015). Though these studies address interesting aspects of electoral representation and party competition, they cannot speak directly to the influence of voter's policy preferences on (1) the positioning of House primary candidates and (2) the outcomes of House primary elections.

The absence of voter preferences from the empirical study of primaries is troubling because they play a crucial role in the dominant theory that relates representation to primary politics. Although the Downsian model of candidate positioning explains the incentives for candidates to stake out moderate policy positions to cater to the ideological "median voter" (Downs 1957), candidates behave differently in the real world. Instead, candidates engage in highly partisan behavior and take divergent issue stances even on salient local issues and in closely competitive districts (Ansolabehere, Snyder, and Stewart 2001; Fowler and Hall 2016). But why? Scholars and political observers have argued that because competing in the general election requires each

candidate to clinch their party's nomination contest, these candidates face a combination of convergence-promoting and divergence-promoting incentives. Primary elections tend to be dominated by partisan voters who are more attentive to politics, hold more non-centrist issue preferences, and "weight" candidates' issue positions more heavily than the average voter in the general election.[1] As a result, the risk that a candidate is defeated in the primary for being too moderate may outweigh the risk of losing the general election for being too partisan. The conflicting incentives imposed by partisan constituency and the general election constituency creates a "strategic-positioning dilemma" that leads candidates to take divergent issue stances rather than targeting a district median voter (Aldrich 1983; Brady, Han, and Pope 2007; Burden 2001; Hill 2015).

The strategic positioning dilemma (SPD) is a central theoretical feature of this project, and tests of the SPD are key empirical contributions in the following chapters. The sections that follow introduce key terms for understanding my critique of the existing research and my contribution to it in this project.

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

1.1.1 Key concept: policy ideology

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

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

greater number of progressive preferences about policy are more ideologically progressive, and vice versa for voters with more conservative policy preferences. Voters who hold a mixture of progressive and conservative beliefs are ideologically moderate.

Policy ideology is different from policy *mood*, since mood measures voter preferences for the government to do more or less than an ever-shifting baseline, while ideology meant to be directly comparable using only issue information (Enns and Koch 2013; McGann 2014; Stimson 1991). Policy ideology is thus a similar concept to any method that measures a hidden ideological summary from one-off issue-based stimuli. This includes ideal point scores for members of Congress, Supreme Court justices, and even individual citizens (Clinton, Jackman, and Rivers 2004; Martin and Quinn 2002; Poole and Rosenthal 1997; Tausanovitch and Warshaw 2013; Treier and Hillygus 2009). Other researchers have called this concept "policy liberalism" (Caughey and Warshaw 2015), which orients the concept so that "larger" values represent "more liberalism." For this project, I prefer to orient the construct as policy *conservatism*, which orients a scale so that larger/more conservative values correspond to "rightward" movements on a number line. I try to be conscious of the difference between *consistent* issue beliefs and *extreme* issue beliefs throughout this project. Consistently conservative issue beliefs do not necessarily imply that an actor is "extremely" conservative (Fiorina, Abrams, and Pope 2005b), and an actor may appear "moderate" even if they hold a mixture of non-moderate progressive and conservative issue beliefs (Broockman 2016).

This project views policy ideology in a measurement modeling context, which we return to in Chapter 2. Policy ideology affects voters' issue beliefs, and while issue beliefs can be measured using a survey, policy ideology itself is not observable. Instead, policy ideology exists in a latent space, an survey items on specific issues reveal only limited information about voters' locations in the latent space. This is different from summarizing policy views by adding or averaging policy responses, which implicitly assumes that all items about all issues are equally informative about ideology. Modern measurement approaches relax this assumption, instead viewing survey items as sources of correlated measurement error across respondents, leading to more careful modeling approaches for estimating a latent signal from noisy survey data

(Ansolabehere, Rodden, and Snyder 2008). Following this modeling tradition, I refer to an individual's location in policy-ideological space as their "ideal point," the point at which their expected utility of a policy is maximized with respect to their ideological preferences.

1.1.2 Key concept: district-party groups

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

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

One important distinction about district-party groups is that they are made of constituents, not organizations. For this reason, it is sometimes helpful to refer to district-party groups as district party "publics," which

emphasizes that the groups are composed of ordinary citizens (Caughey and Warshaw 2018). There is no formal registration requirement to be a member of a party group, only a partisan identification. This construction of district-party publics aligns most closely with Key's "party in the electorate" rather than "party as organization" (Key 1955). This distinguishes party publics from interest groups, policy groups, "intense policy demanders," or the "extended party network," which are concepts that describe organizations or maneuvers by political elites rather than rank-and-file constituents (Cohen et al. 2009; Koger, Masket, and Noel 2009; Masket 2009). Although recent research has underscored the importance of elite actors in shaping party nominations, this project focuses specifically on testing the SPD, which is a voter-centric view of primary representation. We bring in important concepts from elite-driven stories of primaries as they apply to particular claims being tested in later chapters.

1.1.3 Key concept: district-party ideology

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

1.1.4 Key concept: candidate campaign positioning

As with individual voters, we can imagine that candidates for Congress have campaign platforms, or at least promises and stated issue positions, that are located in ideological space as well. The study of United States politics most commonly places elite political actors in ideological space using their voting

records, including members of Congress, Supreme Court justices, federal judges, and state legislators (Clinton, Jackman, and Rivers 2004; Epstein et al. 2007; Martin and Quinn 2002; Poole and Rosenthal 1997; Shor and McCarty 2011). Researchers have extended the modeling intuitions to estimate ideal points from unconventional sources of data, including surveys of congressional candidates, campaign finance transactions, interest group ratings, text from political advertisements, and even Twitter activity (Ansolabehere, Snyder, and Stewart 2001; Barberá 2015; Bonica 2013; Burden 2004; Burden, Caldeira, and Groseclose 2000; Henderson 2016).

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

This project measures primary candidates' campaign positioning using CFscores from Bonica's (2019b) *Database on Ideology, Money in Politics, and Elections* (DIME) database. CFscores use campaign contributions to measure the political ideologies of contributors and recipients of campaign contributions, including candidates for office, party organizations, PACs, and individual donors. The estimation method assumes that a donor makes financial contributions to political actors to maximize their utility over all potential contribution choices, which is affected by the ideological similarity between donors and potential recipients (Bonica 2013, 2014). These scores have been used in other studies of primary candidate ideology by Thomsen (2014), Thomsen (2020), Rogowski and Langella (2015), Ahler, Citrin, and

Lenz (2016), and Porter and Treul (2020), and similar donation-based ideal point measures by Hall and Snyder (2015) have been used by Hall (2015) and Hall and Thompson (2018). As I discuss in future chapters, CFscore are not without controversy as indicators of elite ideology, especially when comparing members of the same party (Hill and Huber 2017; Tausanovitch and Warshaw 2017), but other research shows that donors differentiate moderate and ideological candidates within the same party (Barber, Canes-Wrone, and Thrower 2016), the ideology component of CFscores outperforms a party-only model of giving (Bonica 2014), and CFscores predict future votes by members of Congress to a similar degree of accuracy as roll-call based scores do (Bonica 2019a).

1.1.5 The strategic positioning dilemma, implications, and research questions

Now that we have defined some key terms, we can see how they relate to previous research on the strategic positioning dilemma. The theory states that candidates balance two competing constituencies during their campaign for office. Candidates face incentives to cater to the median voter in the general election, but they do not progress to the general election without first catering to partisan voters in the primary election. As a result, their campaign position is tailored to split the difference between the two constituencies, perhaps leaning more to the partisan base in safe districts and to the median voter in competitive districts. This section unpacks this intuition in detail and argues that existing research does not test the key claims.

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

Second, if candidates anticipate partisan voters' policy views and position themselves accordingly, this suggests that candidates believe partisan voters

are capable of voting in accordance with their policy views. If this is true, we should expect that district-party groups that are more conservative should be more likely to nominate conservative nominees in primary elections.

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

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

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

1.2 Does the Strategic Positioning Dilemma Describe Primary Representation?

1.2.1 Theoretical concerns

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

over candidate platforms.

The notion of the SPD emerges from a clash between idealized candidate positioning in formal models and the candidate positioning we observe in the real world. Classic formal models highlight a strategic logic for candidates to position themselves by "converging" to the location of the median voter: if constituents vote primarily with policy-based or ideological considerations, then candidates maximize the probability of electoral victory by positioning themselves as closely to the median constituent as possible (Black 1948; Downs 1957).[2]

Empirical work finds evidence in partial support of both convergent and divergent candidate incentives. Candidates who run in electorally competitive districts are more moderate than co-partisans who are running in districts that run in electorally "safe" districts (Ansolabehere, Snyder, and Stewart 2001; Burden 2004), and even candidates who run in safe districts are marginally rewarded for taking more moderate issue positions than a typical party member would (Canes-Wrone, Brady, and Cogan 2002). Extremist candidates, meanwhile, earn fewer votes and are less likely to win in Congressional elections, and this tendency is stronger in competitive districts than in safe districts (Hall 2015). Despite these incentives to take moderate campaign positions, candidates nonetheless take divergent rather than convergent stances by and large. Republican and Democratic members of Congress vote very differently from one another, and this partisan divergence increased in recent years (McCarty and Poole 2006; Poole and Rosenthal 1997). The difference in legislative voting behavior across parties isn't simply because Republicans and Democrats represent different districts, since Republicans and Democrats who represent similar districts (or the same state, in the case of U.S. Senators) nonetheless vote differently from one another (Brunell 2006; Brunell, Grofman, and Merrill 2016; McCarty, Poole, and Rosenthal 2009). Even among Congressional races in the exact same district, there is a sizable gap between Republican and Democratic candidate positions (Rogowski and Langella 2015). And although qualitative evidence from decades past suggests that candidates take careful positions on issues of local concern (Fenno 1978), more recent systematic tests find mixed evidence of localized, particularistic position-taking. (Canes-Wrone, Minozzi, and Reveley 2011; Fowler and Hall 2016). In total, even though there is some evidence that candidates benefit

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

by positioning themselves as marginally more moderate or more in line with local public opinion, the dominant finding is that candidates take divergent positions that are more closely aligned with a national party platform than with a set of local issue priorities.

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

These incentives for candidates to diverge from median positions are possible without considering primary elections whatsoever. Even if we introduce primary elections into the theoretical story, many plausible explanations for divergence do not rely on outward pressures from ideological primary voters either. Many scholars of political parties maintain that parties retained their gatekeeping roles over party nominations even as the direct primary ostensibly removed their formal powers over candidate selection. Although primary campaigns take place, these scholars argue that an informal network of party

actors wields enormous influence behind the scenes, controlling which candidates obtain access to the party's resources, donor lists, and partisan campaign labor (Cohen et al. 2009; Masket 2009). Through these mechanisms, candidates can live or die by the nomination process long before primary *voters* ever enter the picture.

One reason to doubt the SPD on theoretical grounds is that it has high demands of voter sophistication in primary elections. It is well understood that learning about the characteristics and issue positions of political candidates is costly for voters, particularly in non-presidential elections. Party labels on the ballot are valuable heuristics for voters to differentiate the issue positions of Republican and Democratic candidates likely hold (Hill 2015). Primary elections, however, occur most of the time between candidates in the same party,[3] which denies voters' the informational shortcut of a candidate's party affiliation (Norrander 1989). Primary elections often occur during months when voters are paying less attention to politics, and the press cover primary campaigns less closely than general election campaigns. Primary voters have a reputation for being more attentive and sophisticated consumers of political information, but in these lower-information environments, they may cast their ballots for non-policy reasons by prioritizing "Washington outsiders" or identity-based candidate features such as gender or race (Porter and Treul 2020; Thomsen 2020). They may also vote for the familiar candidate instead of the ideologically proximate one, in which case asymmetric campaign expenditures or news coverage may advantage one candidate over the other. For example, Bonica (2020) attributes lawyers' numerical prominence in Congress to their ability to raise early money from their wealthy social networks. Furthermore, despite the disproportionate news coverage received by primary candidates who challenge incumbents on ideological grounds, the absolute number of explicitly ideological primary challenges in a given election cycle is low (Boatright 2013), so primary voters are unlikely to experience a deluge of policy-focused campaign messages even if they are attentive and sophisticated to receive and process those messages. In short, the claim that voters' policy preferences affect their choices in primary campaigns sounds straightforward, but the information environment of primary campaigns makes it difficult for constituents to vote foremost with their policy ideologies.

The SPD also requires candidates to perceive the policy ideologies of their

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

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

1.2.2 Empirical ambiguity

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

Much of the interest in primary elections and representation comes from a focus on candidate divergence and partisan polarization. Why do candidates who stand for general election take divergent stances from one another, and do the competitive dynamics of primary elections increase this divergence? Prominent studies of candidate positioning in general elections initially found conflicting evidence about the influence of stiff primary competition on candidate extremity. Using survey data from congressional candidates during the 2000 campaign, Burden (2004) finds that general election candidates take more extreme policy positions in their campaigns if they also faced stronger primary competition. This makes sense especially if primary candidates care more about the candidate's ideological positioning than general election voters do, the latter of whom are also receptive to non-policy appeals. Ansolabehere, Snyder, and Stewart (2001) find the reverse pattern using 1996 survey data. The gap between major party candidates was actually smaller when one of the candidates faced stiffer primary competition. This counter-intuitive finding makes sense if the presence of a primary challenger is itself a consequence of candidate positioning. If an incumbent maintains a partisan reputation, this may fend off credible primary challengers who have less room to wage an ideological campaign against the incumbent. As a result, the *threat* of a primary challenge exerts a centrifugal force on candidate

positioning, even if a primary challenger never actually appears (Hacker, Pierson, and others 2005). Hirano et al. (2010) study this threat-based hypothesis by measuring potential primary threat as the average presence of primary competitors in down-ballot races. In district with high levels of latent primary threat, we might expect the incumbent to take more extreme stances in Congress. Although the idea that incumbents vote as party faithfuls to preempt opportunistic challengers is intuitive and supported by other research (e.g. Mann 1978), this measure was not meaningfully related to the extremity of an incumbent's voting record in Congress (Hirano et al. 2010). In short, the evidence of the polarizing effects of primary challenges is mixed and unclear.

Researchers interested in the polarizing effects of primaries on candidates and legislators has also examined primary "rules." Political parties are private organizations, and nominees are intended to represent the parties' priorities and governing values, but participation in primary elections is not always restricted to party members only. Primary "openness" rules that govern who can participate in a partisan primary are managed by state election law, with some allowances for parties to set rules within those limits. States with "closed" primaries restrict participation in primaries only to individuals who are registered as Republicans or Democrats in their state registration records. States that allow third-party or non-partisan voters to participate in partisan primaries are "partially" open, and states where any voter can participate in any primary are regarded as "open" primaries. I discuss finer details of primary rules in later chapters. Researchers seeking to exploit state-level variation in primary rules hypothesize that states with more restrictive participation criteria might select more ideologically extreme primary nominees, and states with more relaxed rules might select relatively moderate nominees. This is because primary voters are commonly believed to hold more ideologically consistent policy views than other constituents, so candidate polarization will respond to the polarization among the voting public (Jacobson 2012). However, the consensus among recent studies finds little evidence supporting the hypothesis that primary rules affect polarization in congress or candidate divergence more broadly. This is because there is little consensus in public opinion research that partisans who participate in primaries are much different from partisans who do not participate in primaries, either demographically or ideologically (Geer 1988; Hill 2015; Jacobson 2012; Norrander 1989; Sides et

al. 2018), though these studies cover many years, and the dynamics of primary voting might have changed. And even recent studies that find that primary voters hold more ideologically consistent views find no evidence that closed primaries nominate candidates that are more ideologically off-center (Hill 2015). This finding appears to hold for the House, Senate, and state legislatures through the past several decades (Hirano et al. 2010; McGhee et al. 2014; Rogowski and Langella 2015). Even reforms that drastically change the primary rules, such as California's recent shift to a blanket primary where candidates in all parties compete for the same limited number of positions on the general election ballot, do not nominate legislators whose voting records are much more moderate than before (Bullock and Clinton 2011).

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

1.2.3 Vote shares do not identify policy ideology

Another important drawback of the existing research on primaries and ideological representation is the way these studies handle voters' policy preferences. The strategic positioning dilemma pits two constituencies in a district against each other: the nominating constituency (district-party group) that contains constituents from one party's base, and the general election constituency that contains constituents from both major parties and with no party affiliation. The former is theorized to prefer ideologically faithful candi-

dates who adhere closely to a partisan policy platform, while the latter prefers moderate candidates in the general election. Studies routinely acknowledge this distinction in theory, but they often abandon the distinction between the two groups in applied studies, instead operationalizing the preferences of all three constituencies—the general constituency and two partisan primary constituencies—using the same measure: the district-level presidential vote.

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

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

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

The panels on the right side of Figure 1.1 show how slight changes to one party's preference distribution affects the aggregate distribution of preferences in the combined constituency and, as a result, the presidential vote share in the district. The composite distribution is again shown in gray, with dark and light shades indicating vote choice as in the bottom-left panel. The underlying partisan distributions are outlined only with red and blue lines to reduce visual clutter. The modifications to the underlying partisan preferences are simple, but even these simple changes reveal the fundamental problem with using district voting as a proxy for policy ideology in the voting population. In each panel, I intervene on only one feature of the Democratic Party ideal point distribution, leaving the Republican distribution untouched (median of 1, standard deviation of 1). Intervening on just one component of one party's distribution is meant to keep the demonstration simple, bearing in mind that the problem is much more complex in the real world, where we can imag-

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

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

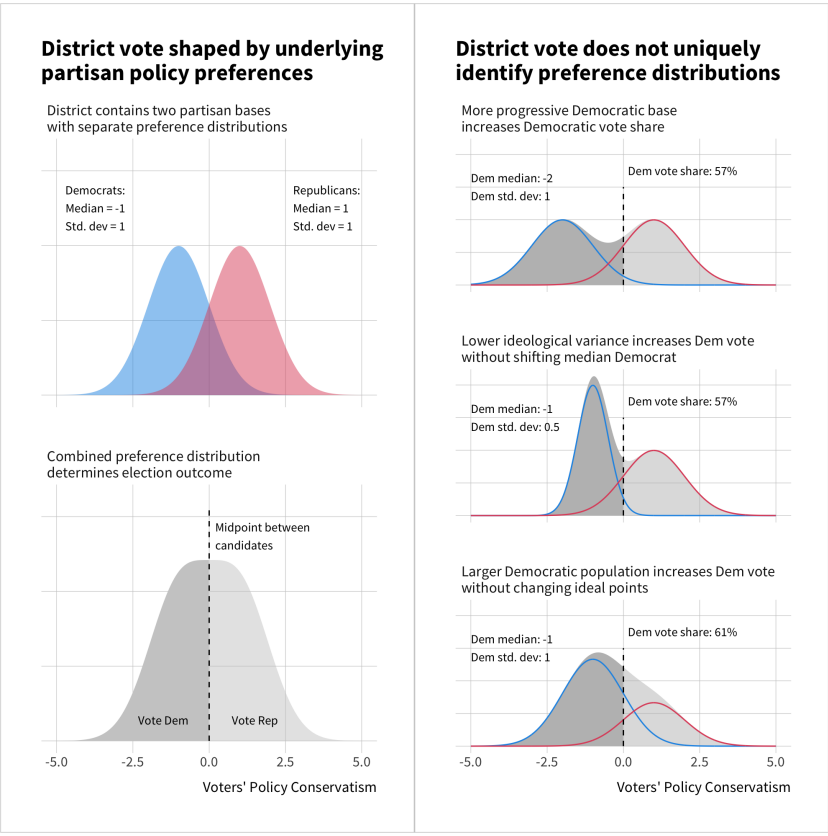

Figure 1.1: Demonstrating how district vote shares from a single election are insufficient to identify underlying policy-ideological features of the district. The left side shows how the policy preference distributions for two parties in a district (top panel) combine to form an aggregate preference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

ine multiple simultaneous changes to both parties at once. The interventions highlight two classes of problems. First, we can perform multiple modifications of the underlying partisan distribution that obtain the same aggregate vote share. This proves that the district vote does not uniquely identify the characteristics of the underlying voter distributions. And second, we can alter the district vote outcome by changing party *sizes* without any change to the ideal point distribution within either party. This proves that vote shares may vary across districts even if partisan ideal points distributions are the same.

In the top-right panel, I shift the location of the Democratic ideal point distribution to the left, from a median of -1 to -2. This location shift results in a greater number of Democratic constituents with ideal points left of the candidate midpoint, increasing the Democratic vote share in the district from 50% to 57%. In the middle-right panel, I shrink the scale of the Democratic ideal point distribution from a standard deviation of 1 to a standard deviation of 0.5. Lower ideal point variance within the Democratic base has the exact same effect on the vote as shifting the location: more Democratic voters left of the midpoint, which increases the Democratic vote share to 57%. This means that compared to a district with a 50% presidential vote split, we would not be able to attribute the increased Democratic vote to a constituency that is *more progressive on average* (location) or simply *less heterogeneous* in its policy preferences (scale). The bottom-right panel in the figure shows how we obtain a different district vote without changing the underlying ideological distribution in either party whatsoever, instead changing only the relative population size of each partisan base. The Democratic base in the final panel is unchanged compare to the original distribution laid out in the top-left: median of -1 and standard deviation of 1. The only difference is that the district contains an unequal balance of partisan voters, two Democratic constituents to every one Republican constituent. This results in an increased Democratic vote from 50% to 61%—ironically, the largest impact on the overall district vote despite not changing the ideological distribution of either party.

To review the lessons of Figure 1.1, observing a Democratic vote share greater than 50% reveals very little about the underlying distribution of voters. In every panel, we observe an increase in the Democratic vote compared to our baseline scenario, but the the median voter in either party does not need to change in order for vote shares to be affected. Since the Republican

distribution is identical in every panel, inferring that Republicans are less conservative in districts with greater Democratic voting would be incorrect in every case. For the Democratic constituents, inferring a more progressive Democratic median voter from greater Democratic voting would be wrong in two of the three cases.

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

The conceptual difference between district vote shares and aggregate ideology appears in real data as well, as shown in Figure 1.2. The figure shows ideological self-placement responses to the Cooperative Congressional Election Study (CCES) as an approximate measure of a citizen policy ideology. I calculate the average self-placement for all respondents in each congressional district, as well as the average self-placement of Republican and Democratic identifiers as separate subgroups within each district. The first two panels use 2018 data to show that the district vote captures variation in ideological self-placement reasonably well when examining congressional districts as a whole, but it does a poorer job capturing variation in self-placement within each party. The first panel shows that districts that voted more strongly for Democratic presidential candidate in 2016 were more liberal on average, and districts that voted more strongly for the Republican candidate in were more conservative, indicated by a positively sloped loess fit line. The middle panel shows that this pattern does not hold as strongly within parties. Among Republican identifiers within each district, a weaker but still positive relationship holds overall, with more conservative Republicans in districts that voted more Republican. Among Democratic identifiers, however, ideological self-placement is not as strongly related to aggregate voting, with a loess fit that is flatter and even negative at several points. The final panel of loess fits

is included to show that this pattern appears in all CCES years and is not particular to 2018 CCES responses: a strong relationship between vote shares and self-placement *on average*, and weak or non-relationships within each party.

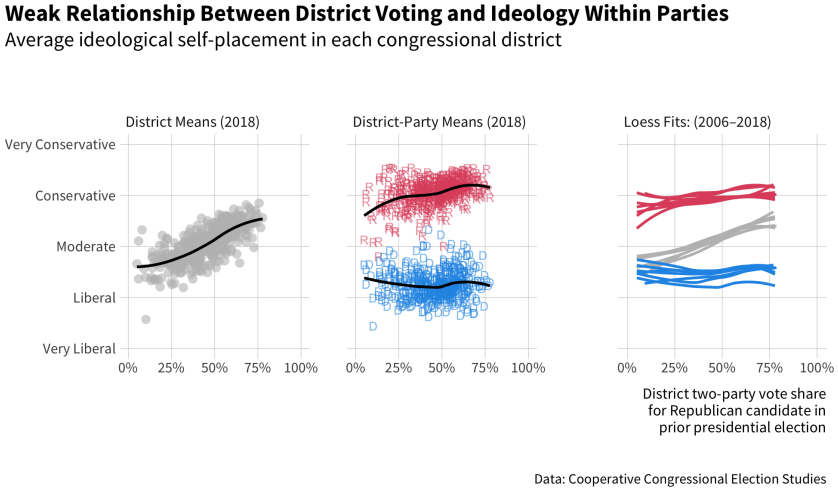

Figure 1.2: Average ideological self placement (vertical axis) and Republican vote share (horizontal axis) in all 435 congressional districts. Mean self-placement is calculated by numerically coding CCES ideological self-placement responses before averaging. The first panel plots average self-placement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

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

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

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

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

1.3 Project Outline and Contributions

1.3.1 Measuring district-party ideology

This chapter has so far identified a shortcoming in the study of primaries that subconstituency preferences are rarely measured. This project rectifies this shortcoming by measuring district-party ideology for Republican and Democratic party groups in Chapter 2. This allows the project to carry out direct tests of SPD hypotheses that were previously impossible in Chapters 4 and 5.

I estimate district-party ideology this using an item response theory (IRT)

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

1.3.2 Empirical tests: how district-party ideology matters

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

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

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

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

rules in Chapters 4 and 5 will provide a more faithful test of the primary rules hypothesis.

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

1.3.3 Causal inference with structural models

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

One inherent limitation of the district-party ideology estimates is that they come from a measurement model. The measurement model smooths estimates with a hierarchical regression, where partial pooling improves the estimate for one unit "borrowing information" from other units. This shrinks estimates toward one another, imposing correlations between estimates that share a common cause. To leverage exogenous variation for design-based causal inference, this variation would likely have to come predominantly through exogenous shocks to raw survey data, which is challenging to conceive of considering that many surveys must be pooled to achieve feasible estimates at the district-party level.

Given these data limitations, this project turns to causal identification through a conditional independence assumption (Rubin 2005), also known as "selection on observables." Although selection on observables is a common approach to quantitative research, many analyses are not careful about their modeling choices, controlling for variables that do not improve causal identification or using modeling approaches that impose fragile or implausible functional assumptions on the data. One guiding ethic for the methodological

contributions in this project is to take observational causal modeling more seriously than the existing research on primary representation by setting up empirical analyses that aspire to do the following:

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

I hope to satisfy these aims by invoking more explicit causal models of potential outcomes (Rubin 2005) and using "structural causal models" (SCMs) to guide model specification choices (e.g. Pearl 1995). The SCM approach makes heavy use of causal diagrams, or "directed acyclic graphs" (DAGs), to visualize a causal structure and identify causal claims. Causal diagrams as heuristic devices for causal inference are not new to political science in general (Gerring 2001), but combining causal diagrams with the formal exactitude of the current causal inference tradition is less common in political science. Furthermore, SCMs and causal diagrams are less common in the literature on primaries and representation, which has not been as explicit about causal assumptions and empirical designs, with some notable exceptions (Fowler and Hall 2016; Hall 2015).

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

this bias using Neyman orthogonalization, a two-stage modeling approach that de-biases causal estimates by reparameterizing the structural causal model (Hahn et al. 2018). Regularization-induced confounding is a serious problem for high-dimensional causal inference, but it has been discussed almost nowhere in political science (Ratkovic 2019).

Selection on observables is a fragile assumption for causal identification, which leads many researchers to speak in "scientific euphemisms" about causality instead of invoking explicit causal language (Hernán 2018). I adopt the position that this "taboo against explicit causal inference" is harmful to the larger aims of a research program because it obscures the dependence of research findings on causal assumptions, whose transparency is essential for credible causal inference, and leads work to be misinterpreted by future audiences who tend to interpret findings as causal regardless of author intent (Grosz, Rohrer, and Thoemmes 2020). No study will ever prove the existence of a causal effect. Researchers should be transparent about causal assumptions so that future readers and researchers have clearer ideas about how to improve previous work. As such, this work will invoke causal language, highlight identification assumptions, and discuss threats to identification assumptions openly.

1.3.4 Bayesian causal modeling

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

Although the Bayesian view of causal inference is not new (Rubin 1978), it appears almost nowhere in political science. Political scientists occasionally use Bayesian technology for analytical or computational convenience (e.g.

Horiuchi, Imai, and Taniguchi 2007; Carlson 2020; Ornstein and Duck-Mayr 2020; Ratkovic and Tingley 2017), but rarely are the epistemic contours of Bayesian analysis explicitly credited for adding value to a causal analysis (Green et al. 2016; in economics, see Meager 2019).

Chapter 3 explores a Bayesian approach to causal inference in political science at length. It lays out a probabilistic model of potential outcomes adapted from Rubin (1978) and discusses how to interpret causal inference research designs through a Bayesian updating framework. I give pragmatic guidance for thinking about priors and specifying Bayesian causal models, and I demonstrate the modeling approaching by replicating and extending a few published analyses in political science, noting where the Bayesian approach leads to different conclusions and interpretations about the findings.

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

1.3.5 Bayesian best practices

Another important contribution of the modeling exercise is the detailed discussion of Bayesian modeling and computational implementation it contains. Classic Bayesian texts for political and social sciences are written for an outdated computational landscape where Metropolis-Hastings and Gibbs sampling algorithms were state-of-the-art estimation approaches (Gill 2014; Jackman 2009). Recent years have seen rapid progress in the development and understanding of Hamiltonian Monte Carlo algorithms, which are faster, more statistically reliable, and easier to diagnose (Betancourt 2017, 2019; Duane et al. 1987; Neal 2012), but they also require renewed attention to the way researchers specify and implement Bayesian models (Betancourt and Girolami 2015; Bürkner and others 2017; Carpenter et al. 2016). Furthermore, this new generation of applied Bayesian modeling has updated best prac-

tices for specifying priors, modeling workflow, and model evaluation that (to my knowledge) have no precedent in the current political science awareness (Betancourt 2018; Gabry et al. 2019; Gelman, Simpson, and Betancourt 2017; Lewandowski, Kurowicka, and Joe 2009; Vehtari, Gelman, and Gabry 2017; Vehtari et al. 2020). One contribution of this project is to highlight the evolving landscape for Bayesian thinking and Bayesian workflow, which has not received its due attention as a new generation of political scientists explores Bayesian analysis.

2

Hierarchical IRT Model for District-Party Ideology

2.1 Testing the Model with Simulated Data

2.2 Data Sources

Describe data

2.3 Model Results

The model was estimated using a remote server at the University of Wisconsin–Madison.[1] I generated posterior samples using MCMC on 5 Markov chains. Each chain was run for 2,000 iterations, divided into 1,000 warmup iterations to tune Stan's adaptive HMC algorithm and 1,000 post-warmup iterations saved for analysis.[2] Following the advice of Link and Eaton (2011), I stored every post-warmup sample with no thinning of chains, resulting in a total of 5,000 samples per parameter across all chains.[3]

Here we can reference Figure 2.1.

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

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

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

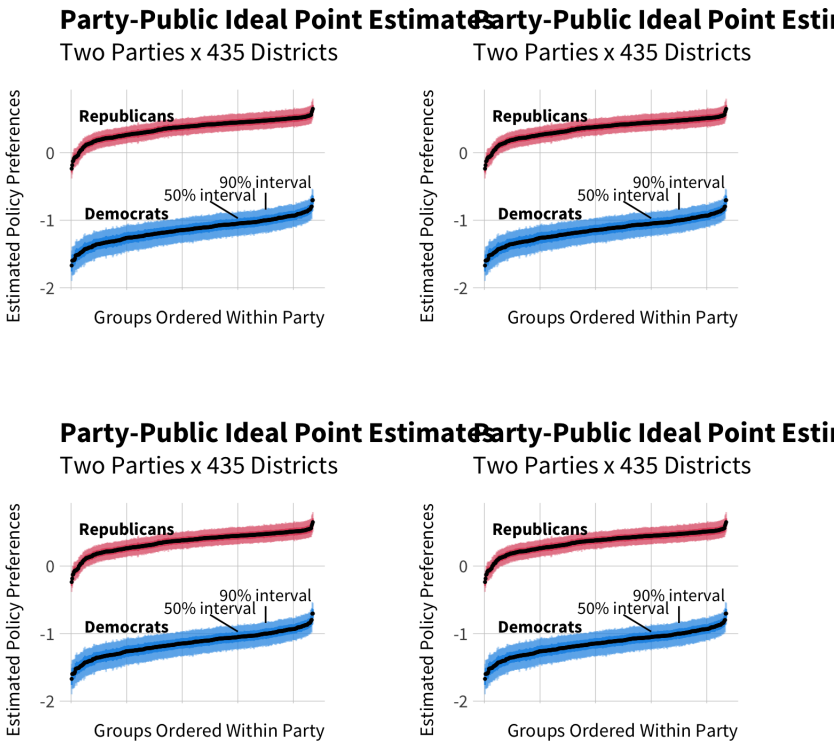

Figure 2.1: Posteriors

3

Bayesian Causal Models for Political Science

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

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

Second, the Bayesian approach incorporates uncertainties about the key independent variable in this project. Causal estimands (to use the Rubin terminology) are comparisons of potential outcomes are two hypothetical values of a treatment, usually a unit's outcome under an *observed* treatment versus an *unobserved* treatment. The data in this project frustrate the typical structure of a causal estimand because the treatment value of interest—policy ideology in a districit-party public—is not observed. Instead, it is estimated up to a probability distribution specified by the measurement model in Chapter 2. Uncertainty about the effect of setting district-party ideology to some new value θ' therefore contains multiple sources of uncertainty: statistical uncer-

tainty about the estimated causal effect, and measurement uncertainty about the original value of θ before a causal intervention. Bayesian analysis provides the statistical machinery to quantify uncertainty in these causal effects as if they were any other posterior quantity, by marginalizing the posterior distribution over any nuisance parameters.

This chapter unpacks these issues according to the following outline. First, I review the notation and terminology for causal modeling in empirical research, where data and causal estimands are posed in terms of "potential outcomes" or "counterfactual" observations. I then describe a Bayesian reinterpretation of these models, which uses probability distributions to quantify uncertainty about causal effects and counterfactual data, conditional on observed data. Because Bayesian modeling remains largely foreign to political science, I spend much of the chapter explaining what a Bayesian appraoch to causal inference means with theoretical and practical justifications: how priors are inescapable for many causal claims, how priors provide valuable structure to improve the estimation of causal effects, and practical advice for constructing and evaluating Bayesian causal models. Finally, I provide examples of Bayesian causal modeling by replicating and extending published studies in political science, showing where priors add value to causal inference.

3.1 Overview of Key Concepts

3.1.1 Causal Models

As an area of scientific development, *causal inference* refers to the formal modeling of causal effects, the assumptions required to identify causal effects, and research designs that make these assumptions plausible. Scientific disciplines, especially social sciences, have long been interested in substantiating causal claims using data, but the rigorous definition of the full causal model and identifying assumptions distinguishes the current causal inference movement from other informal approaches.

Contemporary approaches to causal inference span several fields, most notably in economics, epidemiology, and computer science. The dominant modeling approach to causal inference in political science is rooted in a notation of *potential outcomes* (Rubin 1974, 2005). This "Rubin model" formalizes

the concept of a causal effect by first defining a space of hypothetical outcomes. The outcome variable Y for unit i is a function of a treatment variable A. "Treatment" refers only to a causal factor of interest, regardless of whether the treatment is randomly assigned.[1] Considering a binary treatment assignment where $A = 1$ represents treatment and $A = 0$ represents control, unit i's outcome under treatment is represented as $Y_i(A = 1)$ or $Y_i(1)$, and the outcome under control would be $Y_i(A = 0)$ or $Y_i(0)$. The benefit of expressing Y in terms of hypothetical values of A allows the causal model to describe, with formal exactitude, the entire space of possible outcomes that result from treatment assignment. The treatment effect for unit i, denoted τ_i, is the difference in potential outcomes when changing the treatment status A_i.

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

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

If τ_i is not 0, then A_i has a causal effect on Y_i. Defining the causal model in terms of unit-level effects provides an exact, minimally sufficient definition of a causal effect: A affects Y if the treatment has a nonzero effect *for any unit*.

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

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

observational causal inference, it is common to invoke the following four assumptions. The *consistency* assumption states that when a unit receives a treatment, the outcome that we observed is the potential outcome for that unit under that treatment.

$$\text{Consistency: } Y_i = Y_i(A_i = a) \mid A_i = 1 \tag{3.2}$$

In other words, observing the outcome does not affect the outcome, and there are no hidden versions of treatment that are not defined in the causal model. The *no interference* assumption states that unit i's potential outcome depends only on i's treatment status, not on any other units' treatment status.

$$\text{No Interference: } Y_i\left(A_i\right) = Y_i\left(A_1, A_2, \ldots, A_n\right) \tag{3.3}$$

The consistency and no-interference assumptions are sometimes grouped into a single assumption, called the "stable unit treatment value assumption" (or SUTVA, Rubin 1980). The *conditional independence* assumption (CIA), also known as conditional "ignorability," states that potential outcomes are independent of treatment status conditional on a unit's membership in a specific stratum of covariates $X_i = x$.

$$\text{Conditional Independence: } Y_i\left(A_i = 1\right),\ Y_i\left(A_i = 0\right) \perp\!\!\!\perp A_i \mid X_i = x \tag{3.4}$$

The CIA is essentially an assumption about unit-level expectations under counterfactuals. Within stratum $X_i = x$, any observed Y_i in the treatment group are independent from the $Y_i(1)$ potential outcome values for units in the control group. This assumption is violated by self-selection processes or other confounding affects that sort units into treatment statuses in a way that is correlated with their potential outcomes. The CIA is usually invoked alongside a *positivity* assumption,

$$\text{Positivity: } 0 < p\left(A_i = a \mid X_i = x\right) < 1,\ \forall a \in \mathcal{A},\ \forall x \in \mathcal{X} \tag{3.5}$$

where $\mathcal{A}$ and $\mathcal{X}$ are respectively the spaces of possible treatments and covariate strata. Positivity stipulates nonzero probability of treatment statuses within each strata, ensuring that causal inferences are conducted only on units that could conceivably have received other treatment statuses.[2] Conditional

[2] See Liao, Henneman, and Zigler (2019) for Bayesian causal estimation for "overlapping populations" under violations of strict positivity.

independence and positivity assumptions are essential for conducting causal inference using methods of covariate adjustment.

All together, these assumptions enable inferences about potential outcomes, which aren't always observed, using only observed data. For instance, from these assumptions it follows that the average Y values observed at $A = a$ are equal to the average potential outcomes that we would obtain by setting $A = a$ ourselves, within strata $X = x$:

$$\begin{aligned} \mathbb{E}\left[Y_i \mid A_i = a, X_i = x\right] &= \mathbb{E}\left[Y_i(A = a) \mid A_i = a, X_i = x\right] \\ &= \mathbb{E}\left[Y_i(A = a) \mid X_i = x\right] \end{aligned} \tag{3.6}$$

The first line of (3.6) equates observed outcomes given treatment to potential outcomes under an imposed treatment, which is enabled by SUTVA (consistency and no interference). Conditional independence allows us to suppresses the explicit conditioning on $A_i = a$ in the second line, and positivity is implied by referring generically to treatment level a and covariate stratum x.

Implications derived from identification assumptions are typically posed in terms of expectations about potential outcomes, $\mathbb{E}\left[Y_i\left(A_i\right)\right]$, instead of unit-level potential outcomes, $Y_i\left(A_i\right)$, because unit-level causal heterogeneity is unidentified without additional assumptions (Holland 1986). This is why causal quantities of interest, also known as causal estimands, are usually posed in terms of expectations as well. This project will generally discuss *conditional average treatment effects* (CATEs), which are expectations of treatment effects averaged over the population of units within a covariate stratum. Let $\tilde{\tau}(a, a', x)$ be the average effect of setting $A = a$, as opposed to some other value a', within stratum $X = x$.

$$\tilde{\tau}(a, a', x) = \mathbb{E}\left[Y_i(a) - Y_i(a') \mid X_i = x\right] \tag{3.7}$$

When deriving estimands such as the CATE with identification assumptions, it is important to note that assumptions describe minimally sufficient conditions for *nonparametric* causal identification (Keele 2015). There is no guarantee that linear regression models, or any parametric models, adequately control for confounders. For this reason, it is important to distinguish causal estimands from estimation methods, the latter of which can introduce statistical assumptions that differ from identification assumptions.

In order to separate the major planks of causal inference methods, I center the discussion of causal inference going forward around a three-part hierarchy of causal methodology.

1. Causal model: the hypothetical model that defines all treatments and potential outcomes at the unit level. The causal model is an omniscient view of the causal system, defining individual-level causal effects even though they cannot be observed in real data.
2. Identification assumptions: the assumptions required to identify causal estimands using only the observed data. These assumptions specify how knowledge of observed Y data can be used to make inferences about counterfactual Y data that is not observed. Inferences enabled by identification assumptions are typically posed as expectations about counterfactual data, unless further assumptions are specified that relate individual unit effects to average effects.[3] Because inferences are posed as expectations, they are abstractions that follow from analytical derivations and do not depend on statistical estimation approaches.
3. Estimation methods: How do we estimate the expectations derived from the causal model and identification assumptions? Do these estimation methods introduce additional statistical assumptions on top of the identification assumptions?

[3] For instance, a *constant additive effects* assumption states that all causal effects are constant for all units and can be combined without interactions (Rosenbaum 2002).

I lay out this hierarchy for two main reasons. First, it clarifies why researchers use certain research designs or statistical approaches to overcome particular problems with their data. Statistical assumptions, we will see, can undermine identification assumptions, which is why causal inference scholars tend to promote estimation strategies that rely on as few additional assumptions as possible (Keele 2015). One way to avoid these assumptions is to use research designs that eliminate confounding "by design" rather than through statistical adjustment, such as randomized experiments, instrumental variables, regression discontinuity, and difference-in-differences (for instance, Angrist and Pischke 2008). If researchers cannot design away these difficult assumptions, other methods are available to adjust for confounders without as many strict assumptions about the functional form of the causal model as are commonly invoked in parametric regression models. Causal inference is not synonymous with the new "agnostic statistics" movement (e.g. Aronow and Miller 2019), but it is animated by a similar motivation to identify statis-

tical methods that rely on as few fragile assumptions as possible. For causal inference problems, these methods include matching, doubly robust models, and machine learning methods for estimating flexible conditional expectation functions that are to varying degrees robust to misspecification, nonlinearities, or non-additivities in the data generating process (Aronow and Miller 2019; Aronow and Samii 2016; Green and Kern 2012; Hill 2011; Samii, Paler, and Daly 2016; Sekhon 2009). This dissertation will employ machine learning methods, in particular Bayesian neural networks (BNNs), to estimate regression functions that rely less on exact, reduced-form model specification choices.

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

3.1.2 Bayesian Inference

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

Whereas many statistical methods begin with a model of data given fixed parameters, Bayesian inference consists of a joint model for all components in a system. The "joint model" is simply a probability model for more than one event. For example, suppose that we are interested in the joint probability distribution of age and vote choice in a population. The joint distribution of these two variables as $p(Age, Vote)$, which can be equivalently expressed by factoring it in two ways:

$$p(Vote \mid Age)p(Age) = p(Age \mid Vote)p(Vote) \tag{3.8}$$

If we observe an individual's vote choice, how does this affect the probability distribution of age? Probability theory says that we divide the joint probability by the probability of the conditioning event.

$$\begin{aligned} \frac{p(Vote \mid Age)p(Age)}{p(Vote)} &= \frac{p(Age \mid Vote)p(Vote)}{p(Vote)} \\ \frac{p(Vote \mid Age)}{p(Age)} &= p(Age \mid Vote) \end{aligned} \tag{3.9}$$

This maneuver reveals Bayes' theorem: the probability of A given B, expressed in terms of B given A. Bayes' theorem provides a formal method for rationally updating a joint probability distribution by conditioning on known information.

The Bayesian paradigm of applied statistical modeling applies Bayes' theorem to data $\mathbf{y}$ and parameters $\boldsymbol{\pi}$. The joint model for the data and the parameters takes the form

$$p(\mathbf{y}, \boldsymbol{\pi}) = p(\mathbf{y} \cap \boldsymbol{\pi}) = p(\mathbf{y} \mid \boldsymbol{\pi})p(\boldsymbol{\pi}), \tag{3.10}$$

where $p(\mathbf{y} \mid \boldsymbol{\pi})$ represents the probability distribution of the data, conditioning on parameters, and $p(\boldsymbol{\pi})$ represents the probability distribution of parameters, marginalizing over the data. The marginal parameter distribution is usually referred to as a "prior distribution," since it describes the researcher's prior information (or, controversially, prior "beliefs") about the parameter. The joint model provides machinery for learning about parameters by conditioning the parameters on the data,

$$p(\boldsymbol{\pi} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\pi})p(\boldsymbol{\pi})}{p(\mathbf{y})} \tag{3.11}$$

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

These basic concepts of conditional probability and Bayesian updating are not foreign to political science, but it will be important to establish an interpretation of Bayesian thinking and Bayesian modeling that is productive for causal inference. This project invokes what McElreath (2017b) calls an "inside view" of Bayesian statistics—Bayesian statistics on its own terms. The inside view is a response to an "outside view" of Bayesian statistics as penalized maximum likelihood. Under the outside view, data follow probability distributions, while parameters are fixed. Bayesian estimation, in turn, is a likelihood model with an additional penalty on the parameters, and because the penalty represents the researcher's subjective beliefs, the penalty feels *ad hoc* and anti-scientific. This view of Bayesian statistics is admittedly confusing, and if we take it at face value, it is no wonder that causal inference in political science has largely avoided Bayesian tools other than for computational convenience.

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

[4] The semantic conventions are often sloppier in practice than many researchers would like to think. Many analyses use data that summarize lower-level processes, such as per-capita income in U.S or the percentage of women who vote for the Democratic presidential candidate, which behave like random variables in that their values could differ under repeated sampling. The semantic distinction between data and parameters has a similar spirit to the Blackwell, Honaker, and King (2017) view of measurement uncertainty, where "measurement error" falls on a spectrum between fully observed data and missing data.

dating data given parameters) or "running the model backward" (updating parameters given data).

From the inside view, Bayesian updating proceeds by considering a variety of possible scenarios that create data and evaluating which scenarios are consistent with the data. The joint prior model, $p(\mathbf{y} \mid \boldsymbol{\pi})p(\boldsymbol{\pi})$ describes an overly broad set of possible configurations of the world. These configurations contain a distribution of possible parameters, $p(\boldsymbol{\pi})$, and possible data given parameters, $p(\mathbf{y} \mid \boldsymbol{\pi})$. Bayesian updating decides which configurations of the world are consistent with the data and are therefore more plausible. The plausibility, or posterior probability, of a parameter value is greater if the observed data are more likely to occur under that parameter value versus another value. In turn, the posterior distribution downweights model configurations that are implausible, or inconsistent with the data (McElreath 2020, chap. 2). This is an important distinction from non-Bayesian statistical inference, since there can is no formal notion of "plausible parameters given the data" without a posterior distribution, which necessitates a prior distribution. For causal inference, this means there can be no formal notion of "plausible causal effects" without a probability distribution over causal effects. The mission in the remainder of this chapter is to establish a framework for causal inference in terms of plausible effects and plausible counterfactuals.

The inside view of Bayesian modeling, and the philosophical unity that it brings to statistical machinery and inference, is possible even if using uninformative prior distributions that are indifferent to possible parameters *a priori*. This is how Bayesian methods tend to appear in political science to date, with noninformative priors that exist primarily to facilitate Bayesian computation for difficult estimation problems. The infamy of Bayesian methods, however, is owed to the ability of the researcher to specify "informative" priors that concentrate probability density on model configurations that are thought to be more plausible even before data are analyzed. There are many modeling scenarios where this concentration of probability delivers results that are almost unthinkable without prior structure: multilevel models that allocate variance to different layers of hierarchy, highly parameterized models with correlated parameters such as spline regression, and sparse regressions where regularizing priors are used to shrink coefficients and preserve degrees of freedom to overcome the "curse of dimensionality" (Bishop 2006; Gelman et

al. 2013). At the same time, many researchers are skeptical of Bayesian methods because supplying a model with non-data information can be spun as data falsification (García-Pérez 2019). As I elaborate in Section 3.3, it is a mistake to equate flat prior "flatness" with prior "uninformativeness," and there are many legitimate sources of prior information that have nothing to do with subjective beliefs.

3.2 Bayesian Causal Modeling

Having reviewed the basics of causal models and Bayesian inference, we now turn to a framework for Bayesian causal modeling. The distinguishing feature of a Bayesian causal model is that the elemental units of the model, the potential outcomes, are given probability distributions. This probability distribution reflects available causal information that exists outside the current dataset. Bayesian inference proceeds updating our information about causal effects and counterfactual potential outcomes in light of the observed data. The headings under 3.2 introduce this modeling framework at a high level. I provide a probabilistic interpretation and notation for potential outcomes models (3.2.1), a connection between causal parameters and model parameters (**??**), and a broad justification for the Bayesian interpretation of causal effects (3.2.3).

3.2.1 Probabilistic Model for Potential Outcomes

As with other causal models, we begin at the unit level. Unit i receives a treatment $A_i = a$, with potential outcomes $Y_i(A_i = a)$. Suppose a binary treatment case where A_i can take values 0 or 1, so the unit-level causal effect is $\tau_i - Y_i(1) - Y_i(0)$. Although τ_i is unidentified, it is possible to estimate population-level causal quantities by invoking identification assumptions. For instance, the conditional average treatment effect at $X_i = x$, $\bar{\tau}(X = x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$, can be estimated from observed data assuming consistency, non-interference, conditional ignorability, and

positivity. Suppressing the unit index i,

$$\begin{aligned}\bar{\tau}(X = x) &= \mathbb{E}\left[Y(A = 1) - Y(A = 0) \mid X = x\right] \\ &= \mathbb{E}\left[Y(A = 1) \mid X = x\right] - \mathbb{E}\left[Y(A = 0) \mid X = x\right] \\ &= \mathbb{E}\left[Y \mid A = 1, X = x\right] - \mathbb{E}\left[Y \mid A = 0, X = x\right]\end{aligned} \tag{3.12}$$

where the third line is obtained by the identification assumptions. The identification assumptions connect *causal estimands* and what I will call *observable estimands*. Causal estimands are the true causal quantities, but they are unobservable because they are stated as contrasts of potential outcomes. Observable estimands are the observable analogs of causal estimands and are equivalent to causal estimands if identification assumptions hold. Other literature refers to observable estimands as "nonparametric estimators" (Keele 2015), but I steer clear of this language because the gap between observable estimands and estimators is important for understanding the contributions of the Bayesian causal approach.

The transition to a Bayesian probabilistic model begins with an acknowledgment that no estimate of the observable estimand, $\mathbb{E}\left[Y \mid A = a, X = x\right]$, will be exact. The assumptions identify causal effects only in an infinite data regime, where the observable estimand is known exactly. Inference about causal effects from finite samples, however, requires further statistical assumptions that link the observable estimand to an estimator or model. Let $f(A_i, X_i, \boldsymbol{\pi}) + \varepsilon_i$ be a model for Y_i consisting of a function $f(\cdot)$ treatment A_i, covariates X_i, and parameters $\boldsymbol{\pi}$, and an error term ε_i. We let the systematic component $f(\cdot)$ be a plug-in estimator for $\mathbb{E}\left[Y \mid A = a, X = x\right]$. This setup is similar to any modeling assumption that appears in observational causal inference to link an estimator to the observable estimand, including parametric models for covariate adjustment, propensity models, matching, and more (Acharya, Blackwell, and Sen 2016; Sekhon 2009).[5] We use the statistical model to generate a CATE estimate, $\hat{\bar{\tau}}(X = x)$, by differencing these model predictions over the treatment.

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

$$\begin{aligned}\bar{\tau}(X = x) &= \mathbb{E}\left[Y \mid A = 1, X = x\right] - \mathbb{E}\left[Y \mid A = 0, X = x\right] \\ \hat{\bar{\tau}}(X = x) &= \mathbb{E}\left[f(A_i = 1, X_i 1 = x, \boldsymbol{\pi}) - f(A_i = 0, X_i = x, \boldsymbol{\pi})\right]\end{aligned} \tag{3.13}$$

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

The Bayesian approach, inspired largely by Rubin (1978), confronts the problem with a joint model for data and parameters: $p(Y, \boldsymbol{\pi}) = p(Y \mid f(A, U, \boldsymbol{\pi}))\, p(\boldsymbol{\pi})$. The data are distributed conditional on the statistical model prediction $f(\cdot)$, which conditions on the model parameters $\boldsymbol{\pi}$. The parameters also have a prior distribution $p(\boldsymbol{\pi})$, or a distribution marginal of the data. These models for data and parameters are added statistical assumptions on top of causal identification assumptions. The data model is similar to any estimation approach that uses a probability model for errors (e.g. any MLE method or OLS with Normal errors). The prior model has no analog in OLS or unpenalized MLE, but this added statistical assumption will be leveraged as a major benefit as we explore Bayesian causal estimation below.

The joint generative model is sufficient to characterize the probability distribution for the conditional average treatment effect as defined in Equation (3.13),

$$p(\bar{\tau}(X = x)) = \int p\left[f(A = 1, X = x, \boldsymbol{\pi}) - f(A = 0, X = x, \boldsymbol{\pi}) \mid \boldsymbol{\pi}\right] p(\boldsymbol{\pi})\, \mathrm{d}\boldsymbol{\pi} \tag{3.14}$$

which is the probability distribution of model contrasts for $A = 1$ versus $A = 0$. This distribution of model contrasts contains two sources of uncertainty: uncertainty about the data given parameters, and uncertainty over the parameters themselves. Integrating over $\boldsymbol{\pi}$ in Equation (3.14) marginalizes the distribution with respect to the uncertain parameters. Because the integrated parameters are the prior distribution $p(\boldsymbol{\pi})$, the expression in (3.14) represents a prior distribution for the CATE. This is an inherent feature of the Bayesian approach: probability distributions of causal quantities even before data are observed.

Conditioning on the observed data returns the posterior distribution for the CATE…

$$p(\bar{\tau}(X = x) \mid Y) = \int p\left[f(A = 1, X = x, \boldsymbol{\pi}) - f(A = 0, X = x, \boldsymbol{\pi}) \mid \boldsymbol{\pi}, Y\right] p(\boldsymbol{\pi} \mid Y)\, \mathrm{d}\boldsymbol{\pi} \tag{3.15}$$

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

The Bayesian causal approach makes sense because it enables *direct inference* about treatment effects given the data: which effect sizes are "more likely" or "less likely" than others. Unlike confidence intervals, the posterior distribution enables the researcher to state the probability of a positive treatment effect, or the probability that the treatment effect is likely to be negligible. Building positive statements about model conclusions is a natural way to think about the scientific aims of any discipline engaged causal inference: the world probably works in this or that way, given this evidence. Although this sounds banal, it isn't statistically coherent to invoke similar language under a non-Bayesian inference paradigm, since characterizing the plausibility of causal effects can't be evaluated with a posterior distribution, which entails a prior distribution. Instead, non-Bayesian methods conduct inference by making probability statements about *data* conditional on hypothesized parameters, not parameters themselves. Data-focused inference statements (e.g. p-values) become conclusions about parameters only after invoking additional decision rules that compress the uncertainty in model results into discrete conclusions.

Another reason why the Bayesian causal modeling approach makes sense is because causal models, at their core, are models for counterfactual data. Because the Bayesian model is a *generative* model for parameters and data, the model contains all machinery required to directly quantify counterfactual potential outcomes using probability distributions. To see this in action, remember that we can "run the model forward" to create a predictive distribution for Y given the model parameters. Denote these simulated observations as $\tilde{Y}$ to distinguish them from the data Y that are actually observed. If we marginalize this predictive distribution with respect to the prior parameters, we obtain a "prior predictive distribution"—the distribution of data we would expect under the prior (Gelman et al. 2013).

$$p\left(\tilde{Y} \mid A = a, X = x\right) = \int p\left(\tilde{Y} \mid A = a, X = x, \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi}\right) \, \mathrm{d}\boldsymbol{\pi} \tag{3.16}$$

We update this distribution by conditioning on new data, delivering a "posterior predictive distribution"—the distribution of data that we expect from the posterior parameters.

$$p\left(\tilde{Y} \mid Y, A = a, X = x\right) = \int p\left(\tilde{Y} \mid A = a, X = x, \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi} \mid Y\right) \, \mathrm{d}\boldsymbol{\pi} \tag{3.17}$$

These predictive distributions are the basis for out-of-sample inference in Bayesian generative models,[6] and they are the basis for counterfactual inference as well. Invoking the causal identification assumptions, we can obtain the probability distribution counterfactual data as predictive distributions as well, setting the treatment $A = a$ to some other value $A = a'$. Denote these counterfactual predictions $\tilde{Y}'$, which I will subscript i to show that this model implies a probability distribution for individual data points as well as aggregate treatment effects.

$$p\left(\tilde{Y}_i' \mid Y, A_i = a', X_i = x\right) = \int p\left(\tilde{Y}_i' \mid A_i = a', X_i = x, \boldsymbol{\pi}\right) p\left(\boldsymbol{\pi} \mid Y\right) \, \mathrm{d}\boldsymbol{\pi} \tag{3.18}$$

Stated more simply: if causal models define a space of potential outcomes, then Bayesian causal models are probabilistic representations of the potential outcome space. Probability densities over potential outcomes are defined in the prior and after conditioning on observed data, and they can be defined all the way to the unit level if the generative model contains a probability density for unit outcomes.[7] In this way, Bayesian causal inference as nothing more remarkable than a missing data model for unobserved counterfactuals (see Rubin 2005).

As mentioned above, this is not the first time causal inference has been frames as a Bayesian modeling problem. Indeed, the Bayesian interpretation of causal inference and missing counterfactuals is at least as old as Rubin (1978). Although it is rarely invoked in practice today, Rubin writes that "the posterior distribution of causal estimands, obtained from the posterior predictive distribution of the missing potential outcomes…is the most natural way to summarize evidence for a scientific question" (Rubin 2005). Unlike other Bayesian setups, however, this chapter will contain more practical guidance for thinking Bayesianly about causal inference, more synthesis of Bayesian causal innovations from other fields, and more examples of Bayesian causal modeling in practice.

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

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

[6] Simulations of this sort are possible under any likelihood-based model that posits a generative probability distribution for the data. Bayesian predictive distributions are different, however, because they also marginalize over the parameter distribution instead of merely conditioning on fixed parameters. This makes Bayesian predictive distributions a more thorough representation of the total statistical uncertainty in the model.

[7] If we can estimate average causal effects with group-level statistics, we can elide the unit-level model in favor of a more robust distributional assumption for the statistics, e.g. the Central Limit Theorem for group means. Naturally, models of this type will stop short of defining probability distributions for counterfactual observations, but if will define probability distributions for counterfactual averages. In some cases, such as binary outcome data, means in each group are sufficient statistics for the raw data, so the unit level model is implied by the group-level model. See Section 3.4.4 for explanation and examples.

the treatment status.

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

Guidelines:

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

Causal model: $y_i(1)$ and $y_i(0)$
Observed data: $y_i = z_i y_i(1) + (1 - z_i) y_i(0)$
Unobserved data: $\tilde{y}_i$
Inference about $\tilde{y}_i$: $p(\tilde{y}_i \mid \mathbf{y}) \propto p(\tilde{y}_i \mid \pi, \mathbf{y}) p(\pi)$
This fits an ML approach:

- ML approach to causal inference recasts the problem as a prediction problem for the treatment assignment.
- Treatment effects are then integrated over the uncertainty in the propensity
- This propagates uncertainty using the exact model machinery, rather than a post-hoc computational workaround like bootstrapping
  - Bootstrapping does have an appealing feature that it isn't making a parametric assumption about errors in the model. It simply uses the "sampling uncertainty" intuition to build intuitive bounds on effect uncertainty.
  - However in many real-world cases the data we have cannot be resampled, so the framework for inference under bootstrapping doesn't make theoretical sense without appealing to a "superpopulation" philosophy.

  - Bayes naturally deals with this by assigning a probability distribution to the unmodeled features, which mechanistically follows a similar assumption as a parametric assumption about errors in any typical regression case: the prior isn't really extra.
  - If this assumption isn't something you want, it is possible to generalize the model by specifying an overdispersion parameter and letting the data inform what which model estimates can be rationalized against plausible overdispersion parameters. For example, Kruske's BEST approach to difference-in-means testing.

- In a Bayesian framework, we have posterior More of a "priors to facilitate posteriors" approach, not a prior information view (since there is some formal data peeking involved)

  Works without repeated sampling.
  Probabilistic treatment status?

3.2.2 Bayesian Inference in the Hierarchy of Causal Inference

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

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

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

the causal estimand. After all, the estimator exists in the space of real data, unlike a causal estimand that belongs to the hypothetical space of potential outcomes. Being merely an estimator, we could evaluate the Bayesian model for its bias and variance like any other estimator.

The realm of "mere estimation" is where many Bayesian causal approaches appear in political science and other fields. The estimation benefits of Bayes tend to fall into three categories: priors provide practical stabilization or regularization, posterior distributions are convenient quantifications of uncertainty, or MCMC provides a tractable way to fit a complex model. We could characterize the use of Bayesian methods for these purposes as practically valuable but theoretically dispensable, in the sense that researchers might prefer non-Bayesian means to the same ends. For instance, Green and Kern (2012) adapt Bayesian Additive Regression Trees (or BART, Chipman et al. 2010) to measure treatment effect heterogeneity in randomized experiments. (See Hill 2011 for a non-political science introduction to causal inference with BART.) The advantages of a regression tree model for treatment heterogeneity is that it can explore arbitrary interactions among covariates while controlling overfitting, but the fact that BART is Bayesian isi an afterthought. We observe a similar pattern in the use of Gaussian process models for fitting the running variable in regression discontinuity designs by Ornstein and Duck-Mayr (2020) and in the development of augmented LASSO estimators for sparse regression models by Ratkovic and Tingley (2017, @ratkovic-tingley:2017:sparse-lasso-plus).[8] These authors use priors to regularize richly parameterized functions, posterior distributions to characterize uncertainty, and MCMC to estimate models, but the theoretical implications of Bayesian causal estimation are not a major focus.

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

What does it mean for Bayesian estimation to have theoretical implications for causal inference? This brings our focus to the first level of the causal inference hierarchy: the definition of potential outcomes Any estimation method that invokes Bayesian tools requires priors for model parameters. These priors model the plausibility of parameter values marginal of any observed data. Because potential outcomes are functions of the parameters that model a causal effect (such as a difference in means), and Bayesian models require priors on parameters, this implies at the very least that causal effects have probability densities before considering the data. If the Bayesian model

contains a unit-level probability density, which is the case for most regression approaches (e.g. Normally distributed regression errors), this directly implies that unit-level potential outcomes also have prior distributions.

This is only interesting if I can convince you that we MUST have priors in a causal model, and I'm going to show you that that is true. This might be no problem if I just give everything a flat prior, but I'm going to show you that that doesn't make sense.

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

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

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

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

Ratkovic and Tingley (2017) Carlson (2020)

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

Lattimore and Rohde (2019)

The Bayesian causal model began as an estimator for the an observable estimand, so it belonged to level 3 in the hierarchy. As an estimation approach, it doesn't obviously change the observable estimand or the causal estimand. Instead, it is an estimator that can be evaluated for its bias and variance like any other estimator.

The model contained a data model conditional on parameters, and a parameter model

What was I trying to say here?

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

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

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

- The estimator (3) of an expectation or a counterfactual is thus a different thing from the hypothetical estimand.

- Some causal models imply nonparametric estimators, so no Bayesian anything will be required. However, estimates themselves are in-sample quantities, and any generalization about the in-sample effect can be augmented with priors.
- Bayes is an *estimation* framework that doesn't change the causal model itself. It is a different way of estimating the quantities in an in-sample causal estimand.
- Oganisian and Roy (2020) "identification" assumptions vs. "statistical" assumptions. Identification assumptions get you to a place where you can express causal effects in terms of expectations about Y given treatment, within confounding strata (perhaps then averaging over confounders). "Statistical" assumptions define how we think we can build $E[Y]$

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

- For nonparametric estimators, structural priors can be helpful for concentrating probability mass in sensible areas of space, since nonparametric estimators may be lower-powered than estimators that get a power boost from parametric assumptions.
- Parametric models, in term, are obvious areas where Bayesian estimates can go, but they are stacking more assumptions on top of the parametric assumptions.
- Semi-parametric models are an interesting middle ground, where we want to be flexible about the exact nature of the underlying relationships, but we want to impose some stabilization to prevent the model from behaving like crazy.

Do battle with the "implied nonparametric estimator" framework.

- Causal models are manipulated to express causal *estimands*.
- the "nonparametric estimators" are usually just averages of treatment and control group outcomes, under ignorability assumptions.
- Parametric models for estimands are usually a byproduct of thinking within an MLE/regression framework. Many causal inf techniques don't do that. Instead they try to simply predict E[y(1) - y(0)] and don't care about the interpretability of the other RHS terms.

- Thinking outside the "typical econometric" framework it's actually kinda easy to see how one flavor of fitting Bayesian models for causal effects. If the technology for prediction is Bayesian, you just get the prediction and the posterior distribution. You then deal with the bias/variance properties of $\hat{y}$ or $\tilde{y}$ (unobserved counterfactuals) but at least your focus is on the predictions rather than coefficients (which, from this view, who cares about those?)
- This actually fits quite nicely with the machine learning approach to causal inference

We can think bigger about Bayesian *inference* for a parameter as distinct from Bayesian *estimation* of the in-sample quantity. This lets us use a non-parametric data-driven estimator for the data, but the "inference" or "generalization" still has a prior. For instance a sample mean estimates a population mean without a likelihood model for the data, but inference about the population mean often follows a parametric assumption from the Central Limit Theorem that the sampling distribution from the mean is asymptotically normal (but doesn't have to, c.f. bootstrapping). Even if the point estimator we use for a mean is unbiased, we can assimilate external information during the interpretation of the estimator (biasing the inference without biasing the point estimator). Restated: the posterior distribution is a weighted average of the raw point estimator and external information, rather than biasing the data-driven estimate directly.

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

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

some hybrid view in the "quasi-experimental" approach to causal model selection. We estimate two models from the same data—one with a treatment parameter and one where we impose no treatment effect—and compare models using some likelihoodist or Bayes factor measure of evidence.

What is THIS project going to do?

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

3.2.3 Inferential Goals

1. posterior isn't about frequency properties

2.

What is "bias?"

- look up in BDA
- "Requiring unbiased estimates will often lead to relevant information being ignored (as we discuss with hierarchical models in Chapter 5)" (94)

Why would we want this? Inference makes more sense.

- What's the probability of a model/hypothesis, given the data
- vs. What's the probability of "more extreme data" (?) given a model that I don't believe.
- posterior probabilities mean what they say they mean
  - conditioning on data (and implicitly the model), this is the distribution of parameters
  - p-values are the "probability of more extreme results." They condition on the model, but they're only useful if they don't.

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

Are we in a repeated sampling framework at all?

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

3.2.4 *Shared Goals, Different Tactics*

Fux with GGK:

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

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

3.3 *Making Sense of Priors in Causal Inference*

3.3.1 *Priors as Data Falsification*

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

$$p(\mu \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mu) p(\mu) \tag{3.19}$$

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

$$p(\mu \mid \mathbf{y}) \propto \prod_{i=1}^{n} l(y_i) p(\mu) \tag{3.20}$$

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

$$\log p(\mu \mid \mathbf{y}) \propto \sum_{i=1}^{n} \log l(y_i) + \sum_{i=1}^{n} \log p(\mu)$$

$$\propto \log l(y_1) + \log p(\mu) + \log l(y_2) + \log p(\mu) + \ldots + \log l(y_n) + \log p(\mu) \tag{3.21}$$

The setup in (3.21) 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 (3.21) 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 (3.21) 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.

$$\log p(\mu \mid \mathbf{y}) \propto \log l(y_1) + \log l(y_2) + \ldots + \log l(y_n) \tag{3.22}$$

If $p(\mu)$ is not flat, however, and $p(\mu)$ varies across values of μ, we can no longer ignore $p(\mu)$ in (3.21) shows that $p(\mu)$ varies across values μ. Not only does this prevent us from dropping $p(\mu)$ from the proportionality, but it also reveals how the prior "adds information" to the posterior by the same mechanism that observations do: adding to the log posterior distribution. This general expression where both data and priors contribute to the posterior distribution has led some researchers to argue the Bayesian inference with non-flat priors is analytically indistinguishable from data falsification (García-Pérez 2019). We can highlight this behavior by obscuring each $p(\mu)$ term with a square □.

$$\log p(\mu \mid \mathbf{y}) \propto \log l(y_1) + \square + \log l(y_2) + \square \ldots + \log l(y_n) + \square \tag{3.23}$$

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

3.3.2 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 $\vec{\alpha}$.

$$p\left(\vec{\alpha} \mid y\right) \propto p(y \mid \vec{\alpha}) p(\vec{\alpha}) \tag{3.24}$$

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

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

Alex meeting takeaways:

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

3.3.3 Priors and Model Parameterization

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

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

In a maximum likelihood framework, equivalent parameterizations are a more benign feature of the modeling framework. Reparameterization may

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

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.

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{2\sigma}\right)^2} = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(x-\mu)^2}{2}} \tag{3.25}$$

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

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

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \text{where } \varepsilon_i \sim \text{Normal}(0, \sigma) \tag{3.26}$$

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

$$y_i \sim \text{Normal}(\alpha + \beta x_i, \sigma) \tag{3.27}$$

Algebraically, these two models are identical. The difference is only a matter of which component has the distributional assumption. In Equation (3.26), the distribution is assigned to ε_i, so y_i is a random variable only by way of ε_i. In Equation (3.27), we assign the distributional assumption directly to y_i, bringing the regression function into the mean rather than "factoring it out" of the distribution.

The linear regression context is one context where the choice of parameterization appears. These two parameterizations are typically called the "centered" and "non-centered" parameterization for a Normal distribution. In the centered parameterization, the random variable is drawn from a distribution "centered" on a systematic component, whereas the non-centered distribution

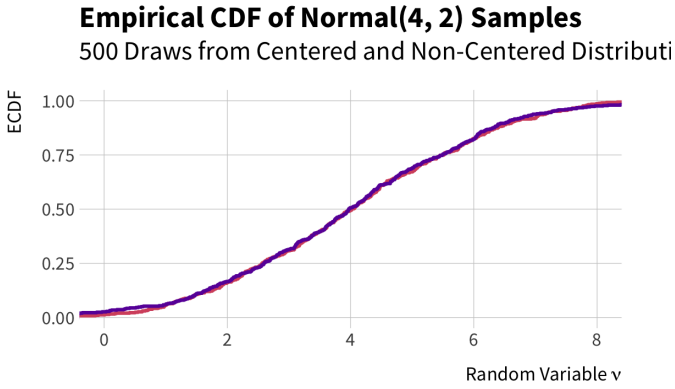

Figure 3.1: Demonstration of centered and non-centered parameterizations for a Normal distribution. The non-centered parameterization is statistically equivalent, but the location and scale are factored out of the distribution.

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 ν with mean 4 and standard deviation 2.

$$\text{Centered Parameterization:} \quad \nu \sim \text{Normal}(4, 2) \tag{3.28}$$

$$\text{Non-Centered Parameterization:} \quad \nu = 4 + 2z, \quad \text{where } z \sim \text{Normal}(0, 1) \tag{3.29}$$

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.

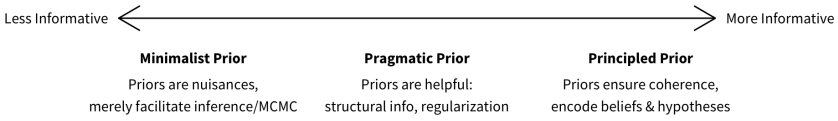

Figure 3.2: A spectrum of attitudes toward priors.

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

3.3.4 Balancing Pragmatic and Principled Approaches to Priors

Various roles that priors take on

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

A general orientation toward priors in this dissertation:

- Not about "stacking the deck" or hazy notions of "prior beliefs"
- information, not belief
- Bayesian view of probability is *more general*, contains information and beliefs. Information is priors, but it's also data. Information is the fundamental unit of uncertainty-quantification

- inference about the thing we care about (counterfactuals)
- structural information when we have it
- Causal inference: "agnosticism" is something valuable generally

  Priors are not de-confounders

- downweighting, not upweighting

3.3.5 Generalization, Big and Small

Models (likelihoods) are priors

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

  Identification assumptions are priors
  Generalization to any population is a prior

- priors are not MY data, but ANY data
- parameters describe ANY data
- Lampinen Vehtari 2001: likelihood as "prior for the data" is the basis for all generalization from any finite model

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

Email to David:

> My first reaction to this is like this: it's probably correct to say that Bayesians may have some shared ideas about how to think about generalizing that might differ systematically from non-Bayesians, but I'm not sure how much of that is because of Bayes per se so much as just…the types of modeling someone is willing to do. By "modeling" I mean, functional form assumptions are you willing to make about data, which is different from "Bayes" in that the former is something required of all modeling and the latter is only what you can say about parameters. For example, a functional modeling thing you might do is specify (using some weights or something) how an estimate in one sample might map to another sample, but whether you do that with Bayesian estimation is a separate choice aside from the functional model itself. That being said, even though functional modeling things can be done with Bayes/non-Bayes point of view, it does seem right to say that certain modeling approaches may feel more

natural in a Bayesian framework, or that Bayesians might construe a problem slightly differently because they are more used to hierarchical modeling.

That's a pragmatic view of things, but I can give you a more theoretically abstract view, and think of situations where Bayesian lets you do things that non-Bayes can't. It all comes down to how "seriously" you want to take the tenets of Bayesian work and what kind of generalization-based claims you want to make. So I will lay out a series of vignettes that start from a world where Bayes is "not unique" and then gets into worlds where Bayes gets more and more necessary to say what you want to say about the out-of-sample world. I will use the example of estimating some parameter, but you can translate this into learning about a "mechanism" however Erica and Nick are defining that, and you can think about it non-statistically as well even if I'll use statistical modeling language.

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

If the study is unrepresentative or "not externally valid" then it's up to the researcher to specify some approach to modeling the invalidity: $\theta \sim D(f(\theta), \sigma)$, where f() is some function that distorts the representativeness of the study. Which is to say, $f(\theta)$ is the expectation for a study with these distortions. Researcher's task is then to learn the form of f(). These distortions might be like sample bias, the country where the study was performed, or whatever, and all you're really doing is reweighting or adjusting the estimate to make more sense for the target population. If you can estimate parameters that determine f(), viola you can infer the posterior distribution for the true θ. But this is what I mean when I say that none of this is really EXCLUSIVELY Bayesian. Reweight-

ing/adjusting happens in non-Bayes world all the time; the main thing that is different is how you write the model and your ability to say that the population estimate is a "posterior of the true parameter, given the information learned from the data." One example of this kind of thing is maybe the multilevel regression and poststratification: we use national surveys to model the attitudes of different demographic groups, and then we use those model predictions plus census information to project estimates for smaller units, for example states or counties, based on the demographic composition of those units! This example goes the other direction from where Erica and Nick want to go (from representative to unrepresentative) but the technology sounds similar: estimate something in the data that you have, and map it into a space where you don't have data. MRP in political science comes from Bayes world and feels natural there, but nothing saying that it HAS to be Bayesian in its overall approach.

Now we get to a world where Bayes is more necessary. If there's something TRULY Bayesian that really makes no sense withoutBayes, it is the fact that I actually don't need data about f() in order to estimate θ. This is because I have priors even in the absence of data. If all I have is priors about f(), then even if I collect data about ONLY μ and NOTHING about f(), I nonetheless update my information about θ. This is because the parameters are functionally related through f(), so if I learn about the subsample then I learn about the population. Stated differently, learning about μ restricts the space of θ because I can basically "solve backward" using my priors about f. This is the stuff that is very natural in Bayes, and I can think of basically no analog in non-Bayes that lets you do something similar (other than picking point estimates for unknowns and simulating, which doesn't have the same theoretical coherence as a prior/posterior distribution). Of course, this means that inference on f() is subject to the priors that go into f(), which is exactly the kind of thing that non-Bayesians are super afraid of despite majorly misconstruing how this works (IMO). For one, the functional form of f() is the kind of thing non-Bayesians would make assumptions about anyway, so that's not unique to Bayes at all. And secondly, the priors that would go into f() would usually be generic enough that researchers aren't "picking their hypothesis" (a common and frankly stupid stereotype) so much as restricting the space of f() to rule out stuff that's frankly impossible. Happy to give you more concrete examples of the kind of "weakly informative priors" that someone would use in a situation like that if it's a route you want to dig into more. It's this kind of stuff that I think non-Bayesians are under-utilizing: how much extrapolation power you get by being willing to place even weak priors on stuff you can't exactly identify with data. And if you REALLY want to give a nod to Bayesian views of extrapolation, this is the area you'd want to dig into, because it's the stuff that doesn't really make sense without Bayes. You can sort of see Ken and I do this in our voter ID

paper (which you can find on my website) though we kind of wimp out of fully placing priors on f().

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

3.3.6 Modeling Cultures in Political Science: Complexity and Agnosticism

Sidestepping priors

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

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

At any rate:

- simple case: sensitivity testing for noisy circumstances
- complex case: stabilizing highly parameterized problems
  - dynamic TCSC models, lots of parameters
  - that hierarchical conjoint thing
  - priors in high dimensions are scary: consider parameterizations and do simulations

3.3.7 Principled Approaches to Model Parameterization

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

Constrained parameters in causal mediation?

For instance, consider a simple experiment with a binary outcome variable y_i and binary treatment assignment $z_i \in \{0, 1\}$. Suppose that the treatment

effect of interest is a difference-in-means, $\bar{y}_{z=1} - \bar{y}_{z=0}$, estimated from a linear probability model. This linear probability model might be parameterized in two ways. First is a conventional regression setup,

$$y_i = \alpha + \beta z_i + \varepsilon_i \tag{3.30}$$

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.

$$y_i = z_i \mu_1 + (1 - z_i)\, \mu_0 + \varepsilon_i \tag{3.31}$$

Although the treatment effect β from Equation (3.30) is equivalent to the difference in means $\mu_1 - \mu_0$ from Equation (3.31), 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 **??**.

- equivalent parameterizations

3.3.8 Structural Priors and Weak Information

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

p doesn't care about your *n*.

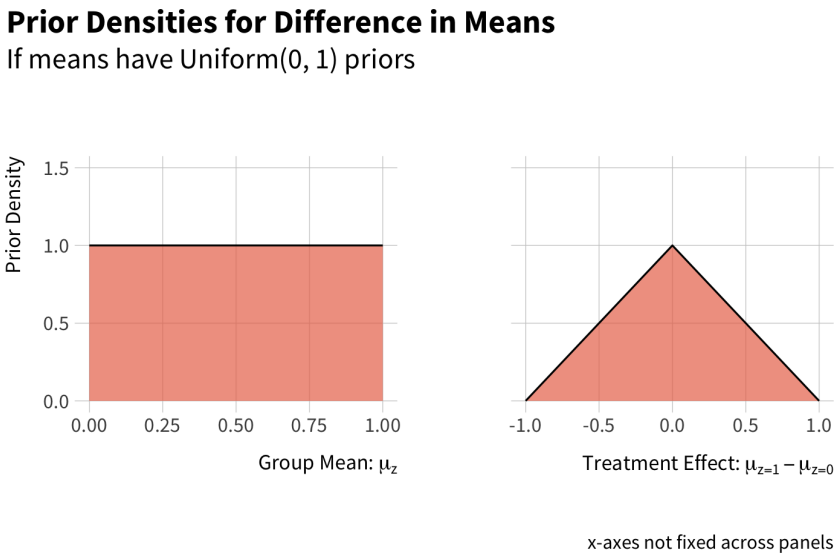

Figure 3.3: Model parameterization affects prior distributions for quantities of interest that are functions of multiple parameters or transformed parameters. The left panel shows that the difference between two means does not have a flat prior if the two means are given flat priors. Note that the *x*-axes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

3.3.9 *Understanding Log Prior* Shape

This is low-key pretty big

3.3.10 *Regularization-Induced Confounding*

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

3.3.11 *Priors for Imperfect Identifiability*

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

This is where Imbens and Rubin push (also Horiuchi et al. example)
Randomization limits the impact of the Bayesian assumptions

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

3.4 Bayesian Opportunities

3.4.1 Full Posterior Uncertainty

Multi-stage models

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

Matching

Predictive/ML models

Multiple comparisons and regularization

3.4.2 Priors for Key Assumptions

Relatedly: priors over models

3.4.3 Application: Weakly Informed Regression Discontinuity

This section presents a reanalysis of Hall's (2015) regression discontinuity study of congressional elections. Portions of the original analysis contain a pathological result where confidence intervals for key parameters of interest contain values that could not possibly occur with nonzero probability. We overcome the pathology using weakly informative priors that contain structural information about the dependent variable only, excluding impossible parameters from the prior but uninformative over possible parameter values. This minor prior intervention successfully guides the posterior distribution away from impossible regions of parameter space, resulting in a posterior distribution that is consistent with the data as well as external structural information about the problem. This intervention does not undermine the main takeaway from the original study, but the Bayesian estimates for the effect of interest are notably smaller and more precisely estimated.

With similar aims to this project, Hall (2015) examines primary elections and their impact on ideological representation in Congress. The study asks if extremist candidates for Congress are more likely or less likely to win the general election contest than candidates who are relatively moderate by comparison. The treatment variable of interest, the ideological extremity of a

party's general election nominee, is confounded by several factors. Competitive districts may lead more moderate candidates to run in the first place, creating selection biases for which candidates represent which district. Conversely, voters in electorally "safe" districts may feel freer to nominate more extreme candidates because the likelihood of their party losing the seat in the general election is sufficiently low. Furthermore, the incumbency advantage in general elections confounds this picture if incumbents tend to be more moderate than challengers who are angling to raise their name recognition (Gelman and King 1990).

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

$$\lim_{x \downarrow x_0} \mathbb{E}\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow x_0} \mathbb{E}\left[Y_i \mid X_i = x\right] = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = x_0\right] \tag{3.32}$$

Equation (3.32) implies that the difference in potential outcomes can be identified from observed data only by observing that everything else about units is continuous at the threshold except for the realized treatment value.

Hall (2015) applies the RD design by assuming that the effect of candidate ideology on vote share and win probability in the general election are iden-

tified locally where the extremist's margin in the primary election crosses 0. For this example, we concentrate on models that predict win probability, since these are the estimates that contain pathological results that we can avoid with Bayesian methods. Hall estimates RD models using a few different specifications, but I replicate his simplest design, which is a linear probability model (LPM) of the following form. The local linear regression is justified by the limit intuition of the key assumption in (3.32); any nonlinear regression function, as long as it is continuous at the cutoff, converges to linearity at the cutoff in the limit. Data were obtained from Hall's replication materials, available on his website.[10] The outcome y_{dpt} is a binary indicator that takes the value 1 if the general candidate running in district d for party p in election year t wins the general election, and it takes 0 if the candidate loses the general election,

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

$$\begin{aligned} y_{dpt} = \beta_0 &+ \beta_1 (\text{Extremist Wins Primary})_{dpt} + \beta_2 (\text{Extremist Primary Margin})_{dpt} \\ &+ \beta_3 (\text{Extremist Wins Primary} \times \text{Extremist Primary Margin})_{dpt} + \varepsilon_{dpt} \end{aligned} \tag{3.33}$$

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

$$\begin{aligned} y_{dpt} &= \alpha_{w[dpt]} + \beta_{w[dpt]} (\text{Extremist Primary Margin})_{dpt} + \varepsilon_{dpt} \\ \varepsilon_{dpt} &\sim \text{Normal}(0, \sigma) \end{aligned} \tag{3.34}$$

where α_w is an intercept for treatment status w, and β_w is the slope for treatment w. This parameterization implies two lines, one line for $w = 0$ and another line for $w = 1$. The treatment effect at the discontinuity is the differ-

ence between the intercepts, $\alpha_1 - \alpha_0$. This parameterization will be helpful for extending the model below.

I plot the OLS win probability estimates around the discontinuity in Figure 3.4. At the discontinuity, we estimate that extremism decreases a candidate's win probability by 0.53 percentage points, which is the same effect found by Hall (2015). The original publication lacks a graphical depiction of these results. Our visualization of the RD predictions reveal that the confidence set for the parameter estimates that compose the treatment effect contain many values that would be impossible to observe. The point estimate for average moderate candidate win probability at the discontinuity is 0.95, which is a possible number to obtain, but the 95 percent confidence intervals includes values as high as 1.24, which far exceeds the maximum possible value of 1.0.

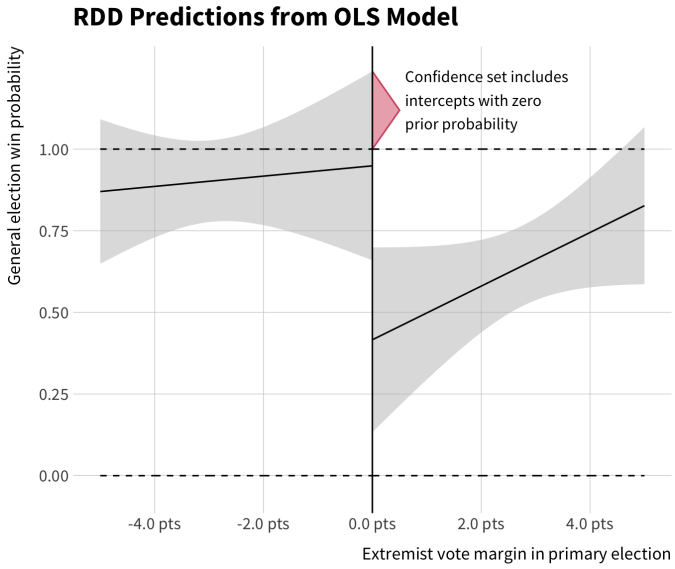

Figure 3.4: OLS estimates of the predicted probability that a candidate wins the general election. Local average treatment effect is estimated at the threshold. Treatment effect is defined by parameter estimates whose confidence sets contain values with no possibility of being true.

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

makes sense, then, to build a model that constrains those extrapolations only to regions of parameter space that are mathematically possible for the problem at hand. Furthermore, the repeated sampling intuition of the frequentist approach does not guide our inferences because the data in the analysis are the population of interest. We have no ability to repeatedly sample this data generating process, so our uncertainty about our inferences must come from some other mechanism. Most importantly, because the intercept estimates are essential for defining the treatment effect of interest, the degree to which this one estimate is corrupted presents a significant problem for the inferences we can draw from the analysis.

To visualize just how much posterior probability this model places in impossible regions of parameter space, Figure 3.5 shows a histogram of posterior samples for the treatment effect from the Bayesian version of this model using (improper) flat priors on all parameters. Because flat priors do nothing to concentrate prior probability density away from pathological regions of parameter space, a large proportion of posterior samples contain intercept estimates that do not and cannot represent win probabilities. Of the 8,000 posterior samples considered by this model fit, 36% of the non-extremist intercepts are "impossible" to obtain because they are greater than 1 or less than 0. A small number of MCMC samples for the extremist intercepts take impossible values as well. As a result, just 64% of MCMC samples for the treatment effect is composed of parameters that are mathematically possible. Even invoking the practical benefits of the LPM, such a high level of corruption in the most important quantity of this analysis is cause to rethink the approach.

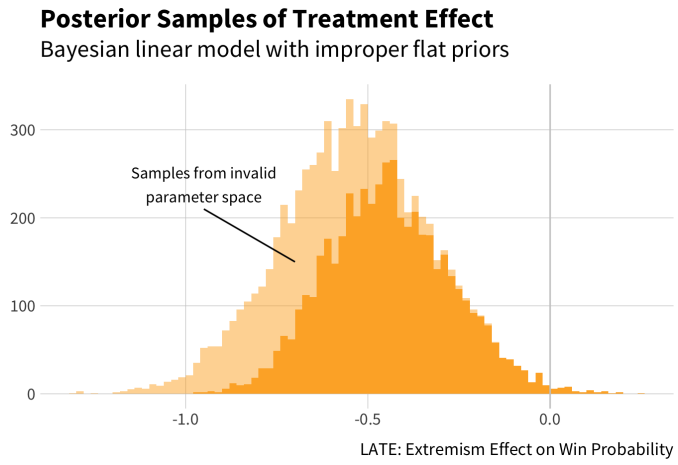

Figure 3.5: Histogram of posterior samples for the treatment effect. Using flat priors for each win probability intercept, a substantial share of the treatment distribution consists of parameters that cannot possibly occur.

The Bayesian approach begins with structural prior information about the intercepts estimated at the discontinuity. In particular, we specify a prior that these constants can only take values in the interval $[0, 1]$. We remain agnostic as to which values within that interval are more plausible in the prior. The result is a uniform prior over possible win probabilities, which we apply to both intercept parameters.

$$\alpha_{w=0}, \alpha_{w=1} \sim \text{Uniform}\,(0, 1) \tag{3.35}$$

The structural information in this prior is indisputable. We know with certainty that no probability can be less than 0 or greater than 1. Accordingly, this prior concentrates probability density away from treatment effects that cannot be true, while maintaining the local linear specification that is justified by the limiting intuition of the key identification assumption. Because we give flat priors to the individual intercepts rather than the treatment effect itself, the implied prior for the treatment effect inherits the triangular shape introduced above in Figure 3.3, which is vague despite not being flat.

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

$$\begin{aligned} y_{dpt} &\sim \text{Normal}\left(\alpha_{w[dpt]} + \beta_{w[dpt]}(\text{Extremist Primary Margin})_{dpt}, \sigma\right) \\ \beta_{\text{w}} &\sim \text{Normal}\,(0, 10) \\ \sigma &\sim \text{Uniform}(0, 10) \end{aligned} \tag{3.36}$$

The Normal model for the outcome data in the first line is equivalent to the Normal error term defined in (3.34). The priors for the β_w slopes and residual standard deviation σ are very diffuse given the scale of the outcome data, $\{0, 1\}$, and the running variable that only takes values in the interval $[-5, 5]$, a bandwidth of ±5 percentage points around the threshold.

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

possible retort by building a Bayesian logit specification as well. This setup considers the binary election result as a Bernoulli variable with a probability parameter specified by a logit model,

$$y_{dpt} \sim \text{Bernoulli}\left(\pi_{dpt}\right) \tag{3.37}$$

$$\text{logit}\left(\pi_{dpt}\right) = \alpha^*_{w[dpt]} + \beta^*_{w[dpt]}(\text{Extremist Primary Margin})_{dpt} \tag{3.38}$$

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

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

$$\begin{gathered}\text{logit}\left(\pi_{dpt}\right) = \alpha^*_{w[dpt]}, \text{ at Extremist Primary Margin}_{dpt} = 0 \\ \text{which implies } \pi_{dpt} = \text{logit}^{-1}\left(\alpha_{w[dpt]}\right)\end{gathered} \tag{3.39}$$

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

$$\alpha^*_w \sim \text{Logistic}\left(0, 1\right) \tag{3.40}$$

which becomes a flat density for $\text{logit}^{-1}\left(\alpha^*_w\right)$. It is startling at first to consider a prior as narrow as $\text{Logistic}\left(0, 1\right)$ as an uninformative prior for a key parameter. But at discussed in Section 3.3.2, the connection between prior vagueness

and prior flatness is not absolute. Flatness is only a shape. The relationship between flatness and informativeness depends on model parameterization and the scale of the data.

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

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

$$\beta_w \sim \text{Normal}\,(0, 10) \tag{3.41}$$

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

These prior interventions in both the Bayesian LPM and the Bayesian logit are minor. They merely encode structural information about the outcome scale. Win probabilities for extremists and non-extremists are constrained to take valid values—between 0 and 1—but the prior is otherwise agnostic about which win probabilities are more likely than others before seeing any data. What effect do these minor interventions have? Figure 3.7 plots the results from these three Bayesian models: the problematic original model with improper flat priors, the Bayesian LPM with structural priors to constrain the in-

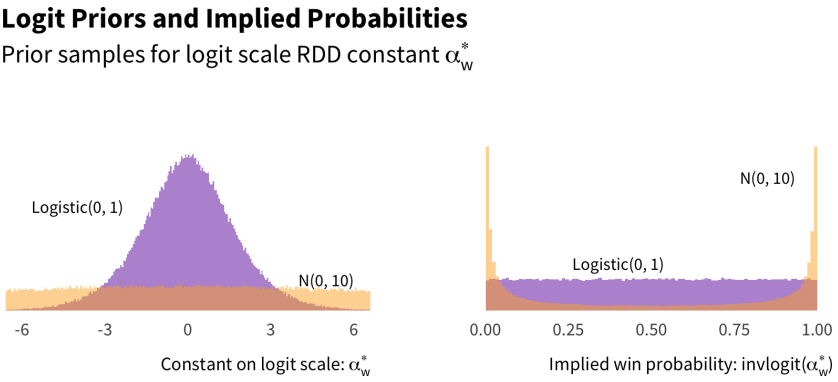

Figure 3.6: Scale invariance of logit model priors. Standard logistic prior on logit scale becomes a flat prior on the probability scale. "Diffuse" priors on logit scale imply priors on probability scale that bias toward extreme probabilities.

tercepts, and the logit model that creates the structural prior using the transformed Logistic distribution. The left panel shows a histogram of posterior MCMC samples for the non-extremist win probability at the threshold. The LPM at the top of the panel included no parameter constraints whatsoever. As a result, we see the pathological behavior where the posterior distribution places positive density on win probabilities that we know with certainty to be impossible to obtain. The histograms in the second and third rows show the LPM and logit with the structural prior. Both models concentrate prior density on possible win probabilities only, resulting in posterior distributions that reflect prior information better than the unconstrained model. The posterior distributions are asymmetric and place a lot of posterior density at high win probabilities, but this should not alarm us. The asymmetry in the distribution reflects the signals obtained from the data, rationalized against weak information encoded in the prior. The asymmetry is direct indicator of the way Bayesian priors added value to the analysis.

The right panel of Figure 3.7 shows how these parameter constrains ultimately manifest in our LATE estimates by plotting posterior means and 90 percent compatibility intervals for each model. As with nearly all Bayesian modeling approaches, our priors have the effect of shrinking important effects toward 0 and reducing the variance of the effect. In this particular case, the posterior mean for the local average treatment effect shrinks from -0.53 with flat/unconstrained priors to -0.44 using the LPM with constrained intercepts: a 17% reduction in the magnitude of the effect. The LATE from the Bayesian logit is , which is a

reduction in magnitude. This shrinkage comes from the fact that some of the largest treatment effects in our original posterior distribution were com-

posed of impossible parameters. This manifested earlier in Figure 3.5, which showed that larger treatment effects were more likely to contain pathological parameters than the smaller treatment effects. The standard deviation of the posterior samples is reduced for the models with structural prior constraints, so these prior interventions are also improving the precision of our estimates. This is because a fair amount of posterior uncertainty in the unconstrained model was owed to impossible parameter values.

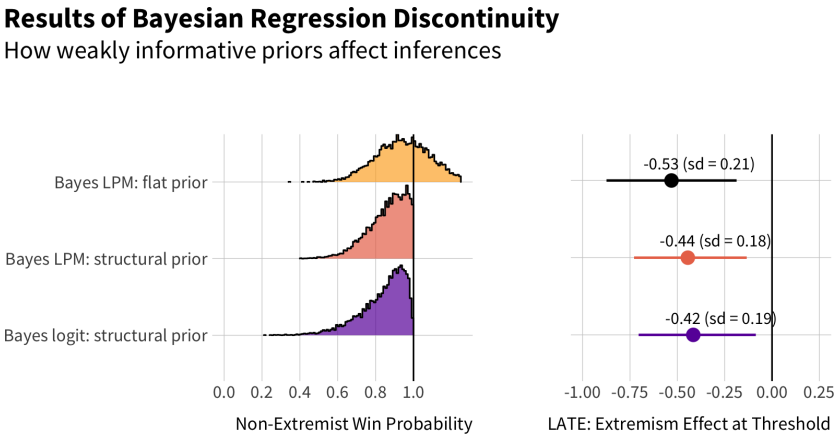

Figure 3.7: Comparison of posterior samples from Bayesian models with improper flat priors and weakly informative priors. Weakly informative models have flat priors over valid win probabilities at the discontinuity. Priors reduce effect magnitudes and variances by ruling out impossible win probabilities.

It bears emphasizing that the prior interventions in this case study were no more controversial than declaring what is already known: probabilities lie between 0 and 1. Since many causal research designs estimate treatment effects on binary variables, and many causal research designs are limited to small numbers of relevant real-world observations or budget-limited experimental samples, simple interventions like this have the potential to substantially improve the precision of research findings in contexts where researchers don't realize how much information they are leaving out of their analyses.

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

3.5 Other Frontiers of Bayesian Causal Inference

3.5.1 Beyond Estimation: Inferences About Models and Hypotheses

Inherit material from earlier section

3.5.2 Priors are the Basis for all Generalization

No-Free-Lunch theorems

3.5.3 Agnostic Causal Inference

Conventional:

$$p(\theta \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)p(\theta)}{p(\mathbf{y})} \tag{3.42}$$

$$p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta)p(\theta) \tag{3.43}$$

Implied:

$$p(\theta \mid \mathbf{y}, \mathcal{H}) = \frac{p(\mathbf{y} \mid \theta, \mathcal{H})p(\theta \mid \mathcal{H})}{p(\mathbf{y} \mid \mathcal{H})} \tag{3.44}$$

4

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

5

How District-Party Ideology Affects Primary Election Outcomes: Combining Subposteriors for Honest Bayesian Causal Inference

Group IRT Model

Colophon

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

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

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

This version of the thesis was generated on 2020-07-31 10:54:17. The repository is currently at this commit:

```
## Commit:  fffda9f5c59685feedeb5ddfacef98bb61f93dae
## Author:  Michael DeCrescenzo <mgdecrescenzo@gmail.com>
## When:    2020-07-30 00:34:26 GMT
##
##      progress in Bayesian causal model
##
## 2 files changed, 235 insertions, 134 deletions
## 30_causality.Rmd               | -134 +192  in 16 hunks
## assets-bookdown/thesis-bib.bib | -  0 + 43  in  2 hunks
```

References

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

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review* 110(3): 512–529.

Ahler, Douglas J, Jack Citrin, and Gabriel S Lenz. 2016. "Do open primaries improve representation? An experimental test of california's 2012 top-two primary." *Legislative Studies Quarterly* 41(2): 237–268.

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.

American Political Science Association, Committee on Political Parties. 1950. *Toward a more responsible two-party system*. Johnson Reprint Company.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Ansolabehere, Stephen et al. 2010. "More democracy: The direct primary and competition in us elections." *Studies in American Political Development* 24(02): 190–205.

Ansolabehere, Stephen, Jonathan Rodden, and James M Jr Snyder. 2008. "The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting." *American Political Science Review*: 215–232.

Ansolabehere, Stephen, James M Snyder, and Charles Stewart. 2001. "Candidate positioning in U.S. house elections." *American Journal of Political Science*: 136–159.

Aronow, Peter M, and Benjamin T Miller. 2019. *Foundations of agnostic statistics*. Cambridge University Press.

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

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

Barber, Michael J. 2016. "Ideological donors, contribution limits, and the polarization of american legislatures." *The Journal of Politics* 78(1): 296–310.

Barber, Michael J, Brandice Canes-Wrone, and Sharece Thrower. 2016. "Ideologically sophisticated donors: Which candidates do individual contributors finance?" *American Journal of Political Science*.

Barber, Michael, and Jeremy C Pope. 2019. "Does party trump ideology? Disentangling party and ideology in america." *American Political Science Review*: 1–17.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data." *Political analysis* 23(1): 76–91.

Bartels, Larry M. 2000. "Partisanship and voting behavior, 1952-1996." *American Journal of Political Science*: 35–50.

Bartels, Larry M. 2009. *Unequal democracy: The political economy of the new gilded age*. Princeton University Press.

Beck, Nathaniel, Gary King, and Langche Zeng. 2004. "Theory and evidence in international conflict: A response to de marchi, gelpi, and grynaviski." *American Political Science Review*: 379–389.

Betancourt, Michael. 2017. "A conceptual introduction to hamiltonian monte carlo." *arXiv preprint arXiv:1701.02434*.

Betancourt, Michael. 2019. "The convergence of markov chain monte carlo methods: From the metropolis method to hamiltonian monte carlo." *Annalen der Physik* 531(3): 1700214.

Betancourt, Michael. 2018. "Towards a principled bayesian workflow."

Betancourt, Michael, and Mark Girolami. 2015. "Hamiltonian monte carlo for hierarchical models." *Current trends in Bayesian methodology with applications* 79: 30.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. Springer.

Black, Duncan. 1948. "On the rationale of group decision-making." *The Journal of Political Economy*: 23–34.

Blackwell, Matthew, James Honaker, and Gary King. 2017. "A unified approach to measurement error and missing data: Overview and applications." *Sociological Methods & Research* 46(3): 303–341.

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

Bonica, Adam. 2019a. "Are donation-based measures of ideology valid predictors of individual-level policy preferences?" *The Journal of Politics* 81(1): 327–333.

Bonica, Adam. 2019b. "Database on ideology, money in politics, and elections: Public version 1.0."

Bonica, Adam. 2013. "Ideology and interests in the political marketplace." *American Journal of Political Science* 57(2): 294–311.

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

Bonica, Adam. 2020. "Why are there so many lawyers in congress?" *Legislative Studies Quarterly* 45(2): 253–289.

Brady, David W, Hahrie Han, and Jeremy C Pope. 2007. "Primary elections and candidate ideology: Out of step with the primary electorate?" *Legislative Studies Quarterly* 32(1): 79–105.

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

Broockman, David E. 2016. "Approaches to studying policy representation." *Legislative Studies Quarterly* 41(1): 181–215.

Broockman, David E, and Christopher Skovron. 2018. "Bias in perceptions of public opinion among political elites." *American Political Science Review* 112(3): 542–563.

Brunell, Thomas L. 2006. "Rethinking redistricting: How drawing uncompetitive districts eliminates gerrymanders, enhances representation, and improves attitudes toward congress." *PS: Political Science and Politics* 39(1): 77–85.

Brunell, Thomas L, Bernard Grofman, and Samuel Merrill. 2016. "Components of party polarization in the us house of representatives." *Journal of Theoretical Politics* 28(4): 598–624.

Bullock, Will, and Joshua D Clinton. 2011. "More a molehill than a mountain: The effects of the blanket primary on elected officials' behavior from california." *The Journal of Politics* 73(3): 915–930.

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

Burden, Barry C. 2001. "The polarizing effects of congressional primaries." *Congressional Primaries and the Politics of Representation*: 95–115.

Burden, Barry C, Gregory A Caldeira, and Tim Groseclose. 2000. "Measuring the ideologies of us senators: The song remains the same." *Legislative Studies Quarterly*: 237–258.

Butler, Daniel M, and Eleanor Neff Powell. 2014. "Understanding the party brand: Experimental evidence on the role of valence." *The Journal of Politics* 76(2): 492–505.

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

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2014. "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica* 82(6): 2295–2326.

Campbell, Angus et al. 1960. New York: John Wiley and Sons 77 *The american voter*.

Canes-Wrone, Brandice, David W Brady, and John F Cogan. 2002. "Out of step, out of office: Electoral accountability and house members' voting." *American Political Science Review* 96(01): 127–140.

Canes-Wrone, Brandice, William Minozzi, and Jessica Bonney Reveley. 2011. "Issue accountability and the mass public." *Legislative Studies Quarterly* 36(1): 5–35.

Carlson, David. 2020. "Estimating a counter-factual with uncertainty through gaussian process projection."

Carpenter, Bob et al. 2016. "Stan: A probabilistic programming language." *Journal of Statistical Software* 20: 1–37.

Caughey, Devin, and Christopher Warshaw. 2015. "Dynamic estimation of latent opinion using a hierarchical group-level irt model." *Political Analysis* 23(2): 197–211.

Caughey, Devin, and Christopher Warshaw. 2018. "Policy preferences and policy change: Dynamic responsiveness in the american states, 1936–2014."

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

Chipman, Hugh A et al. 2010. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics* 4(1): 266–298.

Clinton, Joshua D. 2006. "Representation in congress: Constituents and roll calls in the 106th house." *Journal of Politics* 68(2): 397–409.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98(02): 355–370.

Cohen, Marty et al. 2009. *The party decides: Presidential nominations before and after reform*. University of Chicago Press.

Cox, Gary W. 1990. "Centripetal and centrifugal incentives in electoral systems." *American Journal of Political Science*: 903–935.

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

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

Duane, Simon et al. 1987. "Hybrid monte carlo." *Physics letters B* 195(2): 216–222.

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

Epstein, Lee et al. 2007. "The judicial common space." *Journal of Law, Economics, and Organization* 23(2): 303–325.

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

Fiorina, Morris P, Samuel J Abrams, and Jeremy C Pope. 2005a. *Culture war? The myth of a polarized america*. Pearson Longman New York.

Fiorina, Morris P, Samuel J Abrams, and Jeremy C Pope. 2005b. *Culture war? The myth of a polarized america*. Pearson Longman New York.

Fowler, Anthony, and Andrew B Hall. 2016. "The elusive quest for convergence." *Quarterly Journal of Political Science* 11: 131–149.

Gabry, Jonah et al. 2019. "Visualization in bayesian workflow." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182(2): 389–402.

García-Pérez, Miguel Ángel. 2019. "Bayesian estimation with informative priors is indistinguishable from data falsification." *The Spanish journal of psychology* 22.

Geer, John G. 1988. "Assessing the representativeness of electorates in presidential primaries." *American Journal of Political Science*: 929–945.

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

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. "The prior can often only be understood in the context of the likelihood." *Entropy* 19(10): 555.

Gelman, Andrew et al. 2013. *Bayesian data analysis*. Chapman; Hall/CRC.

Gerring, John. 2001. *Social science methodology: A criterial framework*. Cambridge University Press.

Gilens, Martin, and Benjamin I. Page. 2014. "Testing theories of american politics: Elites, interest groups, and average citizens." *Perspectives on Politics* 12(3).

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

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

Green, Donald P, and Holger L Kern. 2012. "Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees." *Public opinion quarterly* 76(3): 491–511.

Green, Donald P et al. 2016. "The effects of lawn signs on vote outcomes: Results from four randomized field experiments." *Electoral Studies* 41: 143–150.

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

Grosz, Michael P, Julia M Rohrer, and Felix Thoemmes. 2020. "The taboo against explicit causal inference in nonexperimental psychology."

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

Hacker, Jacob S, Paul Pierson, and others. 2005. *Off center: The republican revolution and the erosion of american democracy*. Yale University Press.

Hahn, P Richard et al. 2018. "Regularization and confounding in linear regression for treatment effect estimation." *Bayesian Analysis* 13(1): 163–182.

Hall, Andrew B. 2015. "What happens when extremists win primaries?" *American Political Science Review* 109(01): 18–42.

Hall, Andrew B, and James M Snyder. 2015. "Candidate ideology and electoral success. Working paper: Https://dl. Dropboxusercontent.com/u/11481940/hall snyder ideology.pdf."

Hall, Andrew B, and Daniel M Thompson. 2018. "Who punishes extremist nominees? Candidate ideology and turning out the base in us elections." *American Political Science Review* 112(3): 509–524.

Henderson, John A. 2016. "An experimental approach to measuring ideological positions in political text." *Available at SSRN 2852784*.

Hernán, Miguel A. 2018. "The c-word: Scientific euphemisms do not improve causal inference from observational data." *American journal of public health* 108(5): 616–619.

Hill, Jennifer L. 2011. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics* 20(1): 217–240.

Hill, Seth J. 2015. "Institution of nomination and the policy ideology of primary electorates." *Quarterly Journal of Political Science* 10(4): 461–487.

Hill, Seth J, and Gregory A Huber. 2017. "Representativeness and motivations of the contemporary donorate: Results from merged survey and administrative records." *Political Behavior* 39(1): 3–29.

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

Hirano, Shigeo, and Michael M Ting. 2015. "Direct and indirect representation." *British Journal of Political Science* 45(3): 609.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81(396): 945–960.

Horiuchi, Yusaku, Kosuke Imai, and Naoko Taniguchi. 2007. "Designing and analyzing randomized experiments: Application to a japanese election survey experiment." *American Journal of Political Science* 51(3): 669–687.

Imbens, Guido W, and Donald B Rubin. 1997. "Bayesian inference for causal effects in randomized experiments with noncompliance." *The annals of statistics*: 305–327.

Jackman, Simon. 2009. 846 *Bayesian analysis for the social sciences*. John Wiley & Sons.

Jacobson, Gary C. 2012. "The electoral origins of polarized politics: Evidence from the 2010 cooperative congressional election study." *American Behavioral Scientist* 56(12): 1612–1630.

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

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

Key, Valdimer Orlando. 1955. "Politics, parties, and pressure groups."

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

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

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

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

Lax, Jeffrey R, and Justin H Phillips. 2009. "How should we estimate public opinion in the states?" *American Journal of Political Science* 53(1): 107–121.

Layman, Geoffrey C., and Thomas M. Carsey. 2002. "Party Polarization and "Conflict Extension" in the American Electorate." *American Journal of Political Science* 46(4): 786. `http://www.jstor.org/stable/3088434?origin=crossref` (Accessed February 22, 2015).

Layman, Geoffrey C et al. 2010. "Activists and conflict extension in american party politics." *American Political Science Review*: 324–346.

Lebo, Matthew J, Adam J McGlynn, and Gregory Koger. 2007. "Strategic party government: Party influence in congress, 1789–2000." *American Journal of Political Science* 51(3): 464–481.

Levendusky, Matthew. 2009. *The partisan sort: How liberals became democrats and conservatives became republicans*. University of Chicago Press.

Levendusky, Matthew S, Jeremy C Pope, and Simon D Jackman. 2008. "Measuring district-level partisanship with implications for the analysis of us elections." *The Journal of Politics* 70(3): 736–753.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. "Generating random correlation matrices based on vines and extended onion method." *Journal of multivariate analysis* 100(9): 1989–2001.

Liao, Shirley, Lucas Henneman, and Corwin Zigler. 2019. "Posterior predictive treatment assignment methods for causal inference in the context of time-varying treatments." *arXiv preprint arXiv:1907.06567*.

Liao, Shirley X, and Corwin M Zigler. 2020. "Uncertainty in the design stage of two-stage bayesian propensity score analysis." *Statistics in Medicine*.

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

MacKay, David JC. 1992. "A practical bayesian framework for backpropagation networks." *Neural computation* 4(3): 448–472.

Mann, Thomas E. 1978. 220 *Unsafe at any margin: Interpreting congressional elections*. Aei Pr.

Martin, Andrew D, and Kevin M Quinn. 2002. "Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999." *Political Analysis* 10(2): 134–153.

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

Mayhew, David R. 1974. *Congress: The electoral connection*. Yale University Press.

McCarty, Nolan, and Howard Poole Keith T. and Rosenthal. 2006. *Polarized america: The dance of ideology and unequal riches*. Cambridge, MA: MIT Press.

McCarty, Nolan, Keith T Poole, and Howard Rosenthal. 2009. "Does gerrymandering cause polarization?" *American Journal of Political Science* 53(3): 666–680.

McClosky, Herbert, Paul J Hoffmann, and Rosemary O'Hara. 1960. "Issue conflict and consensus among party leaders and followers." *The American Political Science Review* 54(2): 406–427.

McElreath, Richard. 2017a. "Bayesian inference is just counting."

McElreath, Richard. 2017b. "Bayesian statistics without frequentist language."

McElreath, Richard. 2020. *Statistical rethinking: A bayesian course with examples in r and stan*. 2nd ed. CRC press.

McGann, Anthony J. 2014. "Estimating the political center from aggregate data: An item response theory alternative to the stimson dyad ratios algorithm." *Political Analysis*: 115–129.

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

Meager, Rachael. 2019. "Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments." *American Economic Journal: Applied Economics* 11(1): 57–91.

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

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

Oganisian, Arman, and Jason A Roy. 2020. "A practical introduction to bayesian estimation of causal effects: Parametric and nonparametric approaches." *arXiv preprint arXiv:2004.07375*.

Ornstein, Joseph T, and JBrandon Duck-Mayr. 2020. "Gaussian process regression discontinuity."

Pacheco, Julianna. 2011. "Using national surveys to measure dynamic us state public opinion: A guideline for scholars and an application." *State Politics & Policy Quarterly*: 1532440011419287.

Park, David K, Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian multilevel estimation with poststratification: State-level estimates from national polls." *Political Analysis* 12(4): 375–385.

Park, Trevor, and George Casella. 2008. "The bayesian lasso." *Journal of the American Statistical Association* 103(482): 681–686.

Pearl, Judea. 1995. "Causal diagrams for empirical research." *Biometrika* 82(4): 669–688.

Pearl, Judea. 2009. *Causality: Models, reasoning, and inference*. 2nd ed. Cambridge University Press.

Petrocik, John Richard. 2009. "Measuring party support: Leaners are not independents." *Electoral Studies* 28(4): 562–572. `http://linkinghub.elsevier.com/retrieve/pii/S0261379409000511` (Accessed April 16, 2015).

Phillips, Anne. 1995. *The politics of presence*. Clarendon Press.

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

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

Porter, Rachel A, and Sarah Treul. 2020. "Reevaluating experience in congressional primary elections."

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

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

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

Ratkovic, Marc, Dustin Tingley, and others. 2017. "Sparse estimation and uncertainty with application to subgroup analysis." *Political Analysis* 25(1): 1–40.

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.

Rosenbaum, Paul R. 2002. *Observational studies*. Springer New York. https://doi.org/10.1007/978-1-4757-3692-2.

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

Rubin, Donald B. 2005. "Causal inference using potential outcomes: Design, modeling, decisions." *Journal of the American Statistical Association* 100(469): 322–331.

Rubin, Donald B. 1980. "Comment on randomization analysis of experimental data: The fisher randomization test by d. Basu." *Journal of the American statistical association* 75(371): 591–593.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66(5): 688.

Rubin, Donald B. 1981. "Estimation in parallel randomized experiments." *Journal of Educational Statistics* 6(4): 377–401.

Samii, Cyrus, Laura Paler, and Sarah Zukerman Daly. 2016. "Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia." *Political Analysis* 24(4): 434–456.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12(1): 487–508. `http://www.annualreviews.org/doi/abs/10.1146/annurev.polisci.11.060606.135444` (Accessed January 13, 2015).

Shor, Boris, and Nolan McCarty. 2011. "The ideological mapping of american legislatures." *American Political Science Review* 105(03): 530–551.

Sides, John et al. 2018. "On the representativeness of primary electorates." *British Journal of Political Science*: 1–9.

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

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

Stimson, James A. 1991. *Public opinion in america: Moods, cycles, and swings*. Westview Press.

Stokes, Donald E. 1963. "Spatial models of party competition." *The American Political Science Review* 57(2): 368–377.

Tausanovitch, Chris, and Christopher Warshaw. 2017. "Estimating candidates' political orientation in a polarized congress." *Political Analysis* 25(2): 167–187.

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

Thomsen, Danielle M. 2014. "Ideological moderates won't run: How party fit matters for partisan polarization in congress." *The Journal of Politics* 76(3): 786–797.

Thomsen, Danielle M. 2020. "Ideology and gender in us house elections." *Political Behavior* 42(2): 415–442.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1): 267–288.

Tomz, Michael, and Robert P Van Houweling. 2008. "Candidate positioning and voter choice." *American Political Science Review*: 303–318.

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

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical bayesian model evaluation using leave-one-out cross-validation and waic." *Statistics and computing* 27(5): 1413–1432.

Vehtari, Aki et al. 2020. "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC." *Bayesian Analysis*.

Warshaw, Christopher, and Jonathan Rodden. 2012. "How should we measure district-level public opinion on individual issues?" *The Journal of Politics* 74(01): 203–219.