Generalizing Survey Experiments Using Topic Sampling: An Application to Party Cues

Scott Clifford[1]
Associate Professor
University of Houston

Thomas J. Leeper
Senior Visiting Fellow
London School of Economics

Carlisle Rainey
Associate Professor
Florida State University

Abstract

Scholars have made considerable strides in evaluating and improving the external validity of experimental research. However, little attention has been paid to a crucial aspect of external validity – the topic of study. Researchers frequently develop a general theory and hypotheses (e.g., about policy attitudes), then conduct a study on a specific topic (e.g., environmental attitudes). Yet, the results may vary depending on the topic chosen. In this paper, we develop the idea of topic sampling – rather than studying a single topic, we randomly sample many topics from a defined population. As an application, we combine topic sampling with a classic survey experiment design on partisan cues. Using a hierarchical model, we efficiently estimate the effect of partisan cues for each policy, showing that the size of the effect varies considerably, and predictably, across policies. We conclude with advice on implementing our approach and using it to improve theory testing.

Word Count: 9,218

Keywords: generalizability, external validity, survey experiments, partisan cues

[1] Corresponding author.

Introduction

Experiments are on the rise in political science, but concerns remain about external validity, whether in terms of sample, context, or treatments. Political scientists have studied how sample characteristics affect generalizability (Berinsky, Huber, and Lenz 2012; Clifford, Jewell, and Waggoner 2015; Coppock, Leeper, and Mullinix 2018; Mullinix et al. 2016), and also the role of experimental context (Barabas and Jerit 2010; Coppock and Green 2015; Jerit, Barabas, and Clifford 2013). Yet, less attention has been paid to another important aspect of external validity that is completely under the researcher's control – the topic of study. Suppose a researcher is interested in whether information changes political attitudes. The typical approach is to select a single, specific topic, such as foreign aid, and design topic-specific stimuli (e.g., Gilens 2001). However, scholars rarely develop narrow theories that apply only to a single topic, instead making a topic selection out of some combination of convenience, theoretical guidance, practical relevance, and personal interests. To what extent does this topic selection affect experimental results and thus the generalizability of the findings?

Given the chosen topic is just one from a larger population of possible topics, many articles conclude with caveats about experimental results being "inevitably bound to some degree by the substantive issues we have chosen" (Chong and Druckman 2010, 678) or "circumscribed by our focus on a single issue" (Chong and Druckman 2012, 14). Researchers sometimes address this threat to external validity by reporting multiple

experiments or multiple arms of an experiment, each on a different topic. Replication by other researchers is also possible, though rare. However, these approaches are costly both in time and resources, and only incrementally increase our confidence in the generalizability of the results across topics. Moreover, the focal topics are typically selected by the researcher as ideal tests of the theory, such as issues that are particularly unfamiliar or where opinions are likely to be particularly malleable (Kam 2005), perhaps inflating the likelihood of supportive findings.

This is not only a problem for external validity, but also for theory development and testing. Indeed, many researchers have theoretical expectations as to how a treatment effect might vary across topics. For example, Bakker, Lelkes, and Malka (2020) suggest that party cues will be less influential on salient and politicized issues. Jerit (2009, 422) speculates that the effectiveness of predictive appeals might be "different for 'old' as opposed to 'new' issues." Chong and Druckman (2010, 678) argue that "[i]ssues that evoke passionately held values should be less susceptible to framing effects." To test these hypotheses, researchers typically randomize respondents into one of two topics selected to represent different levels of an expected moderator (e.g., easy vs. hard issues). Yet, the question remains as to how well each topic represents the intended, broader collection of topics.

In this paper, we propose and implement a novel experimental design and modeling approach to overcome these problems. In short, the proposed design involves randomly selecting a sample of topics from a larger population, designing an arm of the experiment

corresponding with each of the sampled topics, and randomizing respondents into (at least) one arm of the experiment. We refer to this approach as "topic sampling." Combined with a hierarchical model of treatment effects, topic sampling allows the researcher to (1) estimate individual treatment effects for each particular topic, (2) summarize the average or "typical" effect in the population, (3) summarize the variability in the treatment effects across topics, and (4) test hypotheses about how treatment effects vary with topic characteristics (e.g., easy vs. hard issues).

In the sections below, we first explain how the selection of a topic (in this case, policies) can influence treatment effects and yield disparate findings. Then, after briefly discussing the shortcomings of existing solutions, we introduce the idea of topic sampling and discuss the implications for research on partisan cues. After identifying a population of relevant topics, we conduct a topic sampling experiment on partisan cues. Using a hierarchical model, we demonstrate substantial heterogeneity in treatment effect size that is predictable by the level of pretreatment on the topic and the type of issue being considered (e.g., social vs. economic). We conclude with advice on how to apply topic sampling to a wide variety of research.

How Topics Vary and Why It Matters

When designing an experiment, researchers typically choose a particular topic to study, though the nature of that topic varies. For example, scholars interested in foreign policy attitudes might present respondents with a hypothetical scenario involving military

intervention in a specific country (e.g., East Timor; Grieco et al. 2011). Or researchers might investigate ideological asymmetries in political tolerance by randomizing between two social groups (e.g., Arabs vs. Americans; Lindner and Nosek 2009). In each case, the researcher picks a single case (e.g., issue or country) from the population of possible topics or attempts to side-step the topic selection by using hypotheticals (e.g., Kertzer and Brutger 2016).[2] The target population will itself vary depending on the research question, but it could be every political issue discussed by candidates during an election year, every potentially hostile foreign country, or every salient social group. Researchers picking only one or two topics to study must assume that their selected topics generalize to the larger population or admit that their findings may have limited generalizability across topics (even if they generalize well in other ways).

To illustrate why the choice of topic matters, we focus specifically on the common case of political issues (e.g., foreign aid) in public opinion research, both for our theoretical discussion and empirical example. Among public opinion researchers, it is widely believed that not all individuals respond to a treatment in the same way. For this reason, researchers have put a premium on collecting nationally representative samples, which allow the direct estimation of the population average treatment effects and conditional

[2] However, respondents often make assumptions about the hypothetical country, which can introduce experimental confounds (Dafoe, Zhang, and Caughey 2018).

average treatment effects (Mutz 2011). Convenience samples, in contrast, do not by design provide unbiased estimates of these quantities unless treatment effect heterogeneity is absent. Examining treatment effect heterogeneity can yield new insights into the nature and breadth of effects. For example, treatment effects often vary across partisan identity, political knowledge, racial attitudes, or gender, and these variables are commonly used as individual-level moderators in experimental studies (for a review, see Kam and Trussler 2016). Nationally representative samples can provide unbiased estimates of conditional average treatment effects and are also more likely to contain greater diversity on these moderating variables. For these reasons, representative samples are often seen as the "gold standard" for survey experiments.

A similar argument can be made for variation in treatment effects across topics of study, but this point has received less attention. Just as two people vary in how they react to a particular treatment, the same person might react differently to a treatment on two different topics (e.g., abortion versus infrastructure). So, just as we should hesitate to generalize from homogeneous samples of respondents, we should hesitate to generalize from studies of only one or two topics to the relevant population of topics. But the variation in effects across topics that threatens generalizability can also inform the theory. For example, variation in effects across topics might help resolve debates over the scope of elite leadership of public opinion (e.g., Lenz 2009; Tesler 2015) or whether information affects policy opinions (e.g., Gilens 2001; Kuklinski et al. 2000).

How Do Issues Vary?

Issues vary in many ways, but we focus here on two broad dimensions that have been frequently invoked by researchers: attitude strength and salience. While attitude strength is typically considered at the individual level, there is evidence for sizeable differences between issues in the strength of the opinions that people hold toward them. Most notably, scholars have long distinguished between "easy" and "hard" issues (Carmines and Stimson 1980). Easy issues tend to be symbolic and based on normative ends, while hard issues are more technical and focus on the means. The result is that people have strong, intuitive responses to easy issues, but are reliant on elite communications to develop attitudes on hard issues. Similarly, scholars have argued that some issues are more "crystallized" in the minds of voters because they are more closely tied to predispositions (Sears and Valentino 1997). Building on this insight, Tesler (2015) uses five cases to show that people are resistant to elite influence on issues that are highly crystallized, contrasting with the findings of Lenz (2009), who focused on four different cases. Similarly, Goren and Chapp (2017) challenge conventional wisdom that issue attitudes are shaped by partisan identification and candidate evaluations, arguing that the direction of causation is reversed for "culture war" topics (operationalized as abortion and gay rights). Other scholars have made a number of similar arguments using different terminology, such as "moral" issues (Mooney and Schuldt 2008) or principled versus pragmatic issues (Tavits 2007). Nonetheless, it's clear that issues vary considerably in the extent to which they generate strong attitudes, which can lead to different substantive conclusions.

A clear commonality in this literature is the contrast between social and economic issues (e.g., Arceneaux 2007; Johnston and Wronski 2015; Simas, Milita, and Ryan n.d.; Tavits 2007). Scholars typically use a social issue to represent the easy, principled, or salient issue, and an economic issue for the opposing category. This argument is made explicitly by Johnston and Wronski (2015, 46), who argue that "on average, social issues are easy issues, and economic issues are hard issues," a measurement choice which they describe as "face valid and intuitive" (for related approaches, see Feldman and Johnston 2014; Johnston, Lavine, and Federico 2017). Similarly, Tavits (2007) argues that social and economic issues fundamentally differ in that the former are inherently principled issues, while the latter are inherently pragmatic issues. Clearly, political scientists see social and economic issues as representing distinct classes of issues that instigate different patterns of public opinion.

The second broad dimension across which issues differ is salience, or the level of attention devoted to the topic by the media. Salience matters for at least two reasons. First, salient issues should be more cognitively accessible in the minds of voters, and thus more likely to be used to evaluate politicians and other political objects (e.g., Bélanger and Meguid 2008; Edwards III, Mitchell, and Welch 1995). This may mean that people can more easily access stored considerations (or a running tally evaluation) about salient issues making them more resistant to influence than on issues they are less familiar with. Moreover, salience is a defining feature of both easy issues (Carmines and Stimson 1980) and moral issues (Mooney 2001). Second, people should be more knowledgeable about

highly salient issues (Barabas and Jerit 2009; Jerit, Barabas, and Bolsen 2006) and more likely to be aware of the parties' stances on the issues. This greater awareness may create stronger, more stable attitudes (Lenz 2012), which matters for reasons described above. But greater awareness also matters for purely practical reasons. Highly salient issues are more likely to be pretreated in the sense that respondents may have already been exposed to the facts, frames, or other issue-relevant stimuli that researchers are manipulating. As a result, highly salient issues tend to generate smaller treatment effects than less salient issues in studies on framing and partisan cues (Druckman and Leeper 2012; Slothuus 2016).

While developing a full theory of how issues vary is beyond the scope of this paper, there are clear theoretical reasons to expect that treatment effects systematically vary across topics. This is the case when the outcome of interest is opinion, but also for other common outcomes like topic-specific knowledge or engagement. Similar arguments could explain how other types of topics (e.g., countries or social groups) vary and thus produce variation in treatment effects. In the next section, we discuss how researchers have attempted to address the challenge of topic-level variation and why these approaches are inadequate, then introduce a new experimental design to address these issues.

What Are the Solutions?

Researchers, of course, have been aware of this problem and have attempted to deal with it using multi-armed studies, systematic literature reviews, and meta-analyses. The most common approach in political science, the multi-armed study, involves selecting two

(or more) issues that differ from each other on some theoretically relevant dimension, then randomize respondents into an issue as well as treatment or control. For example, in a study on the use of ambiguous political rhetoric by politicians, Simas, Milita, and Ryan (n.d.) randomize between transgender rights and business incentives, which are intended to represent principled and pragmatic issues, respectively. In a study on partisan cues, Arceneaux (2007) randomized respondents between abortion and environmental regulation in the federal system, which represented high and low salience issues, respectively. This approach aims to both test theoretical claims about issue differences and to increase the generalizability of the findings by contrasting two different issues.

Of course, the multi-armed study faces substantial shortcomings. If the goal is generalizability, including a second issue offers only a marginal increase beyond a single issue. If the goal is theory testing, then it raises concerns about how well each issue represents the broader category. For example, it's unclear that the topic of business incentives represents the broader class of pragmatic, or economic issues. As we show below, the common practice of relying on social and economic issues to represent fundamental divisions (such as easy vs. hard issues) can yield highly variable results depending on the particular issues that are selected. Thus, while two issues are certainly better than one, multi-armed studies offer only a very modest improvement in the generalizability of the findings.

Systematic literature reviews and meta-analyses promise to leverage more data but face a number of problems. First, the available set of studies is likely subject to publication

bias. For example, Franco, Malhotra, and Simonovits (2014) analyzed the Time-sharing Experiments in the Social Sciences (TESS) database and found that while approximately 60% of studies with strong results were published, only 20% of studies with null results were published. Moreover, 65% of null results were never written up. Thus, in the absence of a study database like TESS, meta-analytic estimates will inevitably be biased toward strong effects, while excluding weaker effects and the corresponding stimuli. A second, related problem stems from researchers' selection of stimuli. Similar to patterns of bias in publication efforts, scholars likely select topics of study that are the most likely to yield strong effects. Researchers also frequently borrow stimuli from previous work (e.g., the ubiquitous hate speech rally; Nelson et al. 1997), providing no new variation in the topic. Third, and perhaps most crucially, it is often difficult to make comparisons between a set of studies because each study typically varies in multiple ways, such as the subject population, the time period, the measurement of the dependent variable, or the implementation of the treatment. These many differences in design make it near impossible to isolate the effect of the selected topic.

Topic Sampling

To address these challenges, we develop a new tool to enhance generalizability that we refer to as "topic sampling." Combined with a hierarchical model, topic sampling allows the researcher to (1) estimate individual treatment effects for each particular topic , (2) summarize the average or "typical" effect across topics, (3) summarize the variability in the

treatment effects across topics, and (4) test hypotheses about how treatment effects vary with topic characteristics (e.g., social vs. economic issues).

As a first step, researchers must identify the topic population of interest and develop a sampling frame (e.g., a list of all salient political issues). Depending on the size of the sampling frame, the next step is to either select a random sample of topics or use the entire population. Then, within the experiment, respondents are first randomized into a topic, then randomly assigned to treatment or control. Finally, the researcher estimates the hierarchical model of the treatment effects, which enables the researcher to accomplish two goals: (1) precisely estimate the treatment effect within each topic by borrowing information across topics and (2) describe how the treatment effect varies across topics.

The topic sampling design enables the researcher to compute several important quantities of interest. First, it enables an estimate of the overall treatment effect—that is, the average treatment effect that would be observed across the full population of topics—perhaps conditional on topic-level explanatory variables. The design also allows the researcher to precisely estimate each treatment effect for the many particular topics. These effects might interest the researcher individually, but they also help the researcher understand how treatment effects vary across topics. Third, and perhaps most importantly, this design allows the researcher to describe how the treatment effects vary with characteristics of the topic. For example, a researcher might investigate whether treatment effects are larger on economic issues than on social issues, or whether political tolerance is more likely to be extended to ideologically similar social groups.

The obvious challenge to our design is that the sample size for any individual topic will be small, rendering less precise estimates of topic-specific treatment effects on our outcome of interest. We address this challenge with a hierarchical model to borrow information across topics. Suppose we have a general conceptual outcome O, but we can only observe particular outcomes o_j for $j \in 1,2, \dots, J$. For example, we might use the particular fact that *Congress needs a two-thirds super-majority to override a presidential veto* to measure the general concept of *political knowledge*. Barabas, Jerit, Pollock, and Rainey (2014) discuss the diversity of questions and facts that researchers can use to measure political knowledge. If researchers are ultimately interested in the general outcome O, they need a *model* to link the particular outcomes they observe $o_1, \dots, o_J$ back to the general outcome O.

We develop an approach that allows researchers to systematically generalize models for particular treatment effects on a particular outcome into a model for the general outcome. For a particular measurement o_{ij} at the respondent-topic level, we have the model $o_{ij} \sim f(\alpha_j, T_{ij})$. That is, the outcome o_{ij} for individual i and outcome j is like a draw from a distribution f that depends on whether the individual received the treatment T_{ij} and parameter α_j (e.g., the treatment effect). In most studies, researchers focus on one or perhaps on a handful of outcomes. For example, Delli Carpini and Keeter (1996) use knowledge of five particular facts to measure the general concept of political knowledge: party control of the House, the percent to override a veto, ideological location of the

parties, definition of judicial review, and identity of the vice president (see Delli Carpini and Keeter 1993, esp. pp. 1198-9).

Rather than assume that a handful of outcomes represent the general concept, we suggest an estimable model to link the models of particular outcomes to a general model. Importantly, we suggest an assumption that the particular models are *different-but-similar*. That is, the model for each particular outcome *j* has a *different* parameter α_j. Nonetheless, we assume that the α_j are *similar* across models, and we estimate the amount of similarity from the data.

Following Bullock, Imai, and Shapiro (2011), we formalize the *different-but-similar* assumption by conceiving of each model parameter (vector) as a draw from a (multivariate) normal distribution, so that $\alpha_j \sim N(\mu_\alpha, \Sigma_\alpha)$. Then, while the treatment effects for particular realizations are different-but-similar, we have a model for those differences (that might include covariates). We can then characterize their approximate value with μ_α and the give-or-take around that value with the Σ_α (i.e., the degree of similarity).

Further, we can leverage the similarity to borrow information across realizations. Rather than limit our focus to a single outcome (with, say, 1,000 observations) or conducting full-powered studies of a handful of outcomes (with a total of 5,000 respondents), we can run a single (carefully designed) full-powered study of *many* outcomes using only, say, 2,000 observations. The stylized model, then, is

$$o_{ij} \sim f(\alpha_j, T_{ij})$$

$$\alpha_j \sim N(\mu_\alpha, \Sigma_\alpha).$$

Importantly, researchers can alter this stylized model to include common modeling features such as control variables and interactions. Researchers can define the probability density (or mass) function f as appropriate. This flexible modeling approach allows researchers to simultaneously estimate a large number of different-but-similar treatment effects with sufficient statistical power and link those to a general model.

An Application to Partisan Cues

Partisan cues have been studied extensively and scholars often speculate that results might differ considerably across issues (e.g., Bakker, Lelkes, and Malka 2020), yet we have little systematic evidence. Many experiments on the topic have been conducted, but the observed variation in treatment effects could be due to many varying features of study design. Of course, researchers have used a variety of issues that are as disparate as food irradiation (Kam 2005) and abortion (Arceneaux 2007). Some scholars intentionally select issues on which respondents have "only weak prior beliefs" (Levendusky 2010, 119). Some studies provide extensive issue information (Boudreau and MacKenzie 2014; Bullock 2011), while others provide very little (Nicholson 2012). These studies also vary in whether the cue focuses on the party, the party leader, the in-party, the out-party, or some combination of these. And, of course, these studies vary in when they were fielded, on which samples, and how the dependent variable was measured. All of these factors make it difficult to ascertain from existing literature how the effects of party cue vary across

particular issues and how well each particular issue generalizes to the larger conceptual outcome. For example, in reviewing findings on the relative impact of party cues and policy information, Bullock (2011, 509) concluded that "variation in these findings defeats most attempts to generalize."[3]

Here, we directly examine how treatment effects vary across issues, while holding all other features of the design constant. We also test two key hypotheses developed from the literature reviewed above on how issues differ. First, building off of the literature on salience and pretreatment, we expect that treatment effects will be smaller when more of the public is already aware of where the parties stand on the issue. Previous work has found some support for this hypothesis. Specifically, Slothuus (2016) found that party cues have the expected effects when citizens were previously unaware of a party's stance on a topic, but have no effect when a party's stance was already widely known. Nonetheless, this evidence is based on only two issues that were selected to maximize differences in

[3] In a recent study, Barber and Pope (2019) study 10 issues at once, providing perhaps the most generalizable findings. Yet, the study was restricted to 10 topics on which former President Trump publicly took stances on both sides of the issue. Moreover, their analysis focuses only on the average effect, while analyses in the supplementary materials suggest meaningful but unexamined heterogeneity in effect size across issues.

pretreatment. Thus, it remains unclear how much the effect of partisan cues varies with pretreatment.

Second, we test the hypothesis that treatment effects will be larger for economic issues than for social issues. This expectation follows from a variety of literature, briefly reviewed above, that describes economic issues as "hard" and "pragmatic," and social issues as "easy" and "principled" (e.g., Carmines and Stimson 1980; Johnston and Wronski 2015; Tavits 2007). Additionally, we examine treatment effects for foreign policy. Following evidence that the public is relatively inattentive and uninformed on these issues (e.g., Almond 1950), we expect that treatment effects will be larger for foreign policy than for social issues.[4]

Finally, we compare our analysis of social vs. economic issues to the common multi-armed study that contrasts just one issue from each category. Taking advantage of the many issues within our study, we show that the selection of issues in a typical multi-armed study can lead to highly variable results, depending on the issues that are selected. These findings underscore the importance of using topic sampling to test hypotheses about differences between types of issues.

[4] We do not have clear expectations about how treatment effects will differ between foreign policy and economic issues.

Defining the Issue Population

One of the major challenges in implementing a topic sampling design is defining the issue population. For any particular substantive question, there does not usually exist a solitary target population. Instead, researchers must choose among various populations to which they can generalize. This choice must be based on a variety of theoretical and practical concerns. The alternative, however, is to select a single issue, or small number of issues, and make no claim about generalizability to other issues (implying either no external validity or universal external validity but leaving which as an exercise to the reader). In contrast, topic sampling provides direct empirical evidence on a much larger set of issues and allows researchers to generalize to a defined population. Of course, some researchers might disagree about the relevant population – both whether it is the relevant population and whether it is correctly operationalized. But without defining and sampling from a population, researchers are left only to speculate about generalizability. Topic sampling allows this debate to progress through empirical evidence.

In the case of political issues, there is clearly no single static population that will be relevant to all studies. In our selection of a population, we sought to balance multiple goals: the topics must be relevant to public opinion, they should be current, and they should not rely overwhelmingly on highly salient issues. To this end, we rely on the Roper Center iPoll database to identify all of the available policy attitude questions asked by public opinion

surveys during 2016 (for a related approach, see Jerit and Barabas 2012).[5] By virtue of appearing in a prior survey, the issue is assured to be of interest to public opinion researchers. And by restricting our search to recent policy questions (and through further refinement, discussed below), we can assure that the questions are currently relevant to politics. Finally, while highly salient issues appear more frequently in the database, the issues cover a wide variety of topics, as shown below. This is because the database includes questions designed and fielded by a variety of organizations, including media outlets, universities, and interest groups. Thus, the iPoll database satisfies all of the qualities we might look for in defining an issue population.

In our application, an alternative approach might rely on Congressional budget codes or the Comparative Agendas Project codes to construct a population of policy issues. While these strategies are plausible, they would include a large number of topics that are not of particular interest to public opinion research (e.g., "copyrights and patents," "maritime issues"). Researchers could instead rely on open-ended responses to "most important problem" questions from ANES, Gallup, or elsewhere to focus on issues that are not just politically salient but publicly salient. However, this approach would exclude a large number of less salient issues, and most responses to these questions fall into a

[5] We used 2016 rather than a later year because 2017 polls were still being added to the Roper database at the time.

handful of broad categories such as "unemployment" or "immigration." Overall, iPoll offers the most appealing option for defining an issue population.

To generate our issue population from iPoll, we searched all questions fielded in 2016 using a string of terms that would commonly be used to measure policy attitudes.[6] A research assistant then downloaded the results and removed any questions that were not designed to measure policy attitudes. For example, we removed all beliefs (e.g., does the death penalty deter crime?), all candidate approval questions, and all questions about vote choice. This process yielded 397 questions but included 243 policy duplicates. After removing redundant questions such that each specific policy question only appeared once, the final dataset contained 154 unique policy questions.[7] While many questions covered hot-button issues, our population also included a variety of less salient issues, such as allowing employees to use their retirement accounts to fund long-term care, government collection of private information on citizens, the trade embargo with Cuba, and regulating the distribution of pornography.

---

[6] Specifically, the terms were "favor or oppose or for or against or should or approve or support." Diagnostic checks suggested that this set of terms included virtually all policy attitudes measured in this time period.

[7] Questions that asked about the same topic but used different question wording were considered redundant.

Next, a research assistant classified each remaining question at three levels. At the lowest level, we classified each question according to the specific *policy*, such as eliminating fossil fuel subsidies. At the next level up, we classified each question's *issue* area, such as energy. And finally, at the top level, we assigned each to one of three *categories* (economic, social, foreign policy). As discussed below, we use these classifications for the purpose of sampling and describing variation in treatment effect size.

Finally, we coded each question according to whether the policy in question tends to receive more support from Democrats or Republicans in the mass public. This coding determined the direction of the treatment effect in order to avoid deception. For salient issues, we relied on our own expertise to determine the direction of partisan support. For cases in which partisan support was unclear, we consulted polling results and assigned support to whichever party exhibited greater support for the policy. Although some issues exhibited only very small partisan differences, establishing the "correct" partisan lean for each issue is not crucial to our design. Our only requirement is that the partisan cue is believable and not deceptive.[8]

[8] We chose not to randomly assign the direction of the partisan cue because the plausibility of the treatment would vary considerably across issues, severely confounding the results.

Experimental Design

A major challenge posed by our design is the creation of a set of comparable stimuli. In the case of party cues, one challenge is that ingroup and outgroup cues may have different effects (e.g., Nicholson 2012). To deal with this challenge, we adopt a standardized question stem, shown below, that provides relative information about the parties' stances. As a result, all respondents in the treatment condition receive information about both the ingroup and outgroup position, though the direction of the treatment depends on the policy. Moreover, by providing *relative* information, we avoid making absolute statements about a party's position (e.g., a majority opposes) that would not be applicable to all policies.

> As you may know, there has been some debate about <*policy*> lately. [*Democrats are more likely to favor <policy>, while Republicans are more likely to oppose <policy > / Republicans are more likely to favor <policy >, while Democrats are more likely to oppose <policy >*]. We'd like to know your opinion. Do you favor or oppose <*policy* >?

Respondents receiving the control condition only received the last sentence of the question above. For our dependent variable, we asked respondents to report their position on a 7-point scale ranging from "strongly favor" to "strongly oppose." The party cue, however, may move respondents in either direction, depending on the combination of the party supporting the policy and the partisan identification of the respondent. Thus, we

reverse the outcome variable for some respondents to create a measure of *Partisan Agreement*, such that higher values always indicate greater agreement with the respondent's party's position.[9] Our design choice entails two assumptions that we find plausible. First, relative partisan cues are equally effective at motivating a partisan to support a policy as they are at motivating opposition to a policy. Second, Democratic and Republican respondents are equally affected by relative party cues (for evidence of symmetry in partisan bias, see Ditto et al. 2019).

Our control group also poses a unique design challenge. The most straightforward application would involve randomly assigning a respondent to one policy, then randomly assigning that respondent to either treatment or control within that policy. However, we opted for a different approach to increase statistical power. Instead, each respondent was randomly assigned to answer six policy questions in random order. To avoid any potential spillover, the first five policy questions asked were all control conditions, while the sixth was always the treatment condition. Thus, we estimate the treatment effect by comparing the levels of partisan agreement on an issue when it is the sixth, treated question to when it is one of the five untreated questions. The benefit of this design is that the five control questions provide additional respondent-level information on their pretreatment levels of

[9] For example, strongly opposing stricter gun control would be coded as high partisan agreement for Republicans and low partisan agreement for Democrats.

partisan agreement.[10] As we discuss in more detail below, we incorporate this information into our model to yield more precise estimates of the treatment effects. This design assumes, however, that there are no order effects on partisan agreement. In other words, answering five control questions does not affect the level of partisan agreement on the sixth question. We consider this a reasonable assumption, though it is not one we can test with the available data.[11]

Manipulation Check and Issue-Level Moderator

Following the measurement of the six policy attitudes, respondents were asked whether they thought Democrats or Republicans are more likely to support each of the six policies. These awareness questions serve as a manipulation check. To assess the level of pretreatment on each issue, we randomly assigned a subset of our sample to an awareness-only module. These respondents did not participate in the focal experiment, but instead answered a series of awareness questions. As discussed in more detail below, we use these questions in the awareness module to produce policy-level estimates of pretreatment that could not be influenced by the experiment. For clarity, from here on out we refer to these estimates as awareness.

---

[10] The five control issues were also recoded so that higher values indicate greater agreement with their party.

[11] Unfortunately, we did not record the order in which the issues were presented.

Topic Randomization

Based on our theory, we expected that treatment effects would vary across policy category (social, economic, and foreign policy). However, policies are not evenly distributed across categories in the population. For example, our population includes 74 social policy questions nested under 14 issues, but only 26 foreign policy questions nested under six issues. We deal with this complication by taking a stratified random sample. Based on a series of simulations used to estimate statistical power, we sought a sample of roughly 50 policies. We sampled proportionately from each of our three categories to create a sample of 24 social policy questions, 16 economic policy questions, and 8 foreign policy questions.

Within each category, we sampled disproportionately because many issues were overrepresented within each category. For example, 24 of the 74 social policies fell under the issue of gun control and 18 of the 26 foreign policies fell under the issue of national defense. We sought to select an even number of policies from within each issue for any given category, but this was impossible as some issues contained only one policy. Instead, we sampled a policy from each issue without replacement and continued this process until we reached the desired number of policies. For example, we needed to sample 8 foreign policy questions. Foreign policy contained 6 issue areas, so we sampled one policy from each of the 6 issue areas. Four of the foreign policy issue areas only contained a single policy, so we randomly selected one more policy question from each of the two remaining issue areas. The resulting sample is shown in Table 1, which consists of 48 policies (e.g.,

banning suspected terrorists from buying guns) nested under 26 issues (e.g., gun control), nested under three categories (e.g., social). Within our sample, 30 of the 48 issues (63%) were coded as being supported by Democrats, which closely corresponds with the distribution in the population of issues (61%).

Table 1. Random Sample of Policies

Category	Issue	Policy	Partisanship
Economic	drug costs	allowing imported drugs from Canada	R
Economic	drug costs	eliminate drug advertisements	D
Economic	drug costs	requiring to pay higher share in drugs	R
Economic	government spending	spending cuts	R
Economic	healthcare	FDA review speed	R
Economic	healthcare	more HIV and AIDS treatment	D
Economic	healthcare	single payer system	D
Economic	inequality	raising national minimum wage	D
Economic	paid family leave	gov't covered the paid leave	D
Economic	paid family leave	gov't provided tax credits to businesses that allow paid leave	D
economic	paid family leave	set up flexible spending account	D
economic	regulation on businesses	regulation on big business	R
economic	retirement plan	setting up retirement plan for workers	D
economic	taxes	higher taxes on the rich	D
economic	taxes	raising property taxes	D
economic	taxes	social program funding	D
foreign policy	international court trials	international criminal court	D
foreign policy	international trade	importing from developing countries	D
foreign policy	international trade	renegotiating trade deals	R
foreign policy	international trade	Trans-Pacific Partnership	D
foreign policy	national defense	allowing Assad to remain in power	D
foreign policy	national defense	arming Syrian militant groups	R
foreign policy	national defense	Iran Nuclear Agreement	D
foreign policy	peace in the middle east	independent Palestinian state	D
social	abortion	end funding for planned parenthood	R
social	abortion	federal funding for abortions under Medicaid	D

social	abortion	ultrasound	R
social	cancer research	gene therapy	D
social	capital punishment	death penalty	R
social	censorship	pornography laws	R
social	drug addiction	requiring drug treatment instead of jail time for illegal drug usage	D
social	equal rights	better sensitivity training for police regarding trans-gender	D
social	equal rights	create programs that allow minorities to get ahead	D
social	equal rights	gender neutral bathrooms	R
social	freedom of the press	allowing news outlets to report on issues pertaining to national security	R
social	gay rights	equal housing opportunity	D
social	gay rights	gay marriage	D
social	gay rights	legally adopt children	D
social	gun control	banning suspected terrorists from buying guns	D
social	gun control	gun show purchase provisions	D
social	gun control	more security in public places	R
social	immigration	ban immigration from terrorist watch listed countries	R
social	immigration	immigrant employment background checks	R
social	immigration	Muslim ban	R
social	marijuana	marijuana legalization	D
social	police violence	body cameras	D
social	sex education	sex education ban in schools	D
social	traditional values	concentration of state or federal gov't	R

Respondent Sample

We recruited 3,500 respondents through Survey Sampling International in the summer of 2018.[12] Respondents were randomly assigned to the primary experimental module (N = 3,250) or the awareness-only module (N = 252). SSI, now Dynata, provides diverse national samples targeted at demographic representativeness. Although it is not a probability sample, the sample is diverse and similar to census demographics on several measures. Due to our focus on partisan cues, we excluded respondents from the experimental module who identified as pure independents (N = 486), leaving a sample of 2,764 respondents.

Modeling Approach

Although we have an ordinal outcome (i.e., "Strongly Oppose" to "Strongly Favor"), we use a normal-linear model, which is easier to understand and estimate and supplies more intuitive quantities of interest. Consistent with the general approach described above, we (1) use a random intercept for each policy question, (2) use a random effect for the treatment effect for each policy question, and (3) allow a correlation between the two. For the numerical outcome $y \in \{1, \ldots, 7\}$, we assume that

[12] This study was reviewed and approved by <university information removed for blind review>. All respondents gave informed consent before participating in the study.

$$y_{ij} \sim N(\mu_{ij}, \sigma_y),$$

and model the location parameter μ_{ij} as a function of (1) whether respondent *i* received the treatment for policy j and (2) the estimated aggregate level of awareness about the parties' positions on policy *j* (i.e., the amount of pretreatment), so that

$$\mu_{ij} = \beta_j^{cons} + \beta_j^T T_{ij} + \beta_j^A A_j + \beta_j^{T \times A}(T_{ij} \times A_j).$$

T_{ij} indicates whether respondent *i* received the treatment for policy *j*, A_j represents the public awareness of the parties' positions on policy *j*. The parameters $\beta_j^{[i]}$ represent *potentially* random effects. According to our theoretical approach, the treatment effects should vary across policies, so that *at a minimum* the intercept β_j^{cons} and treatment effect $\beta_j^T T_{ij}$ should vary across policies. However, it is worth comparing this model to alternatives that vary in their complexity, but are consistent with the different-but-similar approach.

There are four levels of complexity of the structure of the random effects, which are summarized in Table 2. And because policies are nested in issues, which are nested in categories, each of these random effects structures can be applied at each level. At the most basic level, our model might hold all parameters fixed (0) and assume no variation in the intercept or treatment effect (e.g., across policies). Next, we can allow only the intercept β_j^{cons} to vary (1). Further, we can allow both the intercept and treatment effect to vary (2). And, finally, we can allow the interaction between awareness and the treatment to vary (3). To be clear, each of these four structures can be applied at each of the three levels (policy,

issue, category), except we do not consider (3) at the policy level giving a total of $3 \times 4 \times 4 = 48$ possible structures.[13]

Table 2. Summary of Random Effects Structures

Code	Description
0	All parameters are fixed.
1	The intercept β_j^{cons} varies. The treatment effect and the interaction with awareness are fixed.
2	The intercept β_j^{cons} and the treatment effect $\beta_j^T T_{ij}$ vary. The interaction is fixed.
3	The intercept β_j^{cons}, and the treatment effect $\beta_j^T T_{ij}$, and the interaction $\beta_j^A A_j$ and $\beta_j^{T \times A}$ vary. (We do not consider this structure at the policy level.)

We use the format "[policy code] : [issue code] : [category code]" to compactly describe a particular random effect structure. For example, "2 : 0 : 2" denotes a model where the intercept and treatment effect vary at the policy and category levels. The simplest structure that allows the effect to vary across policies is 2 : 0 : 0, in which the intercept and treatment effect vary at the policy level. The most complex structure is 2 : 3 : 3, which additionally allows the intercept, treatment effect, and interaction with awareness to vary at the issue and category levels. In total, we consider 48 possible structures.

[13] Because each policy has a single level of awareness, we do not consider models where the interaction between the treatment effect and awareness vary across policies.

To adjudicate among the structures, we use the predictive accuracy of the model. In particular, we use Vehtari, Gelman, and Gabry's (2017) method to efficiently approximate the leave-one-out cross-validation (LOO-CV) for each model. Lower LOO-CV scores (which are on the deviance scale) indicate a model with higher predictive accuracy. Figure 1 shows the approximate LOO-CV for each specification. It suggests that the 2 : 0 : 2 structure best describes the data.

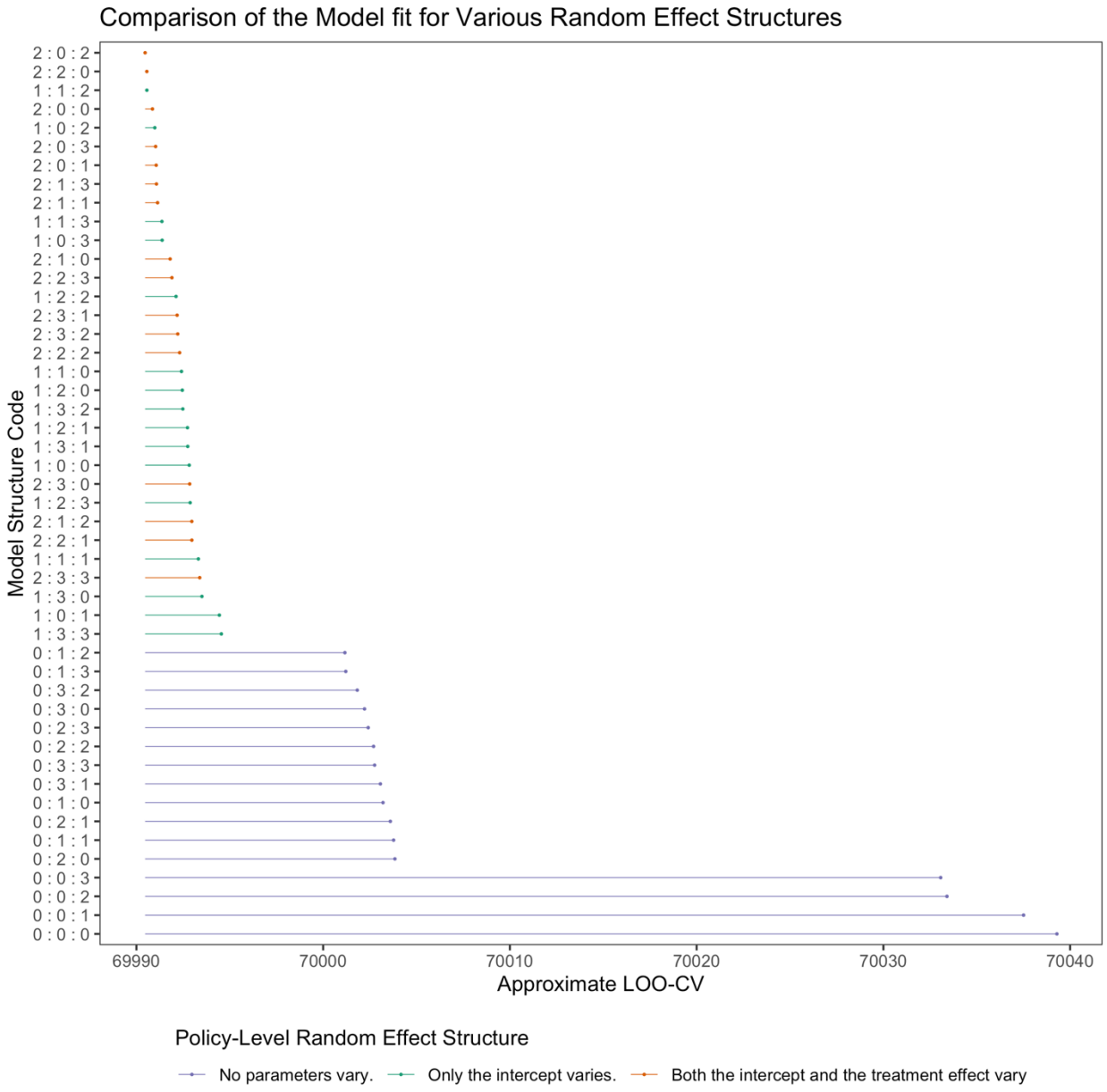

Figure 1: This shows the approximate LOO-CV fit criterion for various random effect structures. We use the best fit (2 : 0 : 2) in the analysis below.

This figure shows considerable policy-level variation *above that explained by awareness* and *above that explained by variation at the issue and category level*. Notice that

structures that allow the treatment effect to vary at the policy level (orange) consistently explain the data better than those that allow no parameters to vary at the policy level (purple) or allow only the intercept to vary (green). Keep in mind, when evaluating these specifications, that we model the treatment effect as a function of awareness, which explains about 60% of the variation in the treatment effects across policies. Thus, the variation in treatment effects at the policy level is *above that explained by awareness.*

Quantities of Interest

To evaluate the hypotheses, we use several quantities of interest derived from the statistical model. Because we have a fully Bayesian approach, we have simulations for each of these quantities of interest. To summarize the posterior distributions, we use the average and the 90% percentile interval (i.e., the 5th and 95th percentiles of the posterior simulations). To evaluate the evidence for each of the hypotheses, we use posterior probabilities (i.e., the percent of the posterior simulations that are consistent with the hypothesis). To assess the evidence for the hypotheses, we use the following guidelines: we consider 95% or more as "strong evidence" for the hypothesis, 90% to 95% as "moderate evidence," and less than 90% as "no or weak evidence." Of course, we have a continuous measure of evidence, so readers should not rely exclusively on the strict trichotomy. We use the tidybayes package in R (Kay 2019) to compute the posterior distributions for all our quantities of interest.

The research design supposes that participant *i* is asked about their support for policy *j* (e.g., allowing imported drugs from Canada) from issue *k* (e.g., drug costs) from

category m (e.g., economic policy). As such, we have a range of possible quantities of interest, depending on whether we focus on a particular policy or summarize across policies.

Estimating the Topic-Level Moderator

To estimate the proportion of respondents aware of the parties' positions on the issue, we rely on the random subsample of respondents (N = 252) who only answered awareness questions and were not exposed to any policy opinion questions. This approach rules out the possibility of spillover between issues causing post-treatment bias due to reliance on the post-treatment measures of awareness. We fit the random effects model $\Pr(\text{Aware}_{ij}) = \text{logit}^{-1}(\alpha_j)$, where $\alpha_j \sim N(\mu, \sigma^2)$ using penalized maximum likelihood (Chung et al. 2013). We use the conditional mode of the random effects (Bates et al. 2015) as our estimate of α_j and transform these estimates from the logit scale to the probability scale. This gives us an estimate of the proportion of respondents aware of the parties' relative positions on each issue.[14]

[14] Using the average of posterior simulations with Stan (Carpenter et al. 2017) produces nearly identical estimates of the awareness of the parties' positions on each issue.

Results

While virtually all previous work has estimated the treatment effect for a handful of policies or perhaps even a single policy, we estimate the treatment effect for 48 different policies. To illustrate the variation in treatment effects, we highlight two policies: marijuana legalization and changing federal standards to speed up the review of prescription drugs. For marijuana legalization, we estimate a treatment effect of about 0.10 [-0.16, 0.32; 90% percentile interval] scale points. The posterior chance that this effect is positive is only 78%. For the review of prescription drugs, we estimate a treatment effect of about 0.66 [0.39, 0.93] points on our seven-point scale—seven times larger than for marijuana legalization. The posterior chance that this effect is positive is 99%. Had we focused on either of these issues alone, we would reach different conclusions about the effect of a partisan cue. Of course, the large difference in the treatment effect immediately raises the question: why?

Figure 2, below, shows the estimates of all 48 treatment effects, along with 90% percentile intervals, for each policy. The policies are grouped by category (social policy, foreign policy, economic policy). Within each category, policies are sorted by the estimate of the treatment effect. The color indicates the level of awareness, which ranges from a low of 46% to a high of 83%.

Notably, the three policies with the highest levels of awareness of the parties' positions were all social policies: banning immigration of Muslims, federal funding for

abortion under Medicaid, and legalizing marijuana. For these three issues, we estimate small treatment effects of about 0.1 points along the seven-point scale.

Three of the issues with the lowest levels of awareness are economic issues that involve allowing imported prescription drugs from Canada, the creation of an optional retirement plan that transfers across jobs, and speeding up the federal review of prescription drugs. For these three issues, we estimate much larger treatment effects of about 0.6 points along the seven-point scale. At first glance, awareness seems clearly related to the size of the treatment effect. Additionally, social issues tend to have the highest levels of awareness and the smallest treatment effects, while economic issues tend to have the lowest levels of awareness and the largest treatment effects. This is consistent with previous research on pretreatment effects in the domain of party cues (Slothuus 2015) and differences in public opinion on social and economic issues (Johnston, Lavine, and Federico 2017).

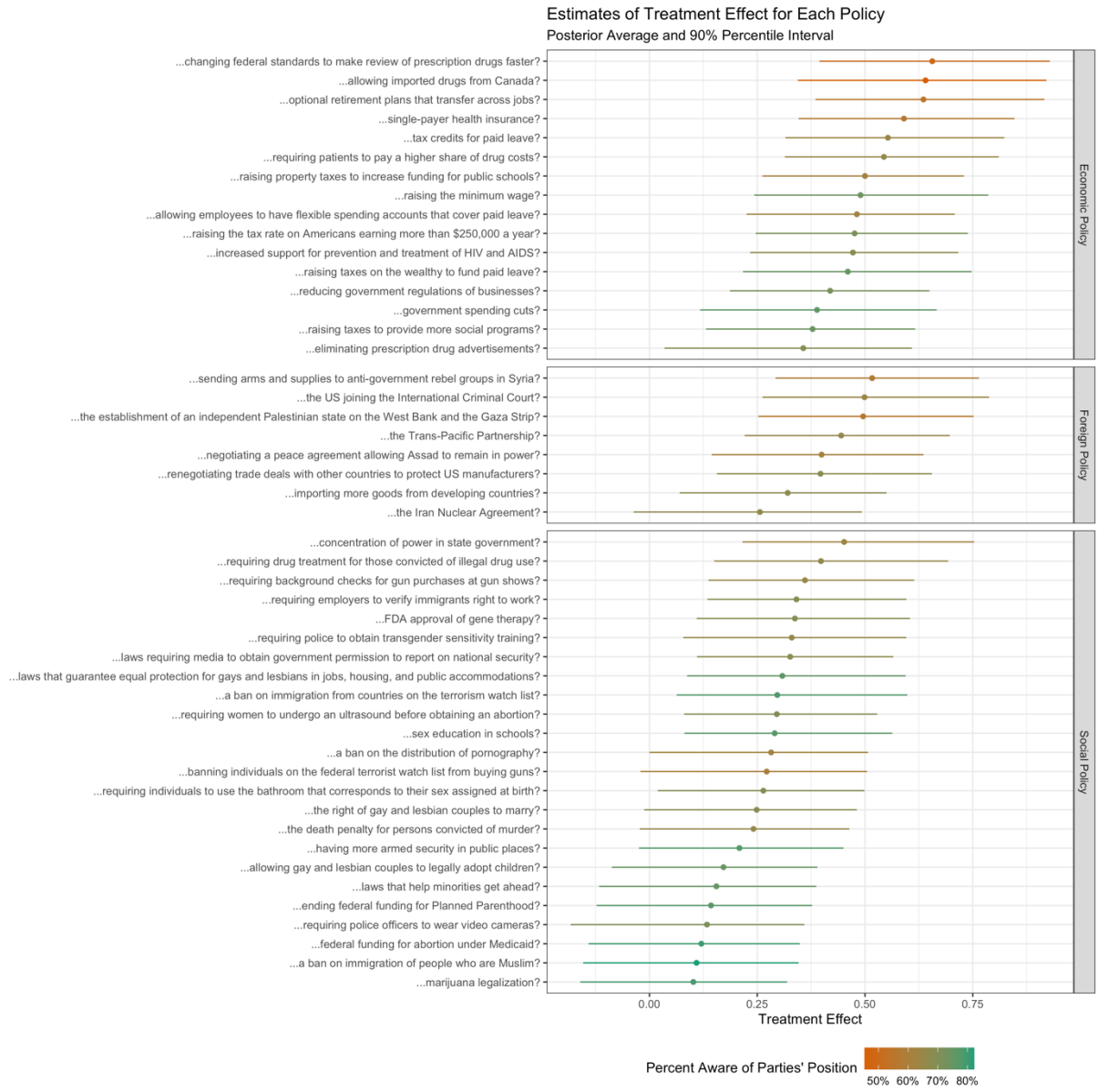

Figure 2: This figure shows the estimates of the treatment effects for each policy. The policy stems are separated into the three policy categories and ordered within each category from the largest estimate (top) to the smallest estimate (bottom). The color indicated the percent aware of the parties' positions on the issues. Green points and lines indicate high awareness and orange points and lines indicate low awareness. While partisan cues have generally positive effects, the magnitude of the effect varies substantially across issues.

Figure 3 shows directly how the treatment effects vary with the level of prior awareness. The scatterplot shows the treatment effect and 90% confidence intervals for each policy across the awareness of the parties' relative positions on the policy. The scatterplot clearly shows a negative relationship between treatment effect size and the level of awareness (Pearson's r = -0.77). For policies with the highest levels of awareness, the treatment effects are about 0.2 points along the seven-point scale, but about 0.6 for the policies with the lowest levels of prior awareness.

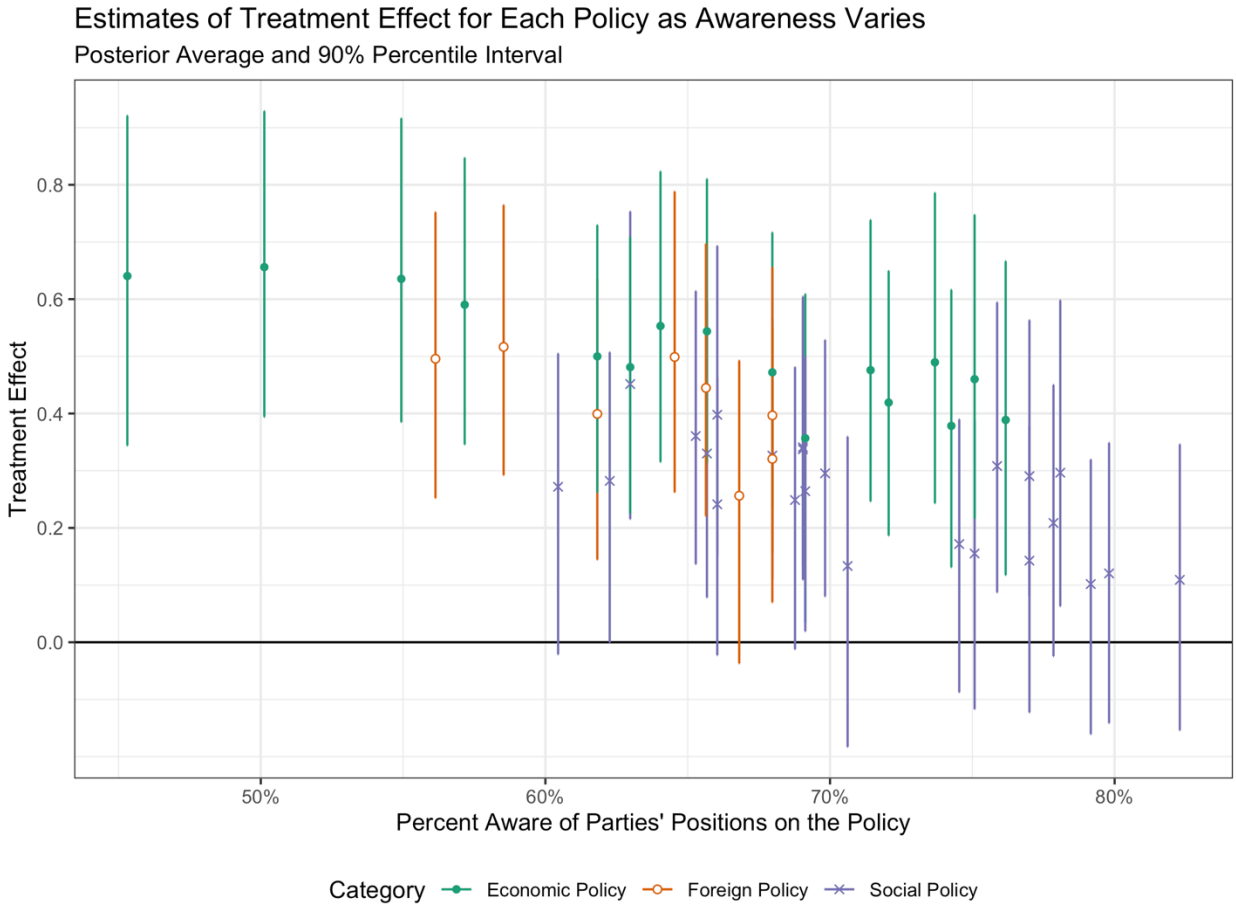

Figure 3: This shows the relationship between estimate of the treatment effect for each policy and the percent of respondents aware of the parties' positions on the issue. The color and shape of the lines and points

indicate the category to which each policy belongs. The effect of the partisan cue varies substantially, and the prior awareness of the parties' positions explains much (about 60%) of that variation.

The product term in the fitted statistical model allows us to formally test the hypothesis that higher levels of awareness are associated with smaller treatment effects. Figure 4 shows the treatment effect, averaging across policies and categories, as awareness varies. A one standard deviation increase in awareness decreases the treatment effect by about 0.16 units, a two standard deviation increase by about 0.31 units, and a minimum-to-maximum increase by about 0.37 units (from a treatment effect of 0.23 to 0.60). The posterior probability that the coefficient for the product term is negative is 0.94—moderate evidence for our interaction hypothesis.

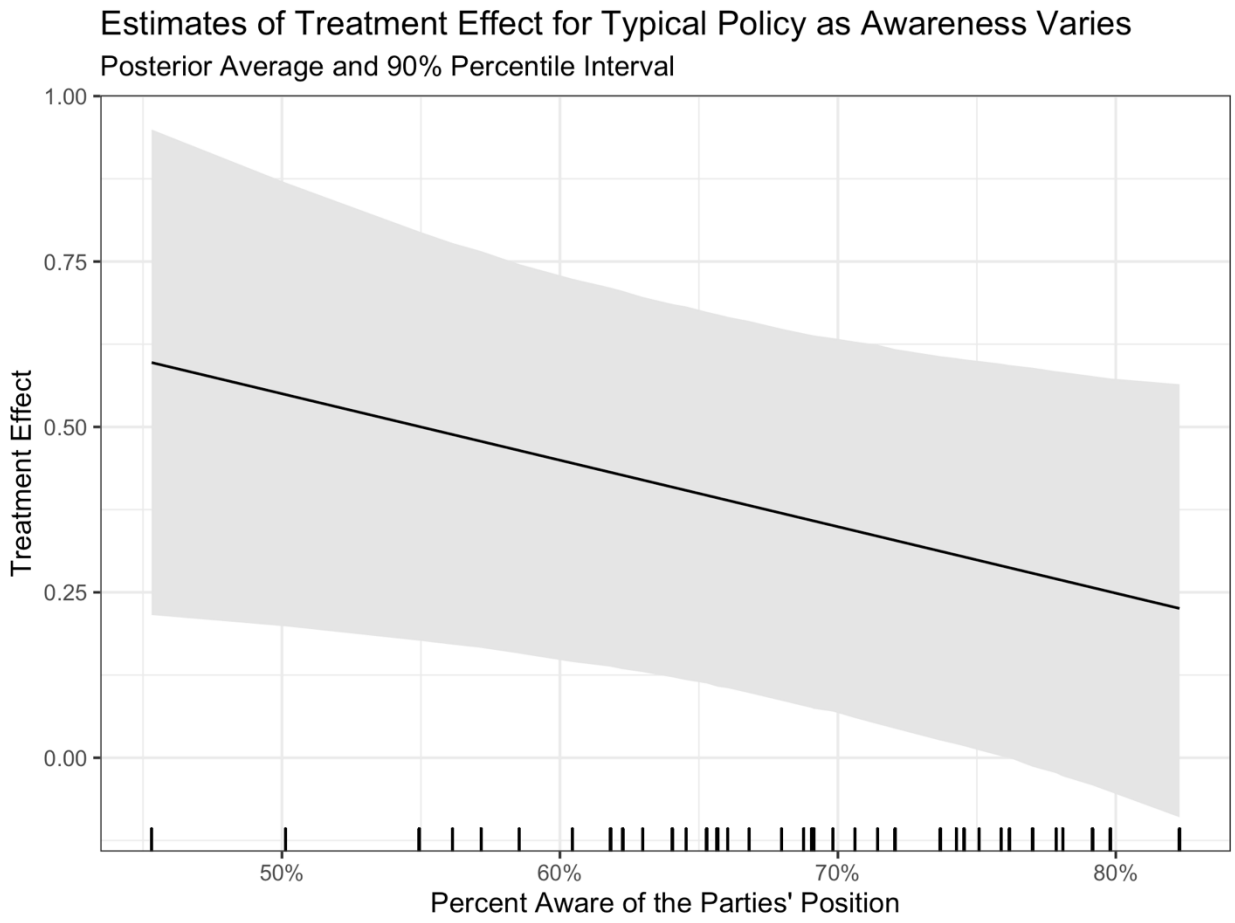

Figure 4: This shows the treatment effect as awareness varies averaging across policies. Notice that the effect is largest for issues with the lowest levels of awareness.

We can also use our model to provide more generalizable estimates of how treatment effects vary across categories of issues. Figure 5 shows the effects by category as awareness varies. The effects are largest for economic policy and smallest for social policy, while the effects for foreign policy fall in between the two. However, after accounting for awareness, the differences across categories are modest. The largest difference is between social policy and economic policy. The treatment effect is about 0.17 units larger for

economic than for social policy, but the evidence for a positive difference is moderate, with a posterior probability of 93%. The treatment effect is about 0.08 units smaller for foreign policy than for economic policy, but the posterior probability that this difference is negative is only 77%. Similarly, the treatment effect is about 0.09 units smaller for social policy than for foreign policy, and the posterior probability of a negative difference is 76%.

Compared to the estimate of the treatment effect of awareness, the differences across categories are quite modest. While there is no good default method to compare the differences in the treatment effects across awareness to the differences across the qualitative categories, we suggest comparing a one-SD increase in awareness to a change in category. In this case, the *largest* difference across categories (economic policy to social policy) is similar to the difference for a one-SD increase in awareness (0.17 versus 0.16). The largest difference across categories is less than half the size of the largest difference across values of awareness (0.17 versus 0.37). Thus, awareness seems to describe the variation in the treatment effects better than the category.

While prior awareness explains about 50% of the variation in the treatment effect, some policy-level variation remains unexplained. Other features of the policies that are not captured by issue category, such as residual variance in how easy, hard, or moral the policy is, might contribute to the magnitude of the treatment effects. We leave this question to further research. However, the key conclusion remains stark: while the treatment effect of a partisan cue is generally positive, the effect varies *substantially* across issues, ranging

from small (but probably positive) to quite large. Treatment effects also vary substantially within issue categories that are often treated as fundamentally distinct from each other.

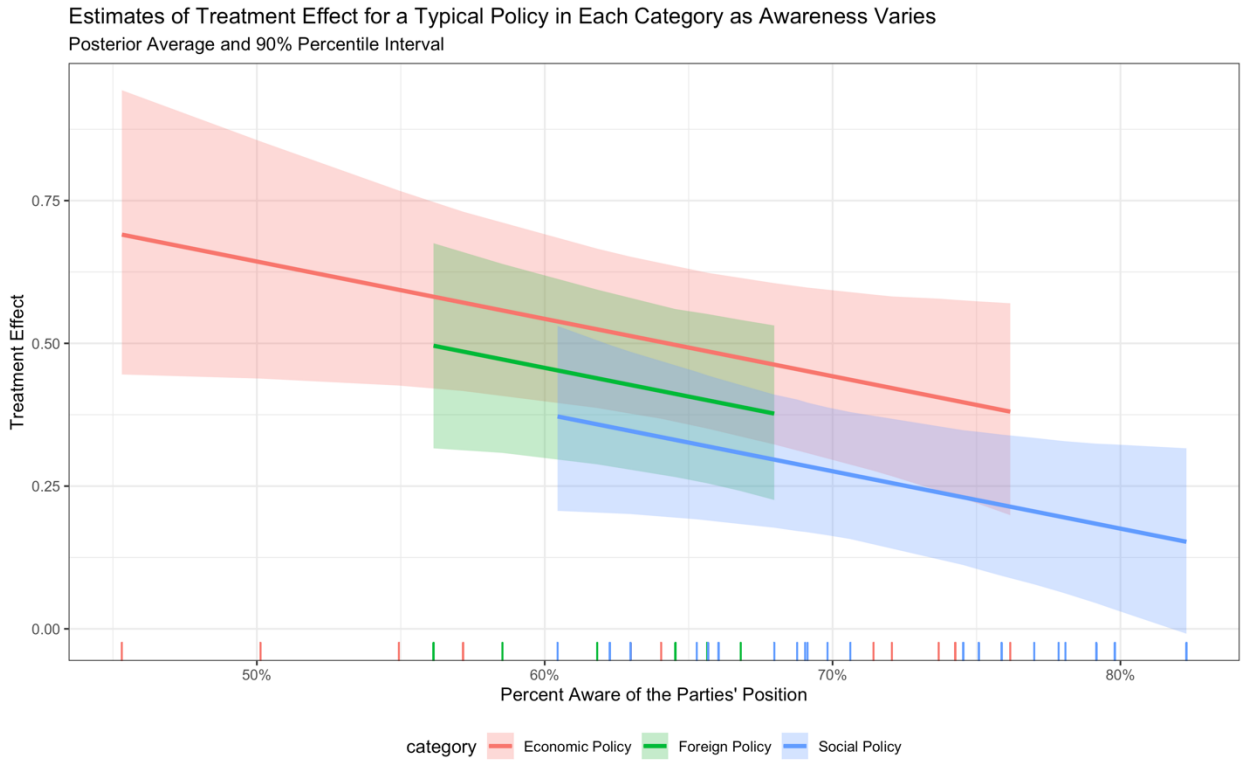

Figure 5: This shows the relationship between the treatment effect and awareness for each category of policies. Notice that while the treatment effect increases with awareness, treatment effects are smallest for social policies and largest for economic policies.

Comparing Topic-Sampling to Multi-Armed Studies

Researchers often wish to make claims about how effects differ across different types of issues, frequently by comparing social and economic issues. Consistent with expectations from past literature, we found evidence of modestly larger effects for economic issues than for social issues (conditional on awareness). With our data, it is

possible to compare our estimates to those that might be reached by a more common approach. In the multi-armed study, researchers typically select one social issue and one economic issue to compare. Yet, the variability of issues within each of these categories means that the results may depend heavily on the particular issues that are selected. To illustrate, we examine all of the hypothetical multi-armed studies from our sample of policies. We have 24 social issues and 16 economic issues, which produces $24 \times 16 = 384$ economic-social policy pairs that could be selected for a hypothetical study. Taking our estimates from Figure 2 as correct, we consider how the treatment effect varies across the possible pairs. Figure 6 shows that the 384 possible studies vary substantially. The economic policy has a larger effect in most pairs (376 of 388; 98%), but the magnitude of the difference varies considerably. About 25% of the possibilities have a difference of less than 0.15 and about 25% have a difference larger than 0.34. For comparison, our approach suggests a difference of about 0.17. Our approach, though, accounts for the level of awareness—recall that respondents have more awareness of the parties' positions social policies than economic policies.

To account for the variation in awareness across policies, we consider the findings that would have been reached by comparing only the economic-social policy pairs in which the levels of awareness fall within five percentage points of each other, what we might think of as "matched pairs." Our sample of policies has 125 of these matched pairs. Again, the economic policy almost always has a larger effect (123 of 125; 98%). The average difference is 0.19, which is closer to our estimate of 0.17. However, there is still

considerable variation across the possible matched pairs. About 25% of the possibilities have a difference of less than 0.12 and about 25% have a difference larger than 0.25.

For context, it is helpful to consider the size of sampling errors of the estimates across these studies. For a multi-armed sample survey with 3,000 respondents (treatment and control conditions in two topics), the sampling error for the difference in the treatment effects between topics would be about 0.14. The researcher would no doubt carefully model and report this source of estimation error. However, the possible errors from topic choice would (by necessity) remain unmodeled and unknown. Because we have estimates for a representative collection of topics, we know that for the partisan cues application, the additional error due to topic choice would be about 0.13 (the SD of the 384 unmatched pairs), which is almost as large as the sampling error. Clever researchers might shrink this to 0.09 (the SD of the 125 matched pairs) by identifying a pair of issues with similar levels of awareness.

In general, researchers using a multi-armed study cannot know the relative contribution of topic selection to the error in the estimates. The contribution error might be small and safely ignored. As in the case of partisan cues, it might be comparable to sampling error. In other applications, it might be much larger than sampling error. Without describing the variation in treatment effects across a representative collection of topics, the researcher cannot know how well their topics represent the general concept they are studying.

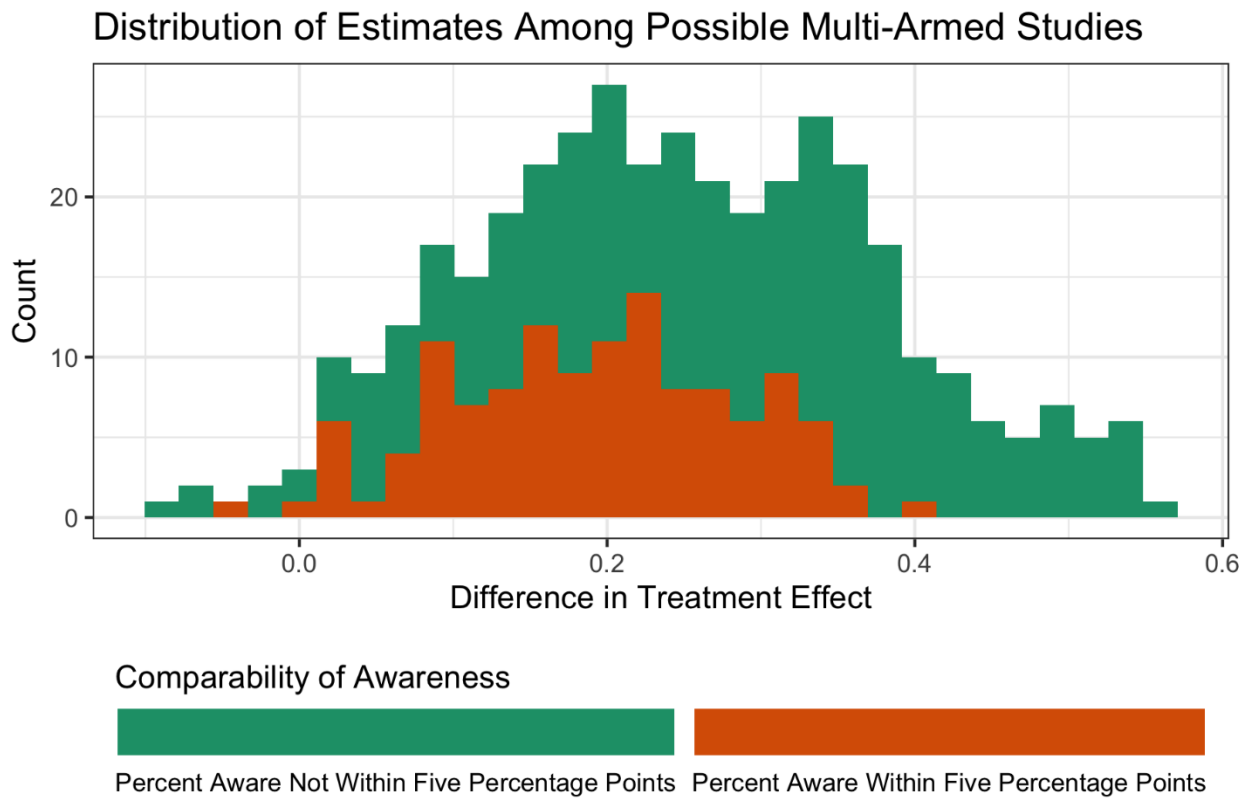

Figure 6: This shows the distribution of differences in treatment effects among the possible policy pairs that researchers might study using a multi-armed design. The orange bars show the distribution for matched pairs with similar levels of awareness. The total height of the orange and green bars shows the distribution for unmatched pairs. Importantly, the variation across policy pairs is similar to the sampling variations in a typically-sized sample survey and thus cannot be safely ignored.

Conclusion

Concerns about the external validity of experiments conducted on convenience samples has been a "near obsession" for political scientists (McDermott 2002, 334). While progress has been made toward addressing this concern (Berinsky, Huber, and Lenz 2012; Coppock 2018; Coppock, Leeper, and Mullinix 2018; Mullinix et al. 2016), much less attention has been paid to another crucial aspect of external validity – the topics that

researchers choose to study. As illustrated by the quotes in the introduction, this challenge to external validity has largely been taken as a necessary limitation to experimental research. By introducing topic sampling, we provide scholars with the tools to directly address this common threat to external validity by estimating the variability and correlates of treatment effect size.

The ability to examine how treatment effects vary across topics promises to yield new insights into a variety of important questions. The results presented here demonstrate that partisan cue effects vary considerably across policies, having large effects in some cases and minimal effects in others. Our analysis suggests that awareness explains much, but certainly not all, of the variance in these effects. Further, we find modest differences in treatment effects by issue category, even after accounting for awareness. We expect that further research using our approach will shed more light on when partisan cues matter more and when they matter less.

There are many other ways in which topic sampling can help move theoretical debates forward. For example, information might change policy attitudes on "hard" issues like foreign aid (Gilens 2001), but not on "easy" issues like immigration and welfare (Hopkins, Sides, and Citrin 2019; Kuklinski et al. 2000). The public might hold politicians and parties accountable for their stances on "crystallized" (Tesler 2015) or "moral" issues (Goren and Chapp 2017), but merely follow the leader on other topics (Lenz 2009). Clearly, the ability to generalize beyond just one or two specific topics does not merely solve a

methodological problem but promises to lend deeper insight into fundamental questions like voter competence and democratic representation.

While we have focused our attention on survey experiments, researchers can implement topic sampling in lab experiments, field experiments, or even observational studies. For example, researchers conducting a field experiment on whether issue appeals motivate voter participation can easily randomize the issue featured in the GOTV message, whether delivered by postcard, phone, or face-to-face. For another example, researchers studying moral conviction, an aspect of attitude strength, are typically forced to rely on observational designs. To increase generalizability and statistical power, moral conviction researchers often study multiple issues at once (Ryan 2014, 2017). These researchers can use topic sampling in their observational studies to increase the generalizability of their findings.

One clear limitation of our approach is that topic-level moderators are observational, rather than experimentally manipulated. As a result, examining topic-level moderators faces all of the inferential problems faced by researchers using observational individual-level moderators (for discussion, see Green and Kern 2012; Kam and Trussler 2016). For example, in our case, it seems likely that levels of awareness covary with attitude strength. Indeed, the issues with the highest levels of awareness involved abortion, marijuana, and immigration – three social issues that could be characterized as salient, moral, or easy issues that likely engender strong attitudes. In contrast, our three issues with the lowest levels of awareness involved retirement plans and prescription drug plans.

These three economic issues would likely be classified as hard or non-salient issues that tend to generate weak attitudes. Thus, while our design lends new evidence as to the generalizability of treatment effects and insight into how these effects vary, it faces the same problems as common moderation designs. Nonetheless, the design gives a clear indication of the amount of variability in treatment effects across topics.

One promising avenue for further development is designing studies that compare one or more pairs of policies that are matched on a variety of characteristics (e.g., awareness), but differ on one dimension (e.g., social vs. economic). As demonstrated by our analysis of social and economic issues, this approach offers more precise and better controlled comparisons between issues. Yet, a major barrier to this approach is our lack of systematic knowledge of the ways in which issues differ. Scholars have come up with a variety of typologies, such as easy vs. hard, principled vs. pragmatic, moral, culture war, and symbolic issues. But many of these categories remain ill-defined and none have laid out a clear way to measure these differences between issues. Thus, to better understand how issues matter in public opinion, scholars will need to devote more effort to studying issues themselves as an object of inquiry and formalizing these issue typologies.

Finally, while we have made an effort to develop a population of policies relevant to public opinion research, there is considerable room for further theoretical and empirical development of the populations of topics that are relevant to particular research questions. Defining the population may be easier for some topics, such as countries, though the relevant set (e.g., democracies) will vary by the research question. The task is more

challenging for more amorphous populations, such as policies, that may evolve over time. For example, the population we defined based on polls in 2016 may not remain relevant for long as new issues emerge into the discussion or old issues are resolved. Clearly, developing and maintaining populations of the topics we consider relevant to our theories will require sustained effort, but this work is necessary to understanding the scope and limitations of our theories, now that we have the tools to systematically do so.

References

Almond, Gabriel A. 1950. *The American People and Foreign Policy*. New York: Harcourt, Brace and Company.

Arceneaux, Kevin. 2007. "Can Partisan Cues Diminish Democratic Accountability?" *Political Behavior* 30(2): 139–60.

Bakker, Bert N., Yphtach Lelkes, and Ariel Malka. 2020. "Understanding Partisan Cue Receptivity: Tests of Predictions from the Bounded Rationality and Expressive Utility Perspectives." *Journal of Politics* 82(3): 1061–77.

Barabas, Jason, and Jennifer Jerit. 2009. "Estimating the Causal Effects of Media Coverage on Policy-Specific Knowledge." *American Journal of Political Science* 53(1): 73–89.

———. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104(02): 226–42.

Barabas, Jason, Jennifer Jerit, William Pollock, and Carlisle Rainey. 2014. "The Question(s) of Political Knowledge." *American Political Science Review* 108(04): 840–55.

Barber, Michael, and Jeremy C. Pope. 2019. "Does Party Trump Ideology? Disentangling Party and Ideology in America." *American Political Science Review* 113(1): 38–54.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statistical Software* 67(1): 1–48.

Bélanger, Éric, and Bonnie M. Meguid. 2008. "Issue Salience, Issue Ownership, and Issue-Based Vote Choice." *Electoral Studies* 27(3): 477–91.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.Com's Mechanical Turk." *Political Analysis* 20(3): 351–68.

Carmines, Edward G., and James A. Stimson. 1980. "The Two Faces of Issue Voting." *American Political Science Review* 74(01): 78–91.

Chong, Dennis, and James N. Druckman. 2010. "Dynamic Public Opinion: Communication Effects over Time." *American Political Science Review* 104(04): 663–80.

———. 2012. "Counterframing Effects." *The Journal of Politics* 75(01): 1–16.

Chung, Yeojin et al. 2013. "A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models." *Psychometrika* 78(4): 685–709.

Clifford, Scott, Ryan M. Jewell, and Philip D. Waggoner. 2015. "Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology?" *Research & Politics* 2(4): 1–9.

Coppock, Alexander. 2018. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods*: 1–16.

Coppock, Alexander, and Donald P. Green. 2015. "Assessing the Correspondence between Experimental Results Obtained in the Lab and Field: A Review of Recent Social Science Research." *Political Science Research and Methods* 3(01): 113–31.

Coppock, Alexander, Thomas J. Leeper, and Kevin J. Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates across Samples." *Proceedings of the*

National Academy of Sciences: 201808083.

Delli Carpini, Michael X., and Scott Keeter. 1996. *What Americans Know about Politics and Why It Matters*. Yale University Press.

Ditto, Peter H. et al. 2019. "At Least Bias Is Bipartisan: A Meta-Analytic Comparison of Partisan Bias in Liberals and Conservatives." *Perspectives on Psychological Science* 14(2): 273–91.

Druckman, James N., and Thomas J. Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4): 875–96.

Edwards III, George C., William Mitchell, and Reed Welch. 1995. "Explaining Presidential Approval: The Significance of Issue Salience." *American Journal of Political Science* 39(1): 108.

Feldman, Stanley, and Christopher Johnston. 2014. "Understanding the Determinants of Political Ideology: Implications of Structural Complexity." *Political Psychology* 35(3): 337–58.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science (New York, N.Y.)* 345(6203): 1502–5.

Gilens, Martin. 2001. "Political Ignorance and Collective Policy Preferences." *American Political Science Review* 95(2): 379–96.

Goren, Paul, and Christopher Chapp. 2017. "Moral Power: How Public Opinion on Culture War Issues Shapes Partisan Predispositions and Religious Orientations." *American*

Political Science Review 111(01): 110–28.

Green, D. P., and H. L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491–511.

Grieco, Joseph M., Christopher Gelpi, Jason Reifler, and Peter D. Feaver. 2011. "Let's Get a Second Opinion: International Institutions and American Public Support for War1." *International Studies Quarterly* 55(2): 563–83.

Hopkins, Daniel J., John Sides, and Jack Citrin. 2019. "The Muted Consequences of Correct Information about Immigration." *The Journal of Politics* 81(1): 315–20.

Jerit, Jennifer. 2009. "How Predictive Appeals Affect Policy Opinions." *American Journal of Political Science* 53(2): 411–26.

Jerit, Jennifer, and Jason Barabas. 2012. "Partisan Perceptual Bias and the Information Environment." *The Journal of Politics* 74(03): 672–84.

Jerit, Jennifer, Jason Barabas, and Toby Bolsen. 2006. "Citizens, Knowledge, and the Information Environment." *American Journal of Political Science* 50(2): 266–82.

Jerit, Jennifer, Jason Barabas, and Scott Clifford. 2013. "Comparing Contemporaneous Laboratory and Field Experiments on Media Effects." *Public Opinion Quarterly* 77(1): 256–82.

Johnston, Christopher D., Howard Lavine, and Christopher M. Federico. 2017. *Open versus Closed: Personality, Identity, and the Politics of Redistribution*. Cambridge University Press.

Johnston, Christopher D., and Julie Wronski. 2015. "Personality Dispositions and Political Preferences Across Hard and Easy Issues." *Political Psychology* 36(1): 35–53.

Kam, Cindy D. 2005. "Who Toes the Party Line? Cues, Values, and Individual Differences." *Political Behavior* 27(2): 163–82.

Kam, Cindy D., and Marc J. Trussler. 2016. "At the Nexus of Observational and Experimental Research: Theory, Specification, and Analysis of Experiments with Heterogeneous Treatment Effects." *Political Behavior*: 1–27.

Kay, Matthew. 2019. "Tidybayes: Tidy Data and Geoms for Bayesian Models."

Kertzer, Joshua D., and Ryan Brutger. 2016. "Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory." *American Journal of Political Science* 60(1): 234–49.

Kuklinski, James H. et al. 2000. "Misinformation and the Currency of Democratic Citizenship." *The Journal of Politics* 62(3): 790–816.

Lenz, Gabriel S. 2009. "Learning and Opinion Change, Not Priming: Reconsidering the Priming Hypothesis." *American Journal of Political Science* 53(4): 821–37.

———. 2012. *Follow the Leader?: How Voters Respond to Politicians' Policies and Performance*. Chicago: University of Chicago Press.

Lindner, Nicole M., and Brian A. Nosek. 2009. "Alienable Speech: Ideological Variations in the Application of Free-Speech Principles." *Political Psychology* 30(1): 67–92.

McDermott, Rose. 2002. "Experimental Methodology in Political Science." *Political Analysis* 10(4): 325–42.

Mooney, Christopher Z. 2001. *The Public Clash of Private Values: The Politics of Morality Policy*. Chatham House Publishers.

Mooney, Christopher Z., and Richard G. Schuldt. 2008. "Does Morality Policy Exist? Testing a Basic Assumption." *Policy Studies Journal* 36(2): 199–218.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2016. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2(02): 109–38.

Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton University Press.

Nelson, Thomas E. et al. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91(03): 567–83.

Ryan, Timothy J. 2014. "Reconsidering Moral Issues in Politics." *The Journal of Politics*: 1–18.

———. 2017. "No Compromise: Political Consequences of Moralized Attitudes." *American Journal of Political Science* 61(2): 409–23.

Sears, David O., and Nicholas A. Valentino. 1997. "Politics Matters: Political Events as Catalysts for Preadult Socialization." *The American Political Science Review* 91(1): 45.

Simas, Elizabeth N., Kerri Milita, and John Barry Ryan. "Ambiguous Rhetoric and Legislative Accountability." *Journal of Politics*.

Slothuus, Rune. 2016. "Assessing the Influence of Political Parties on Public Opinion: The Challenge from Pretreatment Effects." *Political Communication* 33(2): 302–27.

Tavits, Margit. 2007. "Principle vs. Pragmatism: Policy Shifts and Political Competition."

American Journal of Political Science 51(1): 151–65.

Tesler, Michael. 2015. "Priming Predispositions and Changing Policy Positions: An Account of When Mass Opinion Is Primed or Changed." *American Journal of Political Science* 59(4): 806–24.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC." *Statistics and Computing* 27: 1413–32.