Bayesian Causal Inference in Political Science

Michael G. DeCrescenzo[*]

This version: April 12, 2019.[†]

Abstract

Causal inference and Bayesian analysis are two powerful and attractive methodological approaches for conducting empirical research, but rarely in political science does a single study employ both approaches. This stands in contrast to other fields—such as psychology, epidemiology, and biostatistics—where Bayesian and causal methods are more commonly applied together. In this paper I argue that the partition between these methodological schools in political science has no inherent basis in their fundamental goals, which are actually quite compatible: generating the best parameter estimates. In fact, Bayesian analysis provides a number of distinct benefits for estimating statistical models for causal inference. The methodological partition instead owes itself to informal norms surrounding each school in empirical political science. I discuss these sources of normative tension, discuss go-to practices doing Bayesian inference for the "skeptical causal inference audience," and demonstrate these practices using real examples from recent political science work. The purpose of the paper is *not* to convince all causal inference practitioners to adopt Bayesian estimation. The purpose is to show that Bayesian methods deserve space in the study of causal effects because they improve causal estimates and provide an appealing framework for evaluating causal evidence.

☺ This paper is aggressively unfinished ☺
☺ Please do not cite without permission or circulate recklessly ☺

[*] Ph.D. Candidate, Political Science, University of Wisconsin–Madison. I have received plenty of great feedback and advice from Matthew Blackwell, Barry Burden, William Christiansen, Andrew Heiss, Devin Judge-Lord, Michael Masterson, Anna Meier, Laura Meinzen-Dick, Evan Morier, Ellie Powell, Daniel Putman, Alex Tahk, Zach Warner, Chagai Weiss, Jessica Weeks, and Dave Weimer.

[†] View most recent committed version here.

Contents

1 Introduction

“Causal empiricism” (Samii 2016) and Bayesian analysis are two growing schools of methodological work in applied political science. Despite the value of each, studies that employ both methodological approaches are exceedingly rare. This is despite the fact that Bayesian analysis and causal inference are regular companions in other fields such as psychology, epidemiology, and biostatistics. It is also despite the fact that political scientists in both schools regularly describe the benefit of their school using language and examples that borrow from the other. Introductory discussions of Bayesian inference often use experiments to demonstrate the intuition of Bayesian updating, incorporating prior information from previous studies to improve the precision of estimates obtained from new data.[1] Advocates of causal inference in turn use Bayesian language to argue that causal modeling provide stronger information for researchers to update their priors about causal effects, since the researcher has a strong prior that an experimental study contains zero or minimal confounding bias.[2]

field cites

Why, then, is “Bayesian causal inference” so rare in political science? This paper argues this division in methodological approaches is needless. Bayesian analysis and causal inference are fundamentally compatible and mutually enhancing. I provide both a positive argument for this perspective and a negative argument against the perceived tensions between these two schools. Although each school uses different tactics, both are united in the goal of getting the best parameter estimates possible. Further, the advantages of one school do not necessarily conflict with the advantages of the other. Tensions arise from the informal norms surrounding the applied work in each school rather than the formal structure of the underlying methods. The division between schools is artificial in the least, if not outright counter-productive to the progress of causal empiricism in political science.

Jstor

After laying out the “positive” and “negative” arguments, I put forth practical advice for doing Bayesian causal inference in “ordinary” empirical contexts, respecting the skeptical norms within causal empiricism while showcasing the theoretical and practical value of Bayesian approaches. Many Bayesian modeling practices are fundamentally conservative—accounting for all sources of parameter uncertainty, regularizing estimates against over-fitting, and down-weighting parameter values that make no sense in the context of the problem—and

[1] A common example case in Bayesian pedagogy is an experiment of parallel randomized experiments across eight schools by Rubin (1981; see also Gelman et al. 2013).

[2] Indeed, the “illusion of learning from observational research” (Gerber, Green, and Kaplan 2004) is an argument that comes from a formal Bayesian model, and the namesake result is only obtained through the particular choice of priors in the model.

should appeal to causal empiricists. Furthermore, Bayesian inference has philosophical and practical advantages that should appeal to causal empiricists even when treatment effects are not guaranteed to be dramatically affected by the Bayesian model. I then implement these points of advice in two replicated analyses: a regression discontinuity study of ideologically extreme House candidates (Hall 2015) and an experimental public opinion toward unilateral actions by U.S. presidents (Reeves et al. 2017).

My goal is not to convince all causal inference practitioners to use Bayesian methods, nor is it to assert the "superiority" of Bayesian methods. Rather, the objective of the paper is to create space for Bayesian methods in political science's push toward rigorous causal approaches. Bayes and causal inference can make each other better even in "ordinary" data analysis scenarios, but these improvements should be demonstrated, and areas of potential disagreement or misunderstanding should be adjudicated.

Throughout the paper I refer to "causal inference" as a category of research designs that allow the estimation of a parameter with a causal interpretation justified by an identification analysis. This includes randomized experiments as well as model-driven designs such as difference-in-differences, instrumental variables, synthetic control, regression discontinuity, and others. I refer to "Bayesian analysis" as statistical procedures that include models for unobserved parameters (prior distributions) alongside models for observed data (likelihood functions). For the time being I avoid coining any term or initialism for Bayesian causal inference because, as I argue below, estimating causal parameters Bayesianly is neither new nor especially remarkable.[3] I assume that a reader of this paper has some conceptual familiarity with either causal inference or Bayesian analysis, but not necessarily both. I provide some background information and notation to ground the discussion as it progresses, but readers looking for comprehensive introductions might consult Angrist and Pischke (2008), Imbens (2004), Imbens and Rubin (2015), or Rubin (2008) for causal inference, and Gelman et al. (2013), Jackman (2009), or McElreath (2015) for Bayesian analysis.

literal

audience?

2 Shared Goals, Different Tactics

Causal inference and Bayesian analysis can "go together" in political science. This is because both schools are fundamentally interested in the same thing: obtaining the best, most reliable parameter estimates that we can get, given the data. Causal inference and Bayesian analysis employ different tactics to achieve this goal, which I discuss below, but it is valuable to

[3] Explicit Bayesian treatments of potential outcomes modeling can be found as far back as Rubin (1978).

construe their goals this way at the outset.

We might ask whether “obtaining good parameter estimates” is a trivial goal. Don’t all statistical methodologies make parameter estimation their primary focus? I think the answer is No. Take null hypothesis significance testing (NHST) as an example. NHST is the prevailing framework for drawing inferences about hypotheses in political science. It operates by measuring the plausibility of the data, given an assumed null hypothesis. Point estimates from statistical models are judged to be “statistically significant” if they were sufficiently unlikely to have occurred by chance ($p < \alpha$) under the null hypothesis. NHST is a protocol for making inferences about hypotheses, but it does not work by making substantive inferences about what parameters actually are. Instead it assumes that the true parameter is a value that was never actually estimated and (if zero) rarely represents a legitimate competing hypothesis, and then it calculates the probability of the data under this separate hypothesis. The ordinary usage of NHST does not guide a researcher’s inference about which non-null parameters are better, since the null is rejected in favor of a poorly specified alternative. Scholarly work on nonparametric NHST underscores this point by describing the researcher’s quantity of interest as “the certainty of the causal inference”—the p-value—rather than the causal parameter itself (Keele, McConnaughy, and White 2012).[4] It does make sense, then, to highlight that causal inference and Bayesian analysis share a fundamental desire to generate the best parameter estimates because they stand in contrast to the dominant inference paradigm in most empirical social science.[5]

What are the “best” parameter estimates? Statisticians and methodologists formalize this concept using loss functions, decision rules about bias and error (e.g. admissibility), and so on. For applied work, we care about answering our research questions with a sufficient degree of confidence. Causal inference increases our confidence by de-biasing the research design. Bayesian analysis increases our confidence by returning a posterior distribution with value added from prior information. I discuss both of these contributions in turn, and then I briefly discuss a Bayesian potential outcomes model originally described by Rubin (1978).

admissible

[4] Point estimates and confidence intervals allow narrower inferences about parameters, but this isn’t how NHST operates. Furthermore, if confidence intervals are interpreted as ranges of parameter values that are supported by the data, this is already Bayesian inference.

[5] Even if a causal model is interpreted under using NHST, the causal model is its own thing. Furthermore, NHST can be also conducted using Bayesian models, so this discussion of NHST should not be read as a tirade against non-Bayesian analysis!

2.1 *Causal Modeling: Better Parameters "by Design"*

To achieve the best parameter estimates, causal inference methods develop or implement careful research designs that (i) identify a causal effect as an inherent feature of the design and (ii) the type of causal effect is as narrowly defined as the design can justify. These research designs typically are rooted in a model of the "potential outcomes" for units in the data (Rubin 1974), and estimated with a statistical model where the variation in the independent variable identifies the effect in the potential outcomes model.

Causal empiricism formalizes causal claims with a carefully specified causal model instead of an informal, verbal discussion of control variables and potential confounders. Formal causal models in political science typically use potential outcomes notation, a simple example being the case where a unit i is assigned a treatment states $Z_i \in \{0, 1\}$, with 0 indicating control and 1 indicating treatment. The outcome $Y_i(z)$ is unit individual i's potential outcomes given $Z_i = z$, with $Y_i(z = 1)$ being the potential outcome under treatment and $Y_i(z = 0)$ being the potential outcome under control. The unit-level treatment effect T_i is the difference between i's potential outcomes: $T_i \equiv Y_i(1) - Y_i(0)$. T_i can never be directly measured, however, because i can only be assigned to one treatment level at a time (Holland 1986). As such, causal models require a set of assumptions to estimate the treatment effect from data. pace?

The research designs derived from causal models identify parameters "by design," meaning that a causal interpretation of the estimated parameter flows necessarily from the research design as long as its assumptions are satisfied. These designs are particularly attentive to the mechanism that assigns units to a treatment status, since the model of the assignment mechanism determines which causal estimands can be identified from the study design (Rubin 1991). Identification strategies ideally seek treatment assignment mechanisms that are uncorrelated with the units' potential outcomes. The canonical method for obtaining unconfounded treatment status is random assignment into treatment. In situations where assignment is not perfectly uncorrelated, treatment uptake is imperfect, or when the causal structure is indirect or multifaceted, researchers have explored the assumptions required to identify local treatment effects, conditional treatment effects, intent-to-treat effects, treatment-on-treated effects, direct and indirect effects, and marginal component effects. cites

Parameter estimates from causal inference designs are "better" because the researcher has confidence that the estimates truly represent a causal effect. The causal interpretation is justified by an identification analysis (Keele 2015; Matzkin 2007) that explicates the minimal set of assumptions required to form this causal interpretation. This is in contrast to regression models that "merely" estimate a conditional mean of Y and are are routinely discussed with vague causal language that is difficult to justify on a "selection-on-observables" identifying assumption. As a result, causal parameters avoid both internal and external validity problems that plague observational research including model misspecification and generalizing beyond common support. When the researcher is confident that bias in a study is small, the researcher is able to update their priors about the causal effect more decisively than in an observational setting where bias could be more pervasive (Gerber, Green, and Kaplan 2004).

2.2 Bayesian Analysis: Better Parameters with Priors and Posteriors

To achieve the best parameter estimates, Bayesian methods use the posterior distribution. The posterior distribution conveys which parameter values are likely, and which are unlikely, having seen the data. The posterior distribution is unique to Bayesian analysis and has distinctive philosophical and practical benefits. Philosophically, the posterior allows researchers to make direct inferences about which parameters are consistent with the data and with what probability. This can only be done indirectly in non-Bayesian inference; confidence intervals cover the true parameter only in "95% of infinitely repeated samples," whereas Bayesian intervals (called "credible" or "compatibility" intervals) have a clear meaning even in one-off samples that are never repeated. The posterior distribution reflects uncertainty over all parameters both marginally and conditionally, with no need for post-hoc standard error corrections or approximations that collapse uncertainty for nuisance parameters. The posterior can also be improved with the specification of the prior distribution, which is discussed more below. Practically, Bayesian models can naturally handle multilevel data structures, missing data, and mixture processes, and they make it easy for researchers to calculate and visualize model estimates with posterior samples.

Mechanically, Bayesian models specify a joint probability distribution over all variables in the model, where a "variable" could be observed data (represented as y) or unobserved parameters (represented as θ). Specifying a full probability model for data and parameters "merely" applies the same intuition to parameters as is routinely applied to data: we can't perfectly predict every instantiation of the process that produces data (or parameters), but we use a probability distribution to represent our estimates of that process. This joint distribution begins with a data model that is a function of model parameters, $p(y \mid \theta)$, and a prior distribution over possible parameter values, $p(\theta)$. This joint distribution can be represented as

$$p(y, \theta) = p(y \mid \theta)p(\theta). \tag{1}$$

The prior distribution $p(\theta)$ inevitably contains parameters that would be unsupported by the data as more data are collected. Researchers learn which parameters are consistent with the data by conditioning the joint distribution on the data using Bayes' Theorem.

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \tag{2}$$

Conditioning on the data allows the researcher to "update" the distribution of parameters. This updated distribution, the posterior distribution, is the represents the parameter values that are most compatible with both the initial model and the data. The exact shape of the posterior is determined by the specificity of the prior and the strength of the signal from the data. When the prior is flat or the data contain a precise signal about likely and unlikely parameters, the posterior more heavily reflects the data, $p(y \mid \theta)$. When the researcher observes enough data, the prior approaches irrelevance.[6] Conversely, when the prior distribution is more precise or the data send a weak signal, the posterior distribution retains more of the prior. Crucially, the intuition of Bayesian updating is agnostic to whether the original model has a causal interpretation; if any model has parameters that can be conditioned on data, then posterior parameter inference is possible.

All of the philosophically appealing intuitions of the posterior hold even when the prior distribution for parameters $p(\theta)$ is vague or flat. Posterior inference can be improved even further by specifying informative priors about the plausible parameter values in the model. Understandably, the issues surrounding priors is where Bayesian analysis is most commonly misunderstood, so I hope to clear up some of these misunderstandings before directly confronting the intersection of Bayes with causal inference.

First, Bayesian analysis considers a joint probability distribution of parameters and data. This means that Bayesian statistics regards parameters as "random" variables. This terminology can be confusing, especially from a non-Bayesian perspective where the true parameter is thought to be an unknown but "fixed" value. This idea of a fixed parameter works in Bayesian analysis too; the difference is that our uncertainty about its exact value is represented using a probability distribution. The parameter is not random in the sense that it

[6] As long as the prior does not assign zero probability to important regions of parameter space.

is fluctuating between ever-changing quantum states. Rather, it is random in the sense that it may be an instantiation of some underlying process that we cannot describe exactly. This is no different from the way we describe data as draws from probability distributions. We can't exactly predict an individual data point, but we have good justification to assume that the accumulation of forces that determine its exact value resemble a probability distribution. Indeed, likelihood functions are exactly like priors for data; the only difference is that "data" is the word we use for a piece of the model that we observe, while "parameter" is a word we use for a piece of the model that is unobserved.[7]

Non-Bayesians are often skeptical of using priors to "stack the deck" in favor of a predetermined conclusion rather than letting the data speak. Luckily, priors are not employed in such a way. Most of the time priors are used to downweight implausible parameters (e.g. coefficients or variance terms that exceed the range of the outcome variable) rather than upweight the researcher's preferred hypothesis. Nearly all modeling problems have natural constraints on the possible parameters values. These natural constraints can be incorporated into the model with "weakly informative priors," which I discuss more in Section 4 (see also Gelman 2006). Flat priors, by contrast, have the primary effect of *upweighting* highly implausible parameter values. Flat priors regard a coefficient of 10 million as equally likely to a coefficient of 1, but a researcher would throw any model away that returned an estimate of 10 million for most outcome scales. By this logic, it is difficult to imagine too many situations where we can't do better than a flat prior.

puffery

notation

elab?

Mike B?

2.3 *Potential Outcomes as Posterior Distributions*

Causal inference improves parameter estimation by de-biasing the research design, and Bayesian inference improves parameter estimation by incorporating prior information to obtain a posterior distribution of parameter values. How would we combine the two perspectives? This section discusses a Bayesian potential outcomes model originally laid out by Rubin (1978) but simplified for this paper. Next is a generic, reduced-form example of an experiment estimated using Bayesian priors.

A key feature of Bayesian analysis is that unknown quantities in our model are represented with probability distributions that represent our uncertainty. Functions of unknown variables, in turn, reflect uncertainty about variables that compose it. This is relevant to causal inference because it changes our interpretation of the unobserved potential outcomes. An unobserved

[7] Richard McElreath, "Understanding Bayesian statistics without frequentist language." Talk delivered at Bayes@Lund2017. Accessed via https://www.youtube.com/watch?v=yakg94HyWdE

potential outcome $\tilde{y}_i$ in a Bayesian framework is represented by a distribution of unknown potential outcomes that reflects posterior uncertainty about model parameters. We obtain the posterior distribution of the unobserved outcome by conditioning on the data. Its marginal posterior distribution is its joint distribution with the updated model parameters, integrating over the parameter values.

$$p(\tilde{y}_i \mid y) = \int p(\tilde{y}_i, \theta \mid y) d\theta \tag{3}$$

This means that we would construe the unit-level treatment effect $T_i = y_i - \tilde{y}_i$ as an expectation that averages over the posterior uncertainty in the model parameters. int

$$p\,(T_i \mid y) = \int p\,(y_i - \tilde{y}_i \mid y)\, d\theta \tag{4}$$

In English, the posterior distribution of the treatment effect is the distribution of differences between the observed and (posterior) unobserved potential outcomes, marginalizing over the model parameters.

Why is the Bayesian potential outcomes setup appealing? Because parameters are given probability distributions, our epistemic uncertainty about the causal effect is an explicit feature of the potential outcomes model rather than a byproduct of estimation under sampling error. By treating unobserved potential outcomes as conditional on the observed data, the Bayesian model effectively is a missing data model for potential outcomes (Rubin 1978). Unobserved potential outcomes can be simulated directly from the posterior distribution and evaluated during posterior predictive checking (an important part of the Bayesian workflow, see Gabry et al. 2019). This provides natural scaffolding to extend causal models with more complex design features such as missing data, non-compliance, and more (e.g. Horiuchi, Imai, and Taniguchi 2007).

Zooming out from the potential outcomes model, we can see what an experiment would look like if it were analyzed from a Bayesian perspective. Consider an example where individuals are randomly assigned to values of $z \in \{0, 1\}$. The difference of means is commonly estimated using a regression form…

$$y_i = \alpha + \beta z_i + \varepsilon_i, \tag{5}$$

where α is the control group mean, β is the difference in the means of each group,[8] and ε_i is

[8] Sometimes the difference in the means of the treatment and control group is not exactly the same as the

response-level error. For Bayesian estimation, we would specify the full probability model for all observed data and unobserved parameters. First, the response data are given a probability model implied by the parametric assumption of ε_i. (This is not Bayesian.)

$$y_i \sim \text{Normal}\left(\alpha + \beta z_i, \sigma_{z[i]}\right) \tag{6}$$

This example assumes unequal variance across levels of z. We would then include priors for the model parameters, which I will black box with generic $p(\cdot)$ statements..

$$\alpha \sim p(\cdot) \qquad \beta \sim p(\cdot) \qquad \sigma_z \sim p(\cdot) \tag{7}$$

A deeper discussion about priors can be found in Section 4. To fit the model, we condition the parameters on the data and achieve posterior distributions for α, β, and σ. If the priors are flat, then the posterior distribution for the parameters will be proportional to the likelihood. Otherwise, the posterior distribution will retain some contribution from the prior. All the parameters have the same meaning as in a non-Bayesian model; β still represents the causal parameter, but we have estimated it using a prior. It is ultimately a modest intervention on the way experiments are ordinarily interpreted.

Digging a little deeper, one practical consideration for Bayesian estimation is the parametric form of the regression, since parameterization affects the intuition of the priors.[9] By estimating a treatment effect with a dummy variable, it is very difficult to place a prior on a difference in means that leads to equal priors for both group means. For this reason, it is more straightforward to write the conditional mean of y_i explicitly as the mean in each treatment group, μ_z:

$$y_i \sim \text{Normal}\left(\mu_{z[i]}, \sigma_z\right). \tag{8}$$

Under this setup, the average treatment effect is $\beta = \mu_{z=1} - \mu_{z=0}$. From that starting point, it is straightforward to place equal priors on both treatment groups.

mean of the unit-level treatment effects. This can depend on the link function but also whether a researcher wishes to distinguish sampling-based or "design-based" uncertainty (Abadie et al. 2017).

[9] This is also not Bayesian, since likelihood-based models have implicitly flat priors that have different implications on different transformed scales. Flat priors on the log-odds scale will behave differently from flat priors in a linear probability model even if the data are the same. Non-Bayesian models don't avoid the issue of priors; they only ignore it.

2.4 Bayesian Opportunities

There are plenty more things a researcher could consider when estimating a causal model with a Bayesian approach. These are circumstances when we have knowledge about how a model should behave that the model is not guaranteed to learn from the data.

- **Bounds:** We often deal with parameters that have natural bounds. Variance parameters will be positive. Probabilities will not exceed 0 or 1.
- **Scale:** If we know the scale of an outcome variable, we know a lot about what parameters we could expect from the model that generates that data. Standardization can aid.
- **Multilevel data:** Lots of datasets have hierarchical structure. In experimental and causal inference contexts this could be included by clustered randomization, repeated observations within-unit, correlated error across time periods or within clusters of data, and more. Bayesian models allow the researcher to directly estimate the variance at each hierarchical level, sidestepping the need for post-hoc standard error corrections.
- **Regularization:** Some experiments use several treatments and either compute marginal treatment effects or complex interactions. As the number of interesting comparisons multiply, priors can be used to do partial pooling, regularize estimates against noisy comparisons, and counter the "multiple comparisons" problem.

Political scientists have put Bayesian analysis to work in plenty of observational work for measurement (Clinton, Jackman, and Rivers 2004; Kernell 2009; Park, Gelman, and Bafumi 2004), text data (Grimmer 2010), heterogeneous effects (Western 1998), meta-analysis and model averaging (Montgomery and Nyhan 2010), time series (Brandt and Freeman 2006), and time-series cross-sectional data (Shor et al. 2007). It is far less common in experimental and observational causal inference work, but not truly absent. Bayesian methods have been employed to tackle heterogeneous treatments (Green and Kern 2012; see also Hill 2011), modeling noncompliance and nonresponse (Horiuchi, Imai, and Taniguchi 2007), multi-stage models such as instrumental variables (Hollenbach, Montgomery, and Crespo-Tenorio 2018), and conjoint experiments (Jensen et al. 2019). Other researchers outside of political science have explored Bayesian regression discontinuity outside of political science (e.g. Chib and Jacobi 2016; Branson et al. 2019)

This paper uses two examples from published political science research to demonstrate how these "Bayesian opportunities" arise and how a researcher might safely address them. I will preview these cases now and implement Bayesian approaches later in Section 5.

be nice

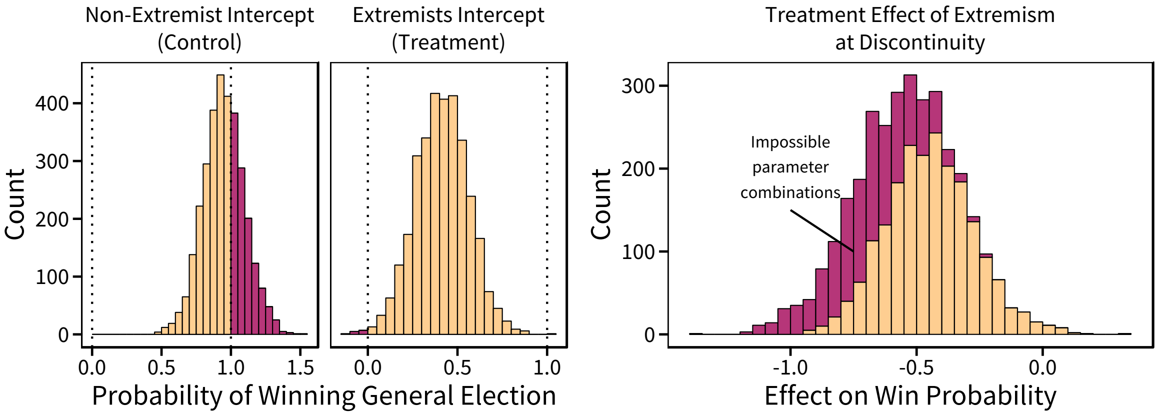

Figure 1: Histogram of posterior samples from Hall's (2015) regression discontinuity design, recreated using a Bayesian model with flat priors. Dashed lines in left two panels indicate the region of possible parameter values.

First is a regression discontinuity study examining the effect of ideological extremism in U.S. House elections Hall (2015). When a primary election between an extremist and a non-extremist candidate is decided by a "coin-flip" result, the discontinuity provides identifying variation in candidate ideology for the general election. The author finds that when the extremist candidate wins the primary, they are less likely to win the general election by 53 percentage points. This result is obtained from a linear probability model (LPM), which is often preferred because it is simple and can recover unbiased treatment effects at the discontinuity under the right assumptions (Lee and Lemieux 2010). In this case, we know that the probability of winning the general election is bounded at 0% and 100%, but the model does not. Researchers often accept this "unrealistic" behavior of LPMs because they are easy to interpret. This is understandable, but if one of the advantages of causal inference is to be realistic (Samii 2016), it is fair to say that the treatment effect is important enough that it should not be contaminated by problematic assumptions (or lack of assumptions) in any model. In this particular analysis—which is only a fraction of the larger study, to be clear—the model happens to place a substantial amount of posterior mass on treatment effects that we would consider impossible.

I demonstrate this behavior of the model in Figure 1, which shows a Bayesian replication of the original model using flat priors. The left two panels show histograms of posterior samples for the the probability of winning at the discontinuity for non-extremists ("control," the first panel) and extremists ("treatment," the second panel). Although the posterior means and modes are within the realistic range of data, a large share of posterior samples for non-extremist candidates exceed 100% probability of winning, and some samples for extremists indicate a probability of winning less than 0% or greater than 100%. The rightmost panel plots the histogram of treatment effects using posterior samples, which is the difference in the intercepts for extremists minus non-extremists. Treatment effects that are composed of at least one "impossible" parameter are indicated with a different color. Of all the treatment effects sampled from the posterior distribution, 36% are a function of at least one impossible parameter value. I correct this behavior with a Bayesian model in Section 5.1, estimating a treatment effect with 20% lower magnitude but also lower variance.

howmuch?

The second application in this paper is an experimental study of public opinion on unilateral actions by presidents (Reeves et al. 2017). The original study uses a series of survey experiments to understand why the citizens' approval of unilateral action varies across certain legal and political contexts. I focus on one experiment that manipulated the specific policy tool or prerogative used by the president to justify unilateral action. The control group was asked to state their agreement with the statement, "Presidents should be able to make new policies without having those policies voted on by Congress." Treatments modified this statement; for example, "Presidents should be able to issue *national security directives* to make new policies without having those policies voted on by Congress" (emphasis mine). The set of policy instruments includes executive agreements, executive orders, national security directives, directing cabinet officials, initiating military operations, issuing proclamations, and issuing memoranda.

The original figure of results (from t-tests) is shown in Figure 2. Treatment effects were all positive, but because most of one-off effects were statistically insignificant the authors initially conclude that information about policy instruments or prerogatives has minimal effects. Their interpretation of the effects veers in an informally Bayesian direction when they note the similarity of the effects overall:

> At the same time, […] the treatment effects displayed in [the figure] are uniformly positive, which suggests that providing more detailed information about the president's behavior increases public support for that action. Given low levels of public knowledge about the different tools of unilateral action, it is possible that providing respondents with even more specific information about each of them would generate larger differences in levels of support.

Interpreting the results this way is partial pooling, done informally. Because the treatments share many similarities, it makes sense to combine some information across treatments

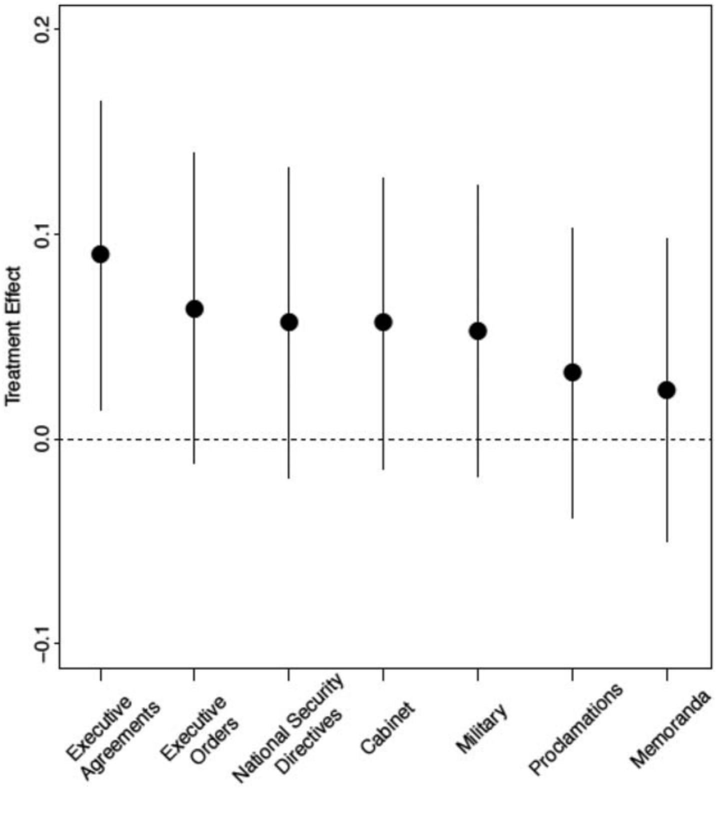

Figure 2: Treatment Effect Estimates from Reeves et al. (2017: Figure 3, p. 458)

to get a general sense of the distribution of similar treatments. This is reasonable; it would be unlikely to observe seven positive estimates if the null hypothesis is true.[10] By building a Bayesian model for these treatment effects, we can do partial pooling formally rather than informally. The results in Section 5.2 show that allowing the model to combine information across treatments leads to more precise estimates for each one-off treatment effect, and it shows that the overall mean of the treatments is positive

3 Tensions and Disagreements

There is some tension between Bayesian practices and causal inference as they are currently performed in political science. While I think much of this is owed to misunderstandings across the divide, open adjudication of the tensions whatever their cause should make any prospect for cooperation easier for both sides. I think this paper will be nothing more than a jumping off point and that much more work can be done to link the applications of Bayes and causal inference in political science. I will start by discussing some hesitations about Bayesian analysis from the viewpoint of causal inference, and then I will offer some words to ease those hesitations. The point is not to invalidate or undermine these hesitations; the hesitations are fair and respectable. The point is to work toward common ground where Bayesian analysis can be done in a way that is respectful of the skeptical view.

Much of the tension comes from the supposed clash between minimalist and maximalist approaches to statistical analysis. Causal inference in minimalist, experiments are simple, and nonparametric estimators rely on fewer assumptions. With all that work to identify a causal effect with as few assumptions as possible, why throw all that parsimony away by adding a bunch of priors?

The response is that I think the view that Bayesian modeling is maximalist is a result of its niche in the field, not because of anything inherent to what Bayesian analysis is. It happens to be the case that Bayesian modeling in political science leans toward maximalism because the flexibility of the framework has made it a tool-of-choice for particular types of problems, but ultimately, Bayesian analysis is "merely" an approach to parameter estimation. It has the flexibility to be maximalist, but it could be put to simple uses as well. And if Bayesians want to participate in causal inference in political science, they should probably get comfortable scaling back the complexity of their models. One way to help normalize Bayesian analysis for

[10] For the curious, if we use a binomial distribution to calculate the probability of observing zero negative estimates out of seven under the null, the probability is approximately 0.008.

"ordinary" data analysis problems would be for Bayesians to do participate in more "ordinary" data analysis.

Causal inference prefers nonparametric models as a default (Keele 2015). Does this rule out Bayesian approaches, which tend to use parametric models? Not necessarily. While there are nonparametric approaches to Bayesian analysis (Ghosal and Van der Vaart 2017; Müller et al. 2015), they tend to be complex and not in the same spirit as nonparametric causal inference. It makes much more sense to model the nonparametric estimate itself as a draw from a the true parameter, which has a prior. I use this approach below to analyze the presidential powers experiment (Section 5.2). This can be especially helpful when nonparametric or semiparametric approaches give results that violate reasonable priors, which I also demonstrate with the House elections RDD (Section 5.1).

Before moving away from the subject of complexity, I think it is appropriate to ask whether experiments will always be so simple. As more political scientists adopt causally rigorous research approaches, experiments and other causal methods will probably become more complex. As a result, demand for statistical tools to deal with these problems would likely increase.

Priors are another area of hesitation. Even if a skeptic of Bayesian methods is convinced that flat priors are unrealistic and upweight extreme estimates, it's fair to question the value of priors in purely a pragrmatic or rhetorical perspective. Suppose that the prior doesn't add anything; the difference of means is statistically and substantively significant: then what's the point of the prior? On the flip side, suppose that the difference of means is noisy: would you trust a prior that showed you something that the data didn't show?

This pragmatic view is understandable, but I think it misunderstands where priors are most effective in applied circumstances. Firstly, priors are not tools for dealing with confounding, so the fact some data come from an experiment is not fundamentally at odds with Bayesian analysis. More importantly, priors are most effective by including weak information. It's helpful to think of priors as *stabilizing* parameter estimates, not by upweighting the researcher's preferred hypothesis but by downweighting parameters that would be unbelievable to the researcher or impossible by the nature of the data. The primary effect of these priors is to provide just enough structure to regularize models against overfitting to noisy data and thus overrejecting null hypotheses.

There is lastly the issue that if Bayes were given space in causal inference, researchers would have to expend a lot of effort to learn how to interpret Bayesian analysis. This is true, but I think it's a fair trade. Causal inference is also very difficult to learn, and it is likely to drastically alter the trajectory of empirical political science. Learning new things is good. It makes us all smarter, more creative, and more open-minded to diverse methodological approaches.

4 Practical Bayes for Skeptical Causal Inference

There are opportunities to improve data analysis using Bayesian approaches in ordinary circumstances as well as in extraordinary circumstances. Bayesian analysis is controversial though, so it should be done with caution. This section discusses a handful of Bayesian tricks that can strengthen a the estimation of causal parameters without stacking too many difficult assumptions atop the design. This section can serve both as advice to modelers and as reassurances to Bayesian skeptics that the tools of the Bayesian trade can be applied conservatively and carefully.

Focus on model-heavy designs and secondary analyses, not the difference in means. I used to have a friend who would get himself into fist fights. I once asked him what to do if I ever found myself in a fight. The first thing he told me was, "Know when *not* to fight." This is a good starting point for thinking about where Bayes could be used to perform causal inference. For instance, always show the difference in means, even if you think it could be improved. Everybody will want to see it without any extra flourishes. Instead, Bayesian analysis could be more effectively used where the researcher relies more on modeling. Some identification strategies are more model-driven than others, such as regression and difference-in-differences. Secondary analysis after the main result also tend to be more model-driven as researchers perform robustness checks and test alternative hypotheses. These areas are where priors can noticeably stabilize an estimate and add clear value to the analysis.

Information, not belief. Priors are sometimes described as a researcher's "beliefs," which is worrisome for empiricism. Most of the time it makes more sense to talk about priors as statements of "information" rather than belief. A prior is an assumption that brings in information that is external to the raw algebra of the model. Although there are subjectivist schools of Bayesian theory, applied work views priors as "just part of the model" and should be "chosen, evaluated, and revised just like all of the other components of the model" (McElreath 2015). Since any model is an inaccurate depiction of the world, the researcher's degree of belief in any prior should be zero; priors should instead be understood as part of the *science* of the model because they can be tested by comparing new data against the model's posterior predictions (Gelman and Shalizi 2013).

Weakly informative priors. A researcher who wants to make causal inference with Bayesian analysis is walking a fine line. Priors are valuable, but most readers will be skeptical of their usage. Researchers should strike a balance with a *weakly informative prior*, which is designed so that "the information it does provide is intentionally weaker than whatever actual prior knowledge is available" (Gelman 2006). The goal is to provide just enough regularizing information to downweight parameters that would be ludicrous or impossible if the researcher were to encounter them, but not so much information to fight against the data.

One helpful approach for developing weakly informative priors is to standardize variables for a model. This makes it much easier to set priors for intercepts, coefficients, and variances by using the standard deviation as a helpful heuristic. We are pretty sure that treatment effects and standard deviations will not exceed 1 in most cases. We are even more assured that neither of these parameters should exceed the full range of the outcome data, so other methods of scaling can be used to improve a prior.

One common example for demonstrating weak information is logistic regression. The log-odds scale extends from $-\infty$ to $+\infty$, but only a small fraction of the log-odds scale maps to probabilities between 1 and 99 percent. For a control group with a predicted probability of 50%, a treatment effect of 3.0 on the log odds scale is an effect of more than 40 percentage points.[11] So for most social science studies, a prior in a logistic regression can safely assume that coefficients of 3 or larger are pretty unlikely. Figure 3 shows how normal priors on the log odds scale translate into the probability scale. The top-right panel shows that a fairly weak $\text{Normal}\left(0, 5\right)$ prior on the log-odds scale drastically upweighting very large and very small probabilities. A truly flat prior would be even more extreme, but the prior is implicitly flat for a traditional MLE logit model. It isn't until the prior standard deviation is near 1.5 that the prior looks approximately flat on the probability scale. Priors like $\text{Normal}\left(0, 1\right)$ and $\text{Normal}\left(0, 0.5\right)$, which sound very narrow at first, are more in line with the expectations we often have in social science.

Weakly informative priors can also be created using principles of maximum entropy. A maximum entropy distribution is the distribution that injects the least amount of prior information given the natural constraints of a problem. For instance, if a process has a mean and a variance, the least informative family of priors for that process is a Normal. These are

[11] $\text{invlogit}(3) = 0.953$

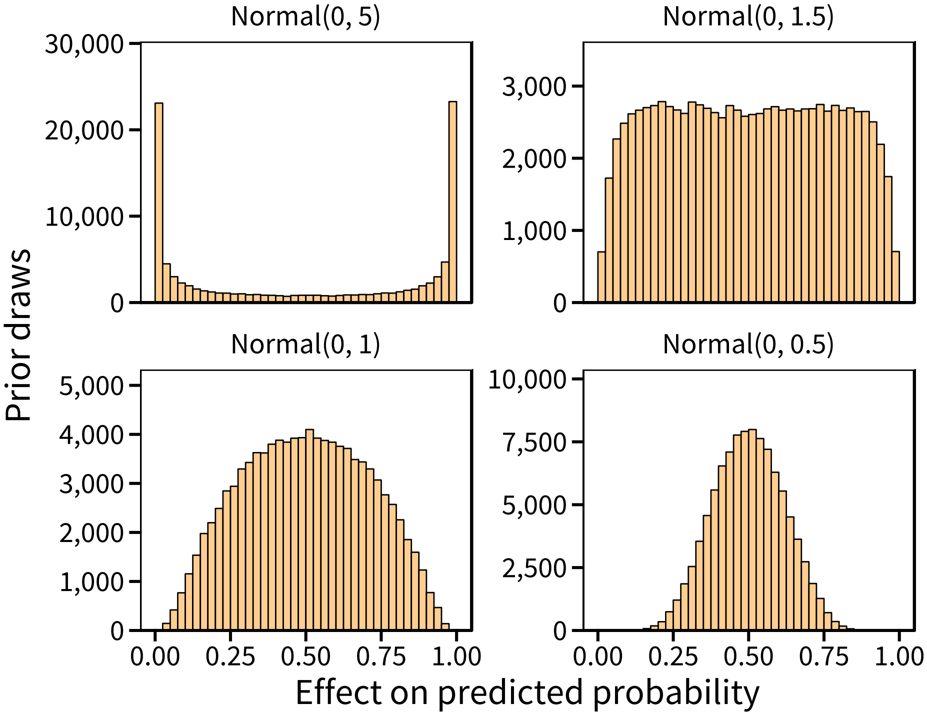

Figure 3: The implications of flat priors. Histograms show prior simulations for logistic regression model coefficients. Prior draws are Normal on the log-odds scale but transformed to the probability scale.

the same properties that justify the use of likelihood families for generalized linear modeling. At the same time, researchers sometimes deviate from maximum entropy priors and opt for some priors because they have desirable properties such as flatter tails than a Normal distribution (Gelman et al. 2008).

cite

Researcher who include weakly informative priors can always simulate data from the prior. This allows the reader to glance the consequences of certain prior choices and facilitates transparency in model-building.

Full uncertainty and regularization with multilevel models. Since many datasets have hierarchical structure, multilevel models are often appropriate to account for multiple sources of variance and correlated error within groups of data. By doing fully Bayesian estimation for multilevel models, all parameter and hyperparameter uncertainty is reflected in the posterior distribution. This allows researchers to build an honest model about multilevel data while being conservative about the amount of actually-existing variation.

Multilevel models are also helpful for partial pooling, which allows a model to learn about the shared characteristics of units and noisy shrink estimates toward the mean of the units. McElreath describes the intuition of partial pooling with a thought experiment about wait times at various cafés.[12] We want to estimate how long it takes to wait in line for a cup of coffee, starting with a vague prior. As we order coffees from various cafés, we gradually learn several things about how long it takes to get a coffee. First, we get a sense how each individual café differs from others. Second, because cafés are pretty similar even if they aren't identical, we learn about the distribution of cafés and use this to pool the information we get from several cafés together. Lastly, because we can pool informationa bout similar cafés, we get a sense of the average and variance of cafés overall, which is what we're usually interested. Even if we haven't collected much data from a particular café, we have reasonably good expectations about it because our priors have been updated by our data from other cafés. This allows us to regularize estimates against noisy and prevent overfitting and type-one errors. We even have reasonable prior estimates for new cafés that we have never visited before because we understand the distribution from which the café is drawn. In short, as we collect data from multiple clusters (cafés), but we are allowed to combine information across clusters in a way that teaches us about variation both within and between clusters. This is unlike the way we ordinarily handle clusters of data in political science, where we estimate all treatment effects

[12] Mcelreath's example is featured in the second edition of his textbook (forthcoming), but is discussed in online lectures as well (see https://github.com/rmcelreath/statrethinking_winter2019).

separately with fixed effects. But there are some situations where pooling information across treatments is natural and intuitive, which I demonstrate in Section 5.2.

5 Estimating Causal Models Bayesianly

5.1 Regression Discontinuity using Weakly Informative Priors

Researchers often encounter modeling scenarios where the problem presents natural constraints on which parameter values are plausible or even possible. I introduced an example with regression discontinuity earlier where flat priors resulted in posterior treatment effect estimates that included "impossible" parameter combinations—predicted win probabilities greater than 100% or less than 0%. This section applies weakly informative priors to overcome this byproduct of the model and return more sensible estimates.

The model we are replicating is the local linear RDD predicting the probability of winning the general election. Hall estimates this binary outcome using a linear probability model, where y_{dpt} equals 1 if the candidate in district d in party p in election t wins the general election. Treatment is assigned when the extremist primary candidate's margin in the general election (m_{dpt}) is greater than 0. I begin by rewriting his model by indexing parameters according to treatment status in order to set priors more easily. Let's index the treatment status with $w \in \{1, 2\}$, which equals 2 if the extremist candidate wins the primary. First the model assumes that y_{dpt} is distributed Normal with a conditional mean that is the prediction from the RD equation,

$$y_{dpt} \sim \text{Normal}\left(\alpha_{w[dpt]} + \delta_{w[dpt]} M_{dpt}, \sigma\right), \tag{9}$$

where α_w are the treatment and control intercepts, and δ_w are the coefficients on the extremist primary margin M_{dpt}. The treatment effect at the discontinuity is the difference between the intercepts for extremist primary winners and losers: $\alpha_2 - \alpha_1$. At this point, the algebra of the model is no different from the original study.[13]

For our prior, we do nothing more than truncate the prior distribution for the intercepts at 0 and 1. Aside from that, we put no additional prior information about which win probabilities are more likely for extremists or not-extremists. We do this with a uniform prior, $\alpha_w \sim$

[13] Hall estimates both conventional and heteroskedasticity-robust standard errors in his paper and reports whichever is largest. For this model, it makes hardly any difference which standard error is used, so for simplicity I make no heteroskedasticity adjustments.

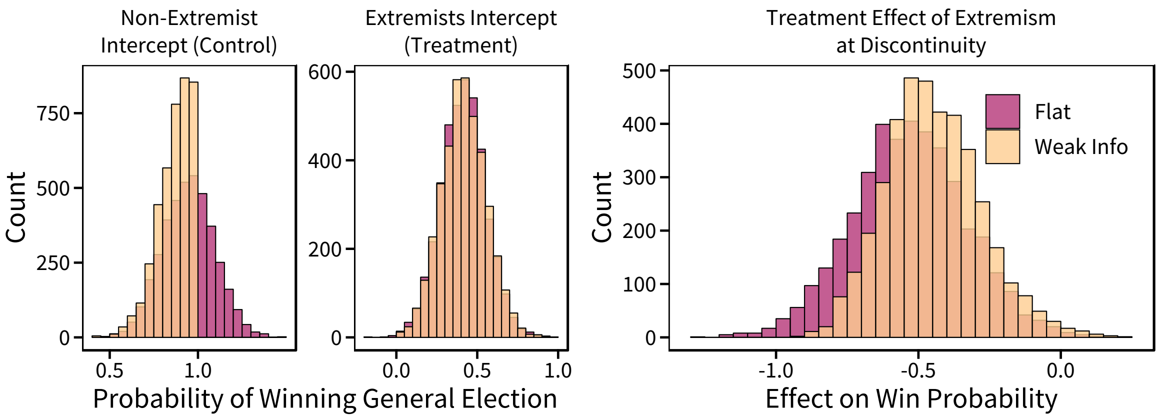

Figure 4: Comparison on posterior draws from Hall's (2015) RDD with flat vs. weakly informative priors.

Uniform $(0,1)$. To isolate the effect of this one change, I will leave the other priors as flat.

Figure 4 compares the posterior distributions for the original flat priors model to this new "weakly informed" model. How did the prior change the posterior inference? The same way that priors normally do: the effect with more prior information is somewhat smaller than the effect with flat priors (by nearly ten points, -0.45 compared to -0.53. The estimate is also more precise. This is because a good deal of posterior uncertainty was owed to model parameters that were impossible and should have been removed from the prior. As a result, the problematic behavior initially identified in the model was removed by merely specifying that the constants should fall between 0 and 1—something that we already knew ahead of time. Causal inference is still possible even with this small prior intervention, and our estimate resembles the original finding even though we've improved it.

5.2 Treatment Effects with Regularization and Partial Pooling

Experiments and other identification strategies often provide researchers with clusters of information that can be combined in a principled way. Section 2.4 describes one such situation encountered by Reeves et al. (2017) in an experimental study that measures how U.S. citizen approval of unilateral action by the president varies with information about the prerogative or policy tool employed. The authors estimate the effects of multiple similar treatments using separate t-tests but interpret the signal from the t-tests jointly (see Figure 2 above). This section demonstrates how these treatment effects can be estimated in a way that preserves information about similar units in order to justify the joint interpretation on statistical grounds. This example also demonstrates how researchers can estimate nonparametric treatment effects as a first cut and model the true underlying effects in a Bayesian framework.

Let $\hat{\beta}_j$ be the estimated treatment effect for treatment $j \in \{1, \ldots 7\}$, which is a difference from the control group mean. We assume (using the central limit theorem, as with all conventional means tests) that these estimates are Normal draws from the true mean β_j and estimated standard error $\hat{\sigma}_j$.[14]

$$\hat{\beta}_j \sim \text{Normal}\left(\beta_j, \hat{\sigma}_j\right) \tag{10}$$

Because these treatments are similar but not identical to one another, it makes sense that the original authors wanted to construct a joint interpretation of them, i.e. what is the "average" or "overall" effect among these variants of the treatment? This interpretation implicitly assumes that each observed treatment is a result of some underlying process. We build a model of that process using relaxed assumptions that combine information across treatment groups. The idea is that citizens have a *general* attitude toward unilateral power, but that the use of unilateral power in specific contexts are slight deviations from that mean. If we assume that deviations from the overall mean result from an additive accumulation of details about each treatment context, we end up with an assumption that the true effects β_j are themselves distributed around a common mean μ with some hierarchical standard deviation ψ.

$$\beta_j \sim \mathcal{N}\left(\mu, \psi\right) \tag{11}$$

I also estimate models that assume that β_j is a draw from a Student's T distribution with 3 degrees of freedom. This is a more relaxed prior that has fatter tails, giving more prior weight to larger deviations from the center of the distribution.

As a result, we have a multilevel model where the true means for each treatment are themselves given a model in order to understand the overall distribution of treatments. This approach allows the model to remember what it learned from treatment 1 when it estimates the effect of treatment 2. It allows the model to synthesize information across all treatments into its own meta-analysis that results in a posterior estimate of the grand mean of all treatments μ.

Naturally we have to specify priors for μ and ψ. I use very weakly informative priors to

[14] In this example we take the value of $\hat{\sigma}_j$ to be known, but we could easily give it a prior and make the model even more uncertain. Since the standard deviations in these data are so, it makes little difference for this demonstration.

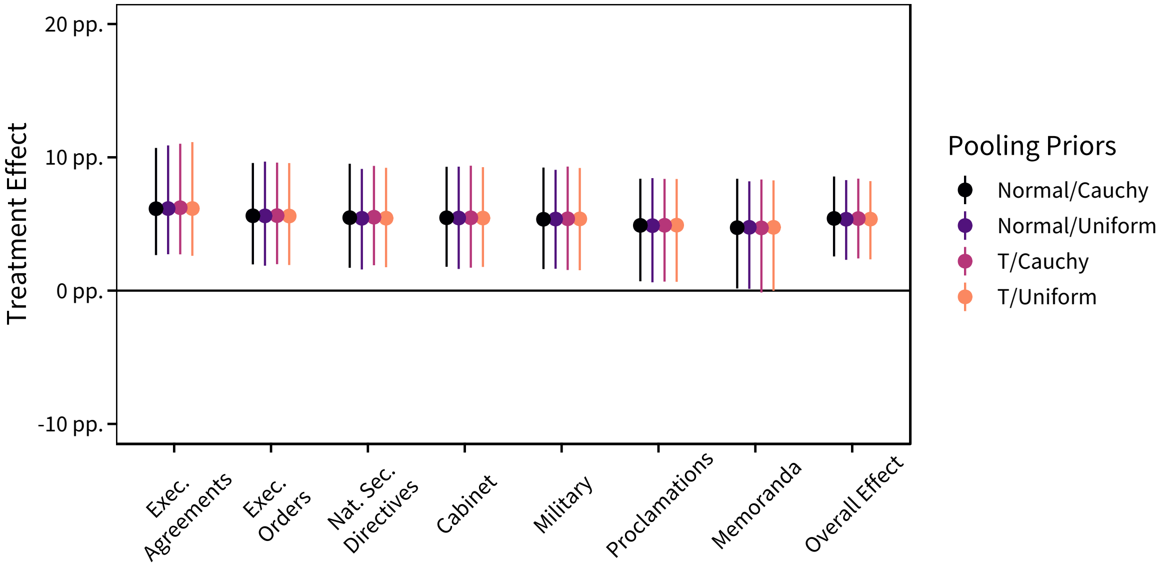

Figure 5: Compare to Figure 2

demonstrate the robustness of the hierarchical model specification. Because the outcome scale is a probability on a 0–1 scale, the maximum possible range of treatment effects is $(-1, 1)$. As a result, I give μ a Uniform $(-1, 1)$ prior, which does nothing but rule out nonsensical parameters. The prior for the scale parameter ψ is also extremely weak, I use either a Uniform $(0, 2)$ or a Half-Cauchy $(0, 1)$ prior. Both priors are wider than the possible variance parameters that could exist; the half-Cauchy in particular encompassing the entire range of possible treatment effects within ±1 scale distance from the center of the distribution.

Figure 5 shows the treatment effect estimates from these hierarchical models and the estimated "Overall Effect" (μ) as well. I show the results of four models for each treatment: two different hierarchical priors for the treatments (Normal and T) and two different priors for the hierarchical scale parameters (Uniform and Half-Cauchy). Right away, it's clear that the estimates are robust to the choice of priors; what matters is that we pooled the treatments using a model at all. The multilevel model has the effect of shrinking estimates toward the mean of the distribution. This is because the prior for each treatment effect is adapting to the information learned about all treatment effects. If the model assumes that the treatments are draws from a common distribution, even if our priors about that distribution are very wide, then the model can rationalize more parameter estimates closer to the mean than farther from it. This makes sense; if all we know about the true parameter was that it was drawn from a distribution, we would place a stronger bet on the possibility that the true parameter is closer to the mean than it is farther away from it. In this example, the shrinkage estimates supported the authors' original interpretation of the study; if the model is allowed to combine information from similar treatments instead of being forced to forget information, the fact that all initial treatment effects were positive *did* strengthen the case in favor of a positive overall treatment effect.

The amount of pooling is not an arbitrary decision introduced by the researcher. Rather, the appropriate amount of pooling is learned from the data because the distribution of groups is updated during model fitting. For this reason, the hierarchical prior is commonly called an "adaptive" prior because its parameters (μ and ψ) are estimated from the data rather than specified exactly by the researcher. The researcher can ask the model for more pooling by placing a tighter prior on the hierarchical scale, but this is often unnecessary because pooling can be achieved even with vague and weakly informative priors (Gelman 2006).[15] At the same time, it is good to remember that the hierarchical prior is just that: a prior. It does not require that the draw of groups in the data has perfectly symmetrical shape. It is more appropriate to imagine that the posterior distribution contains an infinite number of group configurations—some more symmetrical than others—that are weighted by their plausibility given the adaptive prior and the data.

Partial pooling is appropriate for this example because it reflects the way we ordinarily learn about the world, appreciating the variation both within groups and across groups. It allows us to accumulate knowledge over clusters of data rather than throwing away information after each cluster. Second, partial pooling is fundamentally *conservative* because its primary consequence is to regularize estimates toward a common mean, protecting against the risks of small samples and noisy data. This regularization goes by other names in non-Bayesian contexts: ridge regression, penalized likelihood, and cross-validation approaches to model selection. It is "regression to the mean" as applied to model parameters.

6 Unfinished Business

The purpose of this paper is not to convert all causal empiricists to Bayesian analysis. Instead it modestly seeks to justify a space for Bayesian analysis in contemporary causal inference work in political science. Research designs with causal identification still present numerous

[15] Researchers sometimes object to multilevel or "random effects" models on the grounds that it assumes that groups are *exchangeable*—there is no information that we have that would lead us to assign a different prior for any group or subset of groups. But this is the same assumption that all regression modeling makes at the unit level, and any regression of groups such as U.S. states or countries makes the same assumption. The assumption is easier to justify if the hierarchical mean is itself a regression on covariates, in which case exchangeability applies only to the regression errors.

opportunities for Bayesian analysis, but applied researchers have little precedent in the literature for seizing those opportunities. My goal for this paper is to help build that precedent by confronting the issue directly. Looking forward, many unresolved issues still remain.

Bayesian analysis is not automatic. Building a model with appropriate assumptions takes effort and can be slow. Although there are many small decisions a researcher could take through the model (a problem that is not unique to Bayesian analysis), the costs of model-building will deter the researcher from building 100 variations of a model and showcasing the nicest one.

Pre-analysis plans (PAPs) are important for Bayesian and non-Bayesian analysis alike to manage the issue of "researcher degrees of freedom." For Bayesian analysis, PAPs should contain prior predictive simulations to check the reasonableness of model assumptions. Future work will showcase how prior predictive simulations can also be incorporated into a `DeclareDesign` workflow to reassure audiences that Bayesian approaches are not being used to p-hack (which is hard to do in Bayesian analysis anyway because most priors regularize against large, noisy effects).

In conclusion, the causal inference movement in political science asks us to think harder and think better about how we estimate parameters. Bayesian causal inference helps us think harder about estimation in addition to research design.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2017. "Sampling-based vs. design-based uncertainty in regression analysis." *arXiv preprint arXiv:1706.01778* .

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Brandt, Patrick T, and John R Freeman. 2006. "Advances in Bayesian time series modeling and the study of politics: Theory testing, forecasting, and policy analysis." *Political Analysis* 14(1): 1–36.

Branson, Zach, Maxime Rischard, Luke Bornn, and Luke W Miratrix. 2019. "A Nonparametric Bayesian Methodology for Regression Discontinuity Designs." *Journal of Statistical Planning and Inference* .

Chib, Siddhartha, and Liana Jacobi. 2016. "Bayesian Fuzzy Regression Discontinuity Analysis and Returns To Compulsory Schooling." *Journal of Applied Econometrics* 31(6): 1026–1047.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(02): 355–370.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian workflow." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182(2): 389–402.

Gelman, Andrew. 2006. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian analysis* 1(3): 515–534.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su et al. 2008. "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics* 2(4): 1360–1383.

Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. "Philosophy and the practice of Bayesian statistics." *British Journal of Mathematical and Statistical Psychology* 66(1): 8–38.

Gelman, Andrew, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis*. Chapman and Hall/CRC.

Gerber, Alan S, Donald P Green, and Edward H Kaplan. 2004. "The Illusion of Learning from Observational Research." In *Problems and Methods in the Study of Politics*, eds. Ian Shapiro, Rogers Smith, and Tarek Masoud. Cambridge University Press , 251–273.

Ghosal, Subhashis, and Aad Van der Vaart. 2017. *Fundamentals of nonparametric Bayesian inference*. Vol. 44 Cambridge 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.

Grimmer, Justin. 2010. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18(1): 1–35.

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

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

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

Hollenbach, Florian M, Jacob M Montgomery, and Adriana Crespo-Tenorio. 2018. "Bayesian Versus Maximum Likelihood Estimation of Treatment Effects in Bivariate Probit Instrumental Variable Models." *Political Science Research and Methods* , 1–9.

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. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and statistics* 86(1): 4–29.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

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

Jensen, Amalie, William Marble, Kenneth Scheve, and Matthew J Slaughter. 2019. "City Limits to Partisan Polarization in the American Public." Paper presented at the Annual Meeting of the American Political Science Association,.

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

Keele, Luke, Corrine McConnaughy, and Ismail White. 2012. "Strengthening the experimenter's toolbox: Statistical estimation of internal validity." *American Journal of Political Science* 56(2): 484–499.

Kernell, Georgia. 2009. "Giving Order to Districts: Estimating Voter Distributions with National Election Returns." *Political Analysis* 17(3): 215–235.

Lee, David S, and Thomas Lemieux. 2010. "Regression discontinuity designs in economics." *Journal of economic literature* 48(2): 281–355.

Matzkin, Rosa L. 2007. "Nonparametric identification." *Handbook of econometrics* 6: 5307–5368.

McElreath, Richard. 2015. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Montgomery, Jacob M, and Brendan Nyhan. 2010. "Bayesian model averaging: Theoretical developments and practical applications." *Political Analysis* 18(2): 245–270.

Müller, Peter, Fernando Andrés Quintana, Alejandro Jara, and Tim Hanson. 2015. *Bayesian nonparametric data analysis*. Vol. 18 Springer.

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.

Reeves, Andrew, Jon C Rogowski, Min Hee Seo, and Andrew R Stone. 2017. "The contextual determinants of support for unilateral action." *Presidential Studies Quarterly* 47(3): 448–470.

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

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

Rubin, Donald B. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics* 6(4): 377–401.

Rubin, Donald B. 1991. "Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism." *Biometrics* , 1213–1234.

Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *The Annals of Applied Statistics* 2(3): 808–840.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3): 941–955.

Shor, Boris, Joseph Bafumi, Luke Keele, and David Park. 2007. "A Bayesian multilevel modeling approach to time-series cross-sectional data." *Political Analysis* 15(2): 165–181.

Western, Bruce. 1998. "Causal heterogeneity in comparative research: A Bayesian hierarchical modelling approach." *American Journal of Political Science* 42(4): 1233.