Political Behavior (2024) 46:1233–1256
https://doi.org/10.1007/s11109-023-09870-1

ORIGINAL PAPER

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

Scott Clifford[1] · **Thomas J. Leeper**[2] · **Carlisle Rainey**[3]

Accepted: 1 March 2023 / Published online: 26 March 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

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.

Keywords Generalizability · External validity · Survey experiments · Partisan cues

✉ Scott Clifford
sclifford@uh.edu

Thomas J. Leeper
thosjleeper@gmail.com

Carlisle Rainey
crainey@fsu.edu

[1] University of Houston, 3551 Cullen Blvd. Rm 447, Houston, TX 77204, USA

[2] London School of Economics, Houghton Street, London, UK

[3] Florida State University, 600 W. College Avenue, Tallahassee, FL 32306, USA

Introduction

Experiments are on the rise in political science, but concerns remain about external validity, whether in terms of sample, context, or treatments. Public opinion researchers have studied how generalizability is affected by sample characteristics (Berinsky, Huber, and Lenz 2012; Coppock, Leeper, and Mullinix 2018) and study context (e.g., Barabas and Jerit 2010; Coppock and Green 2015; Jerit, Barabas, and Clifford 2013). Yet, less attention has been paid to an 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 topic – in this case, an issue, such as foreign aid – and design topic-specific stimuli (e.g., Gilens 2001). In other cases, the topic might involve a specific country (e.g., East Timor; Grieco et al. 2011) or social group (e.g., Arabs; Lindner and Nosek 2009). 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 the 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 external validity. Continuing with the example of political issues, authors have described their results as "circumscribed by our focus on a single issue" (Chong and Druckman 2012, 14). Researchers sometimes address this threat to external validity by reporting multiple studies or multiple arms of a study, each on a different topic. However, this approach is costly both in time and resources, and only incrementally increases our knowledge of the generalizability of the results across topics. Moreover, the focal topics are typically selected by the researcher as ideal tests of the theory, perhaps inflating the likelihood of supportive findings.

This is not only a problem for external validity, but also for theory development and testing. Indeed, in the public opinion literature, many researchers have theoretical expectations as to how an effect might vary across issues. 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 study corresponding with each of the sampled topics, and randomizing respondents into (at least) one arm of the study. We refer to this approach as "topic

sampling" (for related discussion, see Wells and Windschitl 2016; Tappin 2022). 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 estimated 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 issues, 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 prior awareness of the parties' positions on the issue and the type of issue being considered (e.g., social vs. economic). We conclude with advice on how to conduct a topic sampling experiment, discussion of application to observational designs and field experiments, as well as sampling other aspects of the stimuli or context of a study.

How Topics Vary and Why It Matters

When designing a study, researchers typically apply their theory to a particular topic, 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 topic (e.g., country or social group) from a population. The target population will itself depend 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 well known that people may respond to the same treatment in different ways and the analysis of individual-level moderators has been central to experimental research (Kam and Trussler 2016). Similarly, any given person might react differently to the same 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 topic to the relevant population of topics. But the variation in effects across topics that threatens generalizability can also inform 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).

What Are the Solutions?

Researchers, of course, have been aware of this problem and have attempted to deal with it using multi-armed studies 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 randomizing 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 (2021) randomize between transgender rights and business incentives to test whether effects differ across "principled" and "pragmatic" issues. In an observational, within-subjects design, Ryan (2017, 5) evaluates the effects of morally convicted attitudes on compromise across five different issues "to accumulate evidence across a broader array of issues… that are both putatively moral and non-moral." Thus, researchers using both experimental and observational designs seek variation across issues to test theoretical claims and to increase the generalizability of their findings.

Of course, the multi-armed study faces substantial shortcomings. If the goal is generalizability, including an additional issue offers only a marginal improvement. If the goal is theory testing, then it raises concerns about how well each issue represents the broader category. For example, it is 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.

Meta-analyses promise to leverage more data but face a number of problems. First, the available set of studies is likely subject to substantial publication bias. For example, an analysis of the Time-sharing Experiments in the Social Sciences (TESS) database revealed that statistically significant findings are dramatically more likely to be written up and published than null results (Franco, Malhotra, and Simonovits 2014). In the absence of a study database like TESS, meta-analytic estimates will 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. 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 and apply a tool to enhance generalizability that we refer to as "topic sampling." Combined with a hierarchical model, topic sampling allows the researcher to (1) precisely estimate individual treatment effects for many particular topics, (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).

Researchers must first identify a population of topics and develop a sampling frame (e.g., a list of all salient political issues), then draw a random sample of topics from that population. Within the experiment, respondents are first randomized into a topic, then randomly assigned to treatment or control. Just as researchers need a representative sample of survey respondents to draw conclusions about a large population of people, researchers need a representative collection of *particular topics* to draw conclusions about the population of topics. Just as a convenience sample of respondents might not represent the population, so might a convenient topic be unrepresentative of the larger collection.

Because topic sampling involves dividing survey respondents across many particular topics, the sample size for any individual topic will be small, rendering less precise estimates of topic-specific treatment effects on the outcome of interest. We address this challenge with a hierarchical model to borrow information across topics. We suggest thinking of the many topics as parallel studies. Across these studies, we expect the treatment effects to be *different, but similar.* Following Bullock, Imai, and Shapiro (2011), we formalize the *different-but-similar* assumption by treating the parameters for each parallel study as a draw from normal distribution. Then, while the treatment effects for each topic are different, we have a model for those differences. We can then characterize the typical treatment effect and the give-or-take around the typical treatment effect (i.e., the degree of similarity). Because we estimate the amount of similarity from the data, we can pool information across topics as warranted by the data (Gelman and Hill 2007, 252–59; Gelman, Hill, and Yajima 2012). Rather than limit our focus to a single topic (with, say, 1,000 observations) or conducting full-power studies of a handful of topics (with a total of 5,000 respondents), we can run a single (carefully designed) full-powered study of *many* topics using only, say, 2,000 observations. When the similarity across topics is high, we can obtain precise estimates for each topic, even with small samples for each topic. As the similarity decreases, so does our ability to obtain precise estimates. However, as similarity decreases, the representativeness of particular issues decreases as well, increasing the importance of estimating the treatment effect across a diverse collection of topics.

Combining this design and modeling strategy enables the researcher to compute several important quantities of interest.[1] 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 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 the 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. Importantly, researchers can modify the core hierarchical model to include other features such as control variables, nonlinear terms, interactions, and limited dependent variables. We summarize the tradeoffs involved in choosing between alternative designs in Table 1, below.

Table 1 Comparing Advantages of Alternative Study Designs

Quantity	Single-topic study	Multiple, separate studies	Topic sampling
Treatment effect for a particular topic	Excellent, but only for a single topic	Excellent, if all *k* studies have large sample sizes	Acceptable, but can produce reasonably precise estimates for a large number of topics. While the estimates for particular topics will be less precise than devoting every subject to a single topic, the researcher can increase the number of topics included in the study from one to 25 or 50 while perhaps only tripling the standard errors of the estimates
Typical treatment effect across topics in the population	Absent	Suggestive, at best	Excellent. Topic sampling gives a principled (either maximum likelihood or Bayesian) estimate of the average of the treatment effects across the topics in the population
Variation in treatment effects across topics	Absent	Suggestive, at best	Excellent. Topic sampling gives a principled (either maximum likelihood or Bayesian) estimate of the variation in the treatment effects across the topics in the population

[1] Modern software makes fitting the hierarchical model easy. For simple models, (restricted) maximum likelihood estimation works well (e.g., Bates et al. 2015). For more complex models, full posterior simulation offers a robust alternative (e.g., Bürkner 2017; Goodrich et al. 2020).

Sample Sizes

In some sense, we should not directly compare the required sample sizes for single-topic studies and topic-sampling studies. Indeed, the two achieve distinct goals. The goal of a single-topic study is modest: estimate the treatment effect for a particular topic. The goal of a topic-sampling study is more ambitious: estimate the typical treatment effect across a large collection of topics. But even though topic-sampling allows the researcher to obtain a more ambitious (and relevant) estimate, topic-sampling requires only a few more respondents to obtain similar precision.

To compare the two approaches, we focus on the sample sizes a researcher needs to achieve a certain standard error. Based on their experience, the literature, and formal analyses, researchers have a good sense of the sample sizes required to obtain sufficient precision for a single-topic study. Unfortunately, the exact precision of the topic-sampling design is difficult to predict because it depends on the variation of the treatment effects across topics. We offer a guideline: if the researcher is primarily interested in estimating the typical treatment effect in the population of topics, then we suggest they use 25 to 50 topics and increase the "usual" number of respondents by 20 to 50%. For example, if the researcher was thinking of running a 500-person study on a particular topic, they could instead implement a topic-sampling design with 25 to 50 topics and 600 to 750 respondents. When the researcher suspects they have an especially heterogeneous collection of topics, it becomes important to include more topics (e.g., 50 rather than 25). This design requires slightly more respondents and much more planning. But the return on the additional investment is large—rather than a precisely-estimated treatment effect for *a particular topic*, the researcher obtains (1) a precise estimate of the typical treatment effect from the population of topics, (2) estimates of the treatment effects for all 25 to 50 particular topics used in the study, and (3) an estimate of the variability across topics. Our sample size recommendations are conservative; if the variation in the treatment effects is low (relative to the error variance), then perhaps the researcher does not even need a larger sample for a topic-sampling design. The Appendix provides more details on these recommendations.

Depending on the application, the researcher can add topic-level predictors of the treatment effect to the model. Below, for example, we use the general public's awareness of the parties' positions on political issues as a predictor of the effect of a partisan cue. This effectively reduces the variability across topics and increases the precision of the estimates. For some applications, researchers can also consider assigning a single respondent to multiple topics for certain applications (for discussion, see Clifford, Sheagley, and Piston 2021).

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). Studies also vary in the amount of contextual information, the nature of the party cue, the sample, the dependent variable, and other factors. This variation makes 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. To illustrate, 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."[2]

Here, we directly examine how treatment effects vary across issues, while holding all other features of the design constant. We also test three key hypotheses developed from the literature reviewed above on how issues differ. First, building off of the literature on "pretreatment," or prior receipt of the treatment (Druckman and Leeper 2012), 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 prior receipt of the treatment. Thus, it remains unclear how much the effect of partisan cues varies with prior awareness.

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 holding that "on average, social issues are easy issues, and economic issues are hard issues" (Johnston and Wronski 2015, 46). In other words, citizens readily connect their predispositions to their views on social issues, but depend on elites to connect their predispositions to most economic issues. As a result, attitudes on social issues are stronger and more resistant to influence, while attitudes on economic issues are more susceptible to elite cues and rhetoric (see also Arceneaux 2007; Simas, Milita, and Ryan 2021; Tavits 2007). Third, we examine treatment effects for foreign policy. Scholars have argued that "[f]oreign affairs are distant from most voters' everyday concerns and thus are especially ripe for cue-giving by elite actors" (Guisinger and Saunders 2017, 425), yet there has been little systematic comparison between partisan cues on foreign policy and other topics. Thus, we expect that treatment effects will be larger for foreign policy than for social issues, though we have no clear expectation for differences between foreign policy and economics.

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

[2] 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.

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.

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 rely on theoretical and practical concerns to choose among various populations to which they can generalize. The alternative, however, is to select a single issue, or small number of issues, and make no claim about generalizability to other issues. 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, researchers might disagree about which 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).[3] By virtue of appearing in a recent survey, the issue is assured to be of interest to public opinion researchers and currently relevant to politics. Finally, the database includes a highly diverse set of questions that were 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.

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.[4] 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 154 unique policy questions.[5] While many questions covered hot-button issues, our population also

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

[4] 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.

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

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]

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.[7] 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.

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*>?

[6] Of the 48 policies we sampled (see below), 40% also showed up three years later in the 2019 Roper population, suggesting considerable stability in the population over time.

[7] To validate our coding, we compared assigned values to partisan differences in the data among untreated respondents. In 77% of the cases, we observed a statistically significant difference in the expected direction. For the remaining 23% of cases, there was no significant partisan difference. Thus, our coding reliably mapped onto partisan differences.

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." 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. Our design choice entails two assumptions. First, relative partisan cues are equally effective at motivating support and 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 baseline levels of partisan agreement. As we discuss in more detail below, we incorporate this information into our model to yield more precise estimates of the treatment effects. In the conclusion, we discuss how topic sampling could be incorporated into a wide variety of experimental designs.

Manipulation Check and Issue-Level Moderator

Following the measurement of the six policy attitudes, respondents were asked whether 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 prior receipt of the treatment 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 prior awareness that could not be influenced by the experiment. For clarity, from here on out we refer to these estimates as awareness.

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 eight foreign policy questions. Within each category, we sampled disproportionately to ensure a roughly even number of policies from within each issue for any given category (for details, see Appendix). The resulting sample 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). Policies are shown in Fig. 1, while further detail is available in the Appendix.

Respondent Sample

We recruited 3,500 respondents through Survey Sampling International (SSI, now Dynata) in the summer of 2018.[8] Respondents were randomly assigned to the primary experimental module (N = 3,250) or the awareness-only module (N = 252). SSI 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 $y \in \{1, 2, ..., 7\}$, we assume that

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

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 prior awareness), so that

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

[8] This study was reviewed and approved by the London School of Economics Research Ethics Committee. All respondents gave informed consent before participating in the study.

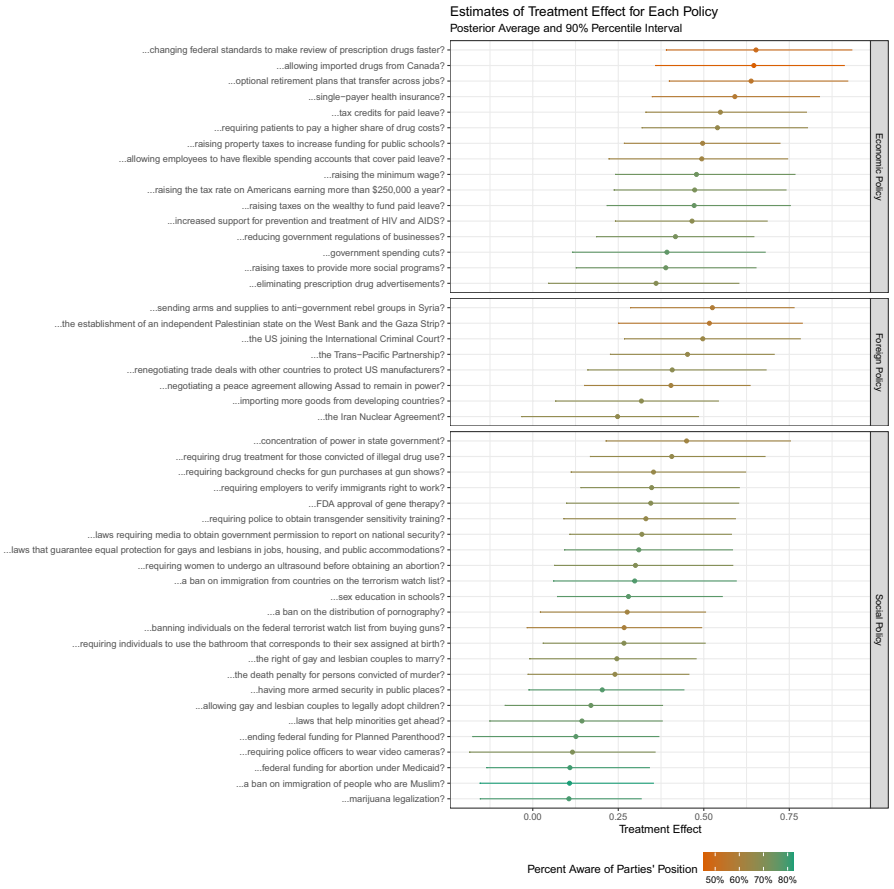

Fig. 1 This figure shows the estimates of the treatment effects for each policy. The policy stems are separated into the three policy categories and ordered within each category from the largest estimate (top) to the smallest estimate (bottom). The color indicates the percent aware of the parties' positions 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 (Color figure online)

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 β_j 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. We use Vehtari, Gelman, and Gabry's (2017) method to efficiently approximate the leave-one-out cross-validation (LOO-CV) and select among the possible models.

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, 90–95% 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 and model 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 used a random effects model (Chung et al. 2013, Bates et al. 2015) to estimate of the proportion of respondents aware of the parties' relative positions on each issue.[9]

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 of a positive effect 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

[9] 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.

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 1 shows the estimates of all 48 treatment effects for each policy along with 90% intervals. 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 the effects of prior receipt of the treatment in the domain of party cues (Slothuus 2016) and differences in public opinion on social and economic issues (Johnston et al., 2017).

Figure 2 shows directly how the treatment effects vary with the level of prior awareness. The scatterplot shows the treatment effect and 90% 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 $r = -0.73$). 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.

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 3 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.

We can also use our model to provide more generalizable estimates of how treatment effects vary across categories of issues. Figure 4 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 policies, but the evidence for

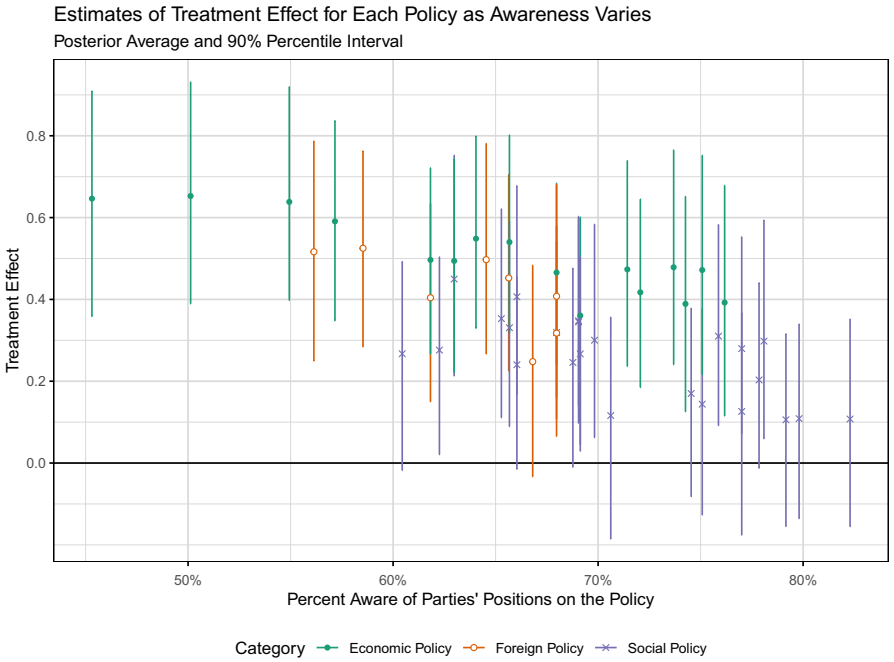

Fig. 2 This shows the relationship between the estimate of the treatment effect for each policy and the percent of respondents aware of the parties' positions on the issue. The color and shape of the lines and points indicate the category to which each policy belongs. The effect of the partisan cue varies substantially, and the prior awareness of the parties' positions explains much (about 60%) of that variation (Color figure online)

a positive difference is moderate, with a posterior probability of 94%. 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 75%. 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 78%.

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 compare a one-SD increase in awareness to a change in category. In this case, the *largest* difference across categories (economic policies to social policies) 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

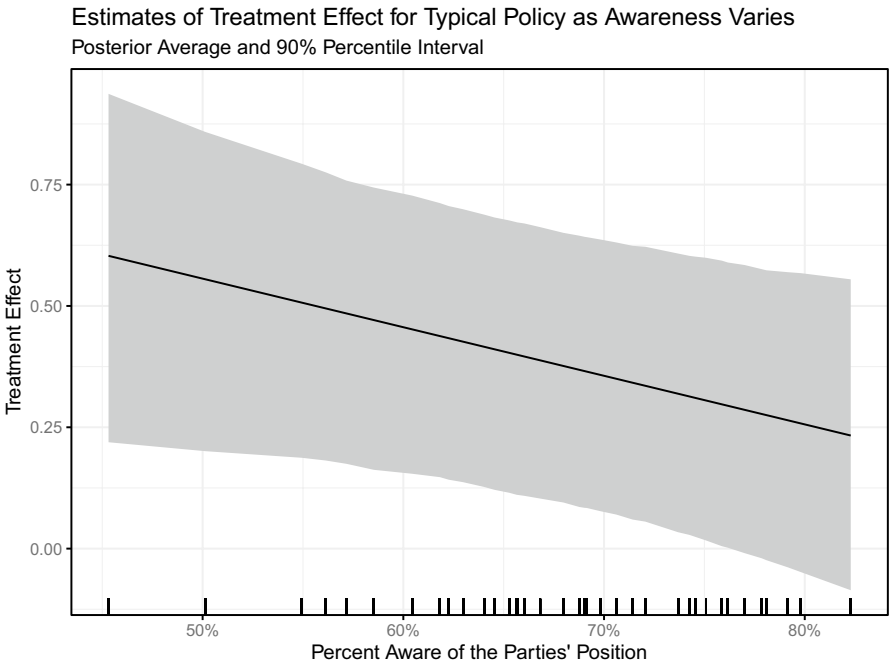

Fig. 3 This shows the treatment effect as awareness varies averaging across policies. Notice that the effect is largest for issues with the lowest levels of awareness

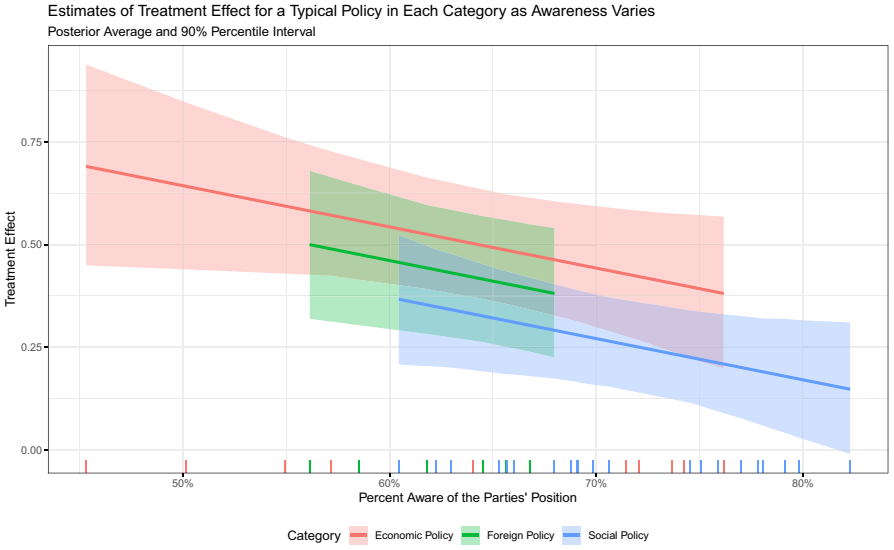

Fig. 4 This shows the relationship between the treatment effect and awareness for each category of policies. Notice that while the treatment effect increases with awareness, treatment effects are smallest for social policies and largest for economic policies

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.

Alternative Estimators

We argue that the hierarchical model provides a rich and helpful summary of the variation across topics, as well as precise estimates. This strategy does assume a particular parametric form, though we view this assumption as relatively weak. However, the hierarchical model is not *essential* for the topic sampling design. Instead, one could use an unbiased nonparametric estimate analogous to the AMCE estimator of Hainmueller, Hopkins, and Yamamoto (2014) in the context of conjoint experiments. To estimate the average treatment effect across topics, one can simply compute the mean of the difference-in-means across topics. Because we use a stratified sample of topics to increase the representativeness of the sample, we use a *weighted* mean of the differences-in-means, where the weights are design weights (i.e., the inverse of the probability of being sampled). For standard errors, we use a two-way cluster bootstrap by topic and respondent (Cameron, Gelbach, and Miller 2011). Because we (1) sample topics from a finite population and (2) use a stratified sample, these confidence intervals are conservative. For our application, the (weighted) average treatment effect across topics is 0.35 with a standard error of 0.13. Using an analogous hierarchical model that does not condition on awareness, we obtain a similar estimate of 0.36 with a standard error of 0.05. Thus, the two approaches offer similar estimates of the average treatment effect across topics, though our hierarchical model provides substantially more precision and a much richer summary of the heterogeneity across topics.

Comparing Topic-Sampling to Multi-Armed Studies

Researchers often use multi-armed studies to make claims about how social and economic issues differ from each other. Consistent with expectations, 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 multi-armed study that compares one social issue to one economic issue. 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 Fig. 1 as correct, we consider how the treatment effect varies across the 384 possible pairs. While the effect is larger for economic policy in most pairs (377 of 384; 98%), 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 0.35 (see Appendix for further detail). For comparison, our approach suggests a difference of about 0.17, after accounting for the level of awareness. Thus, multi-armed studies typically yield an answer in the correct direction, but the magnitude of the effect depends heavily on the issues selected.

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, much less attention has been paid to another crucial aspect of external validity – the context and stimuli that researchers choose for their study. When researchers design an experiment, they must fix many details or, equivalently, choose among many possible different-but-similar experiments that would test the same general claim. We argue that, when feasible, researchers should not select a single experiment from this collection, but many experiments. Researchers can use a hierarchical model to efficiently aggregate these many experiments in a way that explicitly generalizes beyond any single set of details without requiring a huge increase in the required total sample size. This approach provides 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 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. Indeed, we find modest differences in treatment effects by issue category, even after accounting for awareness. We expect that further research using topic sampling will shed more light on the crucial question of when the public is more likely to follow the leader (Lenz, 2009) and when it will hold politicians and parties accountable (Tesler, 2015).

While a wide variety of studies focus on a specific political issue, and thus would benefit from the form of topic sampling approach we describe here, we expect the method will have much wider application. Indeed, researchers often implicitly sample from a variety of constructs to set the context for their designs. For example, scholars studying foreign policy often attach an experimental manipulation to a specific country, and there is debate over whether this choice affects the results (for discussion, see Brutger et al. 2022). Others may randomize between two social groups to test the effects of shared ethnicity and ideological distance (e.g., Lindner and Nosek 2009), but effects might vary considerably depending on the group selected, making topic sampling particularly important (for discussion, see Brandt and Crawford 2019). There has also been extended debate over whether and when factual corrections are effective or backfire (Guess and Coppock 2020; e.g., Nyhan and Reifler 2010; Swire-Thompson, DeGutis, and Lazer 2020; Wood and Porter 2019). While numerous tests have been conducted, scholars might reach clearer conclusions if

they are able to first define the population of topics that might be subject to correction, then sample from that population.

The examples above all involve sampling features of the context of the experiment, but the same design and method can be used to sample experimental stimuli as well. For example, scholars studying framing often carefully select a "strong" frame and a "weak" frame from media coverage of a topic (e.g., Chong and Druckman 2007). Rather than handpicking a pair of frames to contrast, researchers could sample from the full population and examine the characteristics that determine the strength of a frame. Similarly, scholars might randomly sample from the population of candidate ads run during a campaign and examine the variability in persuasiveness. There are surely many more cases in which experimental stimuli can be sampled from a defined population.

While we think topic sampling has broad applicability, it is not always a necessary or appropriate design feature. Topic sampling is likely best thought of as answering "second generation" questions (Kam and Trussler 2016). That is, when researchers are seeking an initial test of a novel theoretical claim, it may make sense to choose a single topic as an ideal test case. However, once the plausibility of that claim has been established, topic sampling is a useful tool for establishing the generalizability of that effect. In the case of partisan cues, the debate is no longer whether they have an effect, but when and why. Similarly, debates over the effects of factual information or corrections are focused on when and why they are (or are not) effective. Research that investigates only a small number of topics at a time will struggle to move these debates forward as effects from different studies are typically incomparable due to many differing design features (e.g., sample, time, question wording). Thus, topic sampling is an ideal tool for moving these debates forward. Nevertheless, there are some cases in which topic sampling may not be appropriate. Some scholars may develop theories that are specific to a particular topic with no intent to generalize. However, most scholars aim to generalize across some set of contexts, topics, or stimuli, and thus topic sampling should be a useful tool in many cases.

Our experiment involved an application to a specific survey experimental design, but creative researchers can also pair topic sampling with other forms of survey experiments, field experiments, and even observational designs. Within surveys, topic sampling could be combined with pre-post or within-subjects designs (Clifford, Sheagley, and Piston 2021). A field experimenter manipulating policy threat (e.g., Miller and Krosnick 2004) might randomly sample the focal issue. Many field experiments on racial bias randomly assign a name to signal racial identity and topic sampling may be a useful method for investigating the extent and correlates of treatment effect variation across names (Butler and Homola 2017). Topic sampling could also be useful when researchers are forced to rely on observational designs. For example, researchers studying moral conviction often use designs that make within-subjects comparisons across a handful of issues (Ryan, 2014, 2017). Rather than handpicking a small number of issues, these researchers could randomly sample a large number of issues 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 & Kern, 2012; Kam and Trussler 2016).[10] 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.

As is surely clear by now, topic sampling requires high-quality descriptive data. In our case, we had to develop a population of policies relevant to public opinion. That population is a moving target, however, and will need to be updated regularly. In some cases, the population may be relatively stable and easy to define, such as democratic countries. In others, the population may be much more challenging to identify, such as the case of political misperceptions. Yet, this is not a problem with topic sampling as a method, but a problem with a lack of descriptive knowledge about the topic of study.

Researchers will also have to make important choices about sampling, such as whether to weight topics by their salience or importance. For example, we opted for a stratified random sample to ensure the diversity of topics, but assigned equal probabilities to topics, regardless of their salience. Given the goal of examining and explaining variation across topics, this sampling strategy is optimal because it maximizes variation. However, if a researcher is more interested in the typical effect than in explaining variation, then it may make more sense to weight topics by their frequency or salience. Choices about sampling strategy only emphasize the need for high-quality descriptive information about the population, including characteristics of each topic (e.g., social vs. economic). We encourage researchers to develop and update these databases and to share them publicly (as we intend to do soon). Clearly, this 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 do so.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11109-023-09870-1.

Acknowledgements The authors would like to thank Brandon de la Cuesta and Brendan Nyhan for helpful comments.

Declarations

Conflict of interest This study was funded by the Danish Council for Independent Research award DFF–4003-00192B. The authors declare they have no financial interests.

[10] This possibility of investigating many moderators underscores the importance of pre-registration. We also note that the topic sampling should be described in the pre-registration document just as a researcher would describe the sampling of participants.

References

Arceneaux, K. (2007). Can partisan cues diminish democratic accountability? *Political Behavior*, *30*(2), 139–160.

Bakker, B. N., & Lelkes, Y., and Ariel Malka (2020). Understanding partisan cue receptivity: Tests of predictions from the Bounded rationality and expressive utility perspectives. *Journal of Politics*, *82*(3), 1061–1077.

Bansak, K., Hainmueller, J., Hopkins, D. J., & Teppei Yamamoto. (2021). "Conjoint Survey Experiments." Ch. 2 in Advances in Experimental Political Science, eds. Druckman, J. N., & Green, D. P. (p. 19). Cambridge: Cambridge University Press.

Barabas, J., and Jennifer Jerit (2010). Are Survey experiments externally valid? *American Political Science Review*, *104*(02), 226–242.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using Lme4. *Journal of Statistical Software, 67*(1), 1–48.

Berinsky, A. J., Gregory, A., Huber, & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.Com's mechanical Turk. *Political Analysis, 20*(3), 351–368.

Brandt, M. J., Jarret, T., & Crawford (2019). Studying a heterogeneous array of Target Groups can help us understand prejudice. *Current Directions in Psychological Science*, *28*(3), 292–298.

Brutger, R. (2022). "Abstraction and Detail in Experimental Design." *American Journal of Political Science,*

Bullock, W., Imai, K., & Shapiro, J. (2011). Statistical Analysis of Endorsement Experiments: Measuring Support for Militant Groups in Pakistan. *Political Analysis, 19*(4), 363–384.

Bullock, J. G. (2011). "Elite Influence on Public Opinion in an Informed Electorate." *American Political Science Review, 105*(03), 496–515.

Bürkner, P. C. (2017). Brms: An R Package for bayesian Multilevel Models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Butler, D. M., and Jonathan Homola (2017). An empirical justification for the Use of racially distinctive names to signal race in experiments. *Political Analysis*, *25*(1), 122–130.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). "Stan: A Probabilistic Programming Language."*Journal of Statistical Software, 76*(1): 1–32.

Cameron, A., Colin, J. B., Gelbach, & Miller, D. L. (2011). Robust inference with Multiway Clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249.

Chong, D., and James N. Druckman (2007). Framing Public Opinion in competitive democracies. *American Political Science Review*, *101*(04), 637–655.

Chong, D., and James N. Druckman (2010). Dynamic public opinion: Communication Effects over Time. *American Political Science Review*, *104*(04), 663–680.

Chong, D., and James N. Druckman (2012). Counterframing Effects. *The Journal of Politics*, *75*(01), 1–16.

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

Clifford, S., Sheagley, G., & Piston, S. (2021). Increasing Precision without Altering Treatment Effects: Repeated measures designs in Survey experiments. *American Political Science Review*, *115*(3), 1048–1065.

Coppock, A., & Green, D. P. (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), 113–131.

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). "Generalizability of Heterogeneous Treatment Effect Estimates across Samples." *Proceedings of the National Academy of Sciences*: 201808083.

Ditto, P. 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), 273–291.

Druckman, J. N., Thomas, J., & Leeper (2012). Learning more from political communication experiments: Pretreatment and its Effects. *American Journal of Political Science*, *56*(4), 875–896.

Franco, A., & Malhotra, N., and Gabor Simonovits (2014). Publication Bias in the Social Sciences: Unlocking the file drawer. *Science (New York N Y)*, *345*(6203), 1502–1505.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1. https://mc-stan.org/rstanarm

Gelman, A., and Jennifer Hill (2007). *Data Analysis using regression and Multilevel/Hierarchical models*. New York: Cambridge University Press.

Gelman, A., Jennifer Hill, and Masanao Yajima (2012). Why we (usually) don't have to worry about multiple comparisons. *Journalof Research on Educational Effectiveness*, *5*(2), 189–211.

Gilens, M. (2001). Political ignorance and collective policy preferences. *American Political Science Review*, *95*(2), 379–396.

Green, D. P., & Kern, H. L. (2012). Modeling Heterogeneous Treatment Effects in Survey experiments with bayesian additive regression trees. *Public Opinion Quarterly*, *76*(3), 491–511.

Grieco, J. M., Gelpi, C., & Reifler, J., and Peter D. Feaver (2011). Let's get a second opinion: International Institutions and American Public support for War1. *International Studies Quarterly*, *55*(2), 563–583.

Guess, A., & Coppock, A. (2020). Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments. *British Journal of Political Science, 50*(4), 1497–1515.

Guisinger, A., and Elizabeth N. Saunders (2017). Mapping the Boundaries of Elite Cues: How elites shape Mass Opinion across International Issues. *International Studies Quarterly*, *61*(2), 425–441.

Hainmueller, J., & Hopkins, D. J., and Teppei Yamamoto (2014). Causal inference in Conjoint Analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, *22*(1), 1–30.

Jerit, J. (2009). How predictive appeals affect policy opinions. *American Journal of Political Science*, *53*(2), 411–426.

Jerit, J. (2012). Partisan perceptual bias and the information environment. *The Journal of Politics, 74*(03), 672–684.

Jerit, J., & Barabas, J. (2013). Comparing contemporaneous laboratory and field experiments on media effects. *Public Opinion Quarterly, 77*(1), 256–282.

Johnston, C. D., Lavine, H., & Federico, C. M. (2017). *Open versus closed: Personality, identity, and the politics of redistribution*. Cambridge: Cambridge University Press.

Johnston, C. D. (2015). Personality dispositions and political preferences across hard and easy issues. *Political Psychology, 36*(1), 35–53.

Kam, C. D., & Trussler, M. J. (2016). At the nexus of observational and experimental research: Theory, specification, and analysis of experiments with heterogeneous treatment effects. *Political Behavior, 39*, 1–27.

Kam, C. D. (2005). Who Toes the Party Line? Cues, Values, and Individual Differences. *Political Behavior, 27*, 163–182.

Kay, M. (2019). "Tidybayes: Tidy Data and Geoms for Bayesian Models." https://doi.org/10.5281/zenodo.1308151

Kuklinski, J. H., et al. (2000). Misinformation and the currency of democratic citizenship. *The Journal of Politics*, *62*(3), 790–816.

Lenz, G. S. (2009). Learning and opinion change, not priming: Reconsidering the priming hypothesis. *American Journal of Political Science, 53*(4), 821–837.

Lindner, N. M., & Nosek, B. A. (2009). Alienable Speech: Ideological variations in the application of Free-Speech Principles. *Political Psychology*, *30*(1), 67–92.

McDermott, R. (2002). Experimental methodology in Political Science. *Political Analysis*, *10*(4), 325–342.

Miller, J. M., & Krosnick, J. A. (2004). Threat as a motivator of political activism: A field experiment. *Political Psychology*, *25*(4), 507–523.

Nicholson, S. P. (2012). "Polarizing Cues." *American Journal of Political Science, 56*(1), 52–66.

Nyhan, B., and Jason Reifler (2010). When Corrections fail: The persistence of political Misperceptions. *Political Behavior*, *32*(2), 303–330.

Ryan, T. J. (2014). "Reconsidering Moral Issues in Politics." *The Journal of Politics, 76*(2), 1–18.

Ryan, T. J. (2017). No compromise: Political consequences of moralized attitudes. *American Journal of Political Science, 61*(2), 409–423.

Simas, E. N., & Milita, K., and John Barry Ryan (2021). Ambiguous rhetoric and legislative accountability. *Journal of Politics*, *83*(4), 1695–1705.

Slothuus, R. (2016). Assessing the influence of political parties on public opinion: The challenge from pretreatment effects. *Political Communication, 33*(2), 302–327.

Swire-Thompson, Briony, J., DeGutis, & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition, 9*(3), 286–299.

Tappin, B. M. (2022). "Estimating the Between-Issue Variability in Party Elite Cue Effects." *Public Opinion Quarterly, 86*(4), 862–885

Tavits, M. (2007). Principle vs. pragmatism: Policy shifts and political competition. *American Journal of Political Science*, *51*(1), 151–165.

Tesler, M. (2015). Priming Predispositions and changing policy positions: An account of when Mass Opinion is primed or changed. *American Journal of Political Science*, *59*(4), 806–824.

Vehtari, A., & Gelman, A. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*, 1413–1432.

Wells, G. L. (2016). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin, 25*(9), 1115–1125.

Wood, T., & Porter, E. (2019). "The elusive backfire effect: mass attitudes' steadfast factual adherence." *Political Behavior, 41*, 135–163.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.