Political Analysis (2024), **32**, 431–444
doi:10.1017/pan.2024.1

PA

ARTICLE

Estimators for Topic-Sampling Designs

Scott Clifford[1] and Carlisle Rainey[2]

[1]Associate Professor, Texas A&M University, College Station, TX, USA; [2]Associate Professor, Florida State University, Tallahassee, FL, USA

Corresponding author: Carlisle Rainey; Email: crainey@fsu.edu

(Received 10 August 2023; revised 21 December 2023; accepted 20 January 2024; published online 13 May 2024)

Abstract

When researchers design an experiment, they usually hold potentially relevant features of the experiment constant. We call these details the "topic" of the experiment. For example, researchers studying the impact of party cues on attitudes must inform respondents of the parties' positions on *a particular policy*. In doing so, researchers implement just *one* of many possible designs . Clifford, Leeper, and Rainey (2023. "Generalizing Survey Experiments Using Topic Sampling: An Application to Party Cues." Forthcoming in Political Behavior. https://doi.org/10.1007/s11109-023-09870-1) argue that researchers should implement *many* of the possible designs in parallel—what they call "topic sampling"—to generalize to a larger population of topics. We describe two estimators for topic-sampling designs: First, we describe a nonparametric estimator of the typical effect that is unbiased under the assumptions of the design; and second, we describe a hierarchical model that researchers can use to describe the heterogeneity. We suggest describing the heterogeneity across topics in three ways: (1) the standard deviation in treatment effects across topics, (2) the treatment effects for particular topics, and (3) how the treatment effects for particular topics vary with topic-level predictors. We evaluate the performance of the hierarchical model using the Strengthening Democracy Challenge megastudy and show that the hierarchical model works well.

Keywords: Bayesian inference; experimental design; heterogeneous treatment effects; hierarchical models; topic sampling

Edited by: Jeff Gill

1. Introduction

When researchers design an experiment, they usually hold important details of the experiment constant. Equivalently, they choose among many different (but similar) experiments that test the same substantive claim. In practice, researchers typically select and run a *single* experiment (or perhaps a *handful* of experiments) from the collection of possibilities. Clifford, Leeper, and Rainey (2023) refer to this collection of possibilities as "topics," and Brutger *et al.* (2023) offer an excellent discussion of how topics might vary across experiments. 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), even though the general claim applies to hypothetical interventions in many

We have written a companion paper (Clifford, Leeper, and Rainey 2023) to this article that focuses on the conceptual and substantive motivation for topic sampling and explores an example application in detail. The companion paper is available at https://doi.org/10.1007/s11109-023-09870-1. We thank Charles Crabtree, Brandon de la Cuesta, Ryan Kennedy, Brendan Nyhan, Diego Reinero, Geoff Sheagley, Ben Tappin, Emily Thorson, Arjun Vishwanath, and Yamil Velez for their helpful comments. We were helped greatly by audiences at several conferences, informal conversations with many colleagues, and an excellent pool of peer reviewers. All data and code to reproduce our results are available on Dataverse at https://doi.org/10.7910/DVN/YBV9Z8 (Rainey 2024).

different countries. Or researchers might investigate ideological asymmetries in political tolerance by randomizing between two specific social groups (e.g., Arabs vs. Americans; Lindner and Nosek 2009), even though the claim applies to other groups as well. The term "topic" suggests *substantive* variation in the details of the experiment, including variation in the treatment, variation in the control, and/or variation in the experimental context. However, our ideas and methods generalize to ancillary variation as well (e.g., the level of detail in a vignette experiment).

This is a practical matter of experimental design. Researchers studying the effects of correcting misperceptions on affective polarization must correct some misperceptions, but which misperceptions should they correct? Researchers comparing a treatment to a placebo must use some placebo, but which placebos should they use?[1] Researchers using names to cue race and gender must use particular names, but which names should they use? We suspect that researchers are most interested in the substantive variation that these examples illustrate—thus our use of the term "topic." However, researchers must also select ancillary details and may care about variation across these sets of details. If a photograph accompanies a mock news article, which photograph? If a survey experiment uses a vignette, how detailed?

We propose that researchers are usually not only interested in the treatment effect for the chosen topic, but in a more general treatment effect averaged over the topics in a population of interest. But when researchers study a single topic, they must either (1) assume that their findings generalize to the larger collection or (2) admit that their findings may have limited generalizability across topics (see Clifford and Rainey 2023). Experimentalists are certainly *aware* that studies of particular topics might not generalize to other conceptually similar topics. Like many others, Chong and Druckman (2013, 14) engage thoughtfully with this limitation:

> Our results are potentially circumscribed by our focus on a single issue and a single approach to operationalizing attitude strength. However, we believe our theory should apply to any issue, including hotly debated issues on which most people hold strong prior opinions; attempts to frame public opinion on such issues will be more difficult or may fail outright.

Clifford *et al.* (2023) propose that researchers use topic sampling to generalize empirically, rather than simply speculate about generalizability (see also Porter and Velez 2022; Tappin 2023).[2] Rather than assigning many respondents to a particular topic, they suggest assigning a slightly larger number of total respondents to many different topics (i.e., about 20% to 50% more respondents and 25 to 50 total topics) and then aggregating those many separate treatment effects into (1) a typical treatment effect across topics and (2) a summary of the heterogeneity across topics. Several recent papers *suggest* a hierarchical model and illustrate its use for this basic design, but do not motivate their estimator in detail (e.g., Clifford *et al.* 2023; Tappin 2023; Tappin, Berinsky, and Rand 2023; Wittenberg *et al.* 2021). In this paper, we thoroughly describe and defend the estimators.

We describe two estimators for topic-sampling designs. To fix ideas, we imagine a collection of experiments indexed by j. For any particular topic j, researchers can use a randomized experiment to estimate the average treatment effect $\delta_j = \overline{Y}_j^T - \overline{Y}_j^C$ for each particular topic, where $\overline{Y}_j^T$ and $\overline{Y}_j^C$ represent the average potential outcomes across respondents for topic j under treatment and control, respectively.[3] While researchers would like to generalize beyond the study of a single topic, the experiment requires fixing details, so the treatment effect is only defined for a *particular* topic j. Much like the average marginal component effect (AMCE) estimand from a conjoint design marginalizes over attributes

[1] See Porter and Velez (2022) for a detailed discussion of this exact question.

[2] Outside political science, see Wells and Windschitl (1999) on "stimulus sampling," Baribault *et al.* (2018) on "radical randomization," Yarkoni (2022) on "the generalizability crisis," and Brandt and Wagemans (2017).

[3] δ_j is an *average* treatment effect (across respondents). For clarity though, we suppress the "average" and refer to δ_j as a "treatment effect" throughout this paper. We omit the "average" in this instance to avoid confusing an average *across respondents* with an average *across topics*.

(Hainmueller, Hopkins, and Yamamoto 2014), we suggest marginalizing across topics. We define a *typical treatment effect* $\bar{\delta}$ as the average of treatment effects for particular topics δ_j (i.e., "typical" across topics). If we conceive of topics as a population of size $\mathcal{J}$ (e.g., that we can randomly sample from), then $\bar{\delta} = \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} \delta_j = \text{avg}(\delta)$. Alternatively, we might consider the δ_j as exchangeable random variables and define $\bar{\delta} = E(\delta)$.

In addition to the typical treatment effect $\bar{\delta}$, researchers might want to describe the *variation* in the treatment effects across topics. At a minimum, we suggest estimating (1) the standard deviation (SD) of the treatment effects across topics and (2) the treatment effects for particular topics. However, the description of the variation across topics becomes especially rich when researchers use topic-level predictors to characterize the variation in treatment effects across topics.

Below, we develop two estimators for topic-sampling designs. First, we use the assumption of random sampling of topics to develop a nonparametric estimator for the typical treatment effect. Second, we use the assumption of exchangeability to develop a parametric, hierarchical model to characterize the heterogeneity across topics: the SD around the (conditional) typical effect, estimates of the treatment effects for particular topics, and the typical effect *conditional on topic-level predictors*. We use the Strengthening Democracy Challenge (SDC) megastudy (Voelkel *et al.* 2023) to demonstrate the effectiveness of the hierarchical model in a real data set.

2. A Nonparametric Estimator of the Typical Effect

First, we sketch a nonparametric estimator from the design-based assumptions of (1) a random sample of topics from a population of interest and (2) random assignment to treatment and control within each of these topics. To motivate these estimators, we assume that researchers recruit N participants. Then, researchers randomly assign each participant to $M \in \{1,2,\ldots,J\}$ topics. For each of these M topics, researchers randomly assign each respondent to treatment and control. This assignment can be independent across topics for each respondent (e.g., Tappin 2023) or each respondent might be randomly assigned to treatment for V topics, where $1 \leq V \leq M-1$, and to control for the other $M-V$ topics (e.g., Clifford *et al.* (2023) use $V = 1$, assigning respondents to treatment for one topic and to control for five topics).

2.1. Point Estimates

Researchers can estimate the treatment effect for the particular topic j as the difference in means between the treatment and control groups for topic j, so that $\widehat{\delta_j} = \bar{y}_j^T - \bar{y}_j^C$, where $\bar{y}_j^T$ and $\bar{y}_j^C$ represent the sample averages of the treatment and control groups for topic j. Then, researchers can estimate the typical treatment effect using the average of the estimates for particular topics, so that $\widehat{\bar{\delta}} = \frac{1}{J} \sum_{j=1}^{J} \widehat{\delta_j} = \text{avg}\left(\widehat{\delta}\right)$ for $\widehat{\delta} = \left\{\widehat{\delta_1}, \ldots, \widehat{\delta_J}\right\}$.

2.2. Variance Estimates and Confidence Intervals (CIs)

We focus on a scenario in which the population of topics is large, so researchers take a simple random sample of J topics to use in their experiment. For example, Clifford *et al.* (2023) take a random sample of 48 policies from a population of 154. In this situation, there is uncertainty due to (1) random assignment of respondents to topics and treatment/control *and* (2) random sampling of topics. In this case, we have $\text{Var}\left(\widehat{\bar{\delta}}\right) = \text{Var}\left(\frac{1}{J} \sum_{j=1}^{J} \widehat{\delta_j}\right) = \frac{1}{J^2} \text{Var}\left(\sum_{j=1}^{J} \widehat{\delta_j}\right) \leq \frac{1}{J^2} \sum_{j=1}^{J} \text{Var}\left(\widehat{\delta_j}\right)$. For example, the estimate of the treatment effect for the first topic $\widehat{\delta_1}$ varies because (1) the first topic is randomly sampled, (2) respondents are randomly assigned to this first topic, and (3) respondents are randomly assigned to treatment and control within this first topic. Fortunately, the experiment offers J replicates of $\widehat{\delta_j}$, so we can plug in the sample variance of $\widehat{\delta} = \left\{\widehat{\delta_1}, \ldots, \widehat{\delta_J}\right\}$, which is $\text{var}\left(\widehat{\delta}\right) = \frac{1}{J-1} \sum_{j=1}^{J} \left[\left(\widehat{\delta_j} - \widehat{\bar{\delta}}\right)^2\right]$. This produces

$\widehat{\text{Var}}\left(\widehat{\bar{\delta}}\right) = \frac{1}{J^2}\sum_{j=1}^{J}\overbrace{\left[\frac{1}{J-1}\sum_{j=1}^{J}\left(\widehat{\delta}_j - \widehat{\bar{\delta}}\right)^2\right]}^{\text{sample variance of the point estimates for particular topics}} = \frac{1}{J}\text{var}\left(\widehat{\delta}\right)$. Researchers can treat $\widehat{\text{Var}}\left(\widehat{\bar{\delta}}\right)$ as following a chi-squared distribution and create $(1-\alpha)\times 100\%$ CIs of the form $\widehat{\bar{\delta}} \pm t_{\frac{\alpha}{2},df}\sqrt{\widehat{\text{Var}}\left(\widehat{\bar{\delta}}\right)}$, where $df = J-1$ (e.g., $\widehat{\bar{\delta}} \pm 1.71\sqrt{\widehat{\text{Var}}\left(\widehat{\bar{\delta}}\right)}$ for a 90% CI when $J = 25$).

2.2.1. Assigning Each Respondent to Multiple Topics

In some applications, researchers may want to assign the same respondent to multiple topics. If this is a sensible option for researchers' application, we strongly suggest it. For example, Clifford *et al.* (2023) assign respondents to six policies at random; the first five are always control, and the sixth is always treatment. Tappin (2023) randomly assigns respondents to six of 34 policies and assigns respondents *independently* to treatment and control across policies. Assigning respondents to multiple topics allows researchers to drastically increase their statistical power for a similar cost. Whether researchers assign respondents to one, several, or all topics does not affect the validity of the point estimators or CIs above.

However, assigning respondents to multiple conditions can potentially introduce order effects (for discussion, see Mutz 2011). Whether order effects are a concern will depend on the application, but these types of concerns seem overstated (e.g., Clifford, Sheagley, and Piston 2021; Mummolo and Peterson 2019). For example, Tappin (2023) makes the case that these order effects are minimal in his experiment on party cues. Similarly, researchers can use a well-powered pilot test on a narrow set of topics to assess the magnitude of order effects. If researchers are worried about spillover effects, they might follow the practice of Clifford *et al.* (2023) and assign respondents to control conditions for multiple topics and a treatment condition for a single topic after the control conditions. If researchers are not worried about spillover, then assigning respondents to several topics and to the same number of treatment and control conditions produces more precise estimates.

3. Extensions: When the Topic Is a Treatment (or a Control, or Both)

In the discussion above, we imagine that potential outcomes for treatment and control exist within each topic, so that $\delta_j = \overline{Y}_j^T - \overline{Y}_j^C$ is defined *for each topic*. However, some applications might have topics that represent *either* treatment conditions *or* control conditions. In this case, the treatment effect is not defined within a topic, so our framework requires adjustment. We consider three extensions.

1. **Extension 1a:** *The topics are distinct treatments.* In this scenario, researchers compare many treatment groups to a single control group. For example, Voelkel *et al.* (2023) consider 25 interventions that might reduce affective polarization. We might think of these many treatments as "topics" in our framework. However, the control group in their study receives no such treatment, and thus, the topic does not vary in the control group.
2. **Extension 1b:** *The topics are distinct controls.* In this scenario, researchers compare a single treatment group to many control groups. For example, Porter and Velez (2022) use a large collection of placebos generated via GPT-2 as a control group. We might think of each placebo as a "topic." However, these topics only vary in the control group; the treatment group receives a single treatment, and thus, the topic does not vary in the treatment group.
3. **Extension 2:** *The topics are either treatments or controls.* In this scenario, researchers compare many treatment groups to many control groups. For example, researchers use collections of names associated with racial groups to cue race (e.g., Bertrand and Mullainathan 2004; Crabtree *et al.* 2022; Elder and Hayes 2023). In this case, we might think of the particular name

(e.g., Octavia, Misty) as the "topic" and think of the racial category of the name (e.g., Black, white) as the treatment and control conditions (see Elder and Hayes 2023). In this case, the topic (i.e., the name) varies within both the treatment and control groups, but the treatment and control groups contain distinct topics.

The intuition from the baseline design generalizes easily to the extensions, but the formalization differs somewhat. For the baseline design, each topic has its own treatment and control groups. In that case, it is natural to think about the "typical" treatment effect $\overline{\delta} = \text{avg}(\delta)$. But when topics do not include both a treatment group and a control group, this motivation no longer works. Instead, we must think about the typical effect as the difference between (1) the average (or "typical") average potential outcome among treatment topics and (2) the average (or "typical") average potential outcome among control topics. In this case, $\overline{\delta} = \overbrace{\left[\frac{1}{\mathcal{J}_T}\sum_{j=1}^{\mathcal{J}_T}\overline{Y}_j^T\right]}^{\substack{\text{typical outcome among}\\ \text{treatment topics}}} - \overbrace{\left[\frac{1}{\mathcal{J}_C}\sum_{j=1}^{\mathcal{J}_C}\overline{Y}_j^C\right]}^{\substack{\text{typical outcome among}\\ \text{control topics}}}$, where $\mathcal{J}_T$ and $\mathcal{J}_C$ represent the number of topics in the treatment and control groups and $\overline{Y}_j^T$ and $\overline{Y}_j^C$ represent the average potential outcomes in the (distinct) jth topics for the treatment and control groups, respectively. When the control group has "no topic," then we can set $\mathcal{J}_C = 1$ so that $\frac{1}{\mathcal{J}_C}\sum_{j=1}^{\mathcal{J}_C}\overline{Y}_j^C$ simplifies to $\overline{Y}^C$. When the treatment group has no topic, $\left[\frac{1}{\mathcal{J}_T}\sum_{j=1}^{\mathcal{J}_T}\overline{Y}_j^T\right]$ simplifies similarly to $\overline{Y}^T$. To obtain point estimates, we can plug the sample means into $\left[\frac{1}{\mathcal{J}_T}\sum_{j=1}^{\mathcal{J}_T}\overline{Y}_j^T\right]$ and $\overline{Y}^C$ to obtain $\left[\frac{1}{J_T}\sum_{j=1}^{J_T}\overline{y}_j^T\right]$ and $\overline{y}^C$, for example.

Randomization and inference also work slightly differently under the extensions. In the baseline design, researchers assign respondents randomly to topics and then to treatment and control within those topics. Thus, the assignment to treatment and topic are independent of the baseline design. For the extensions, there is no treatment and control within topics, and there are only topics that are considered as "treatments" or "controls." To maintain the intuition in the extensions, we assume that respondents are first randomly assigned to treatment or control and then randomly assigned a single topic.

Consider the variance estimates for Extension 1a first. We can estimate $\text{Var}\left[\frac{1}{J_T}\sum_{j=1}^{J_T}\overline{y}_j^T\right]$ using $\frac{1}{J^T}\text{var}\left(\overline{y}^T\right)$, where $\overline{y}^T = \left\{\overline{y}_1^T, \ldots, \overline{y}_{J^T}^T\right\}$ and J^T represents the number of topics in the treatment group. We can then estimate $\text{Var}\left(\overline{y}^C\right)$ using $\frac{\text{var}(y^C)}{N^C}$, where N_C is the number of respondents in the single control group, which is the usual variance estimator for randomized experiments. Finally, we can estimate the variance of the point estimate $\widehat{\overline{\delta}} = \left[\frac{1}{J_T}\sum_{j=1}^{J_T}\overline{y}_j^T\right] - \overline{y}^C$ using $\widehat{\text{Var}}\left(\widehat{\overline{\delta}}\right) = \frac{1}{J^T}\text{var}\left(\overline{y}^T\right) + \frac{\text{var}(y^C)}{N^C}$.

The estimates are analogous for $\left[\frac{1}{\mathcal{J}_C}\sum_{j=1}^{\mathcal{J}_C}\overline{Y}^C\right]$ and $\overline{Y}^T$. Table 1 shows the point and variance estimates for the extensions as well as a single-topic design for comparison.

Similar to the baseline design, researchers can treat $\widehat{\text{Var}}\left(\widehat{\overline{\delta}}\right)$ as following a chi-squared distribution and create $(1-\alpha)\times 100\%$ CIs of the form $\widehat{\overline{\delta}} \pm t_{\frac{\alpha}{2},df}\sqrt{\widehat{\text{Var}}\left(\widehat{\overline{\delta}}\right)}$. However, the degrees of freedom calculation can be improved slightly. Following the Welch–Satterthwaite approach, the generic expression for the degrees of freedom adjustment in the extensions is $df = \frac{\left(\widehat{V}_1+\widehat{V}_2\right)^2}{\frac{\widehat{V}_1^2}{df_1}+\frac{\widehat{V}_2^2}{df_2}}$. For Extension 1a, we have $\widehat{V}_1 = \frac{1}{J}\text{var}\left(\overline{y}^T\right)$, $df_1 = J-1$, $\widehat{V}_2 = \frac{\text{var}(y^C)}{N^C}$, and $df_2 = N_c - 1$. For Extension 1b, we have $\widehat{V}_1 = \frac{\text{var}(y^T)}{N^T}$, $df_1 = N_T - 1$, $\widehat{V}_2 = \frac{1}{J}\text{var}\left(\overline{y}^C\right)$, and $df_2 = J-1$. For Extension 2, we have $\widehat{V}_1 = \frac{1}{J_T}\text{var}\left(\overline{y}^T\right)$, $df_1 = J_T - 1$, $\frac{1}{J_C}\text{var}\left(\overline{y}^C\right)$, and $df_2 = J_C - 1$.

The nonparametric estimators above are relatively straightforward and depend only on design-based assumptions. However, the nonparametric approach above is limited—it only estimates the typical treatment effect. To summarize the heterogeneity across topics, we suggest a hierarchical model.

Table 1. This table shows the point estimates for the typical effect and the associated variance estimates for the baseline design, the three extensions, and the single-topic design.

Description	Point estimate	Variance estimate
Baseline: Each topic contains a treatment group and a control group	$\widehat{\delta}_j = \bar{y}_j^T - \bar{y}_j^C$ $\widehat{\bar{\delta}} = \text{avg}\left(\widehat{\delta}\right)$	$\widehat{\text{Var}}\left(\widehat{\bar{\delta}}\right) = \frac{1}{J}\,\text{var}\left(\widehat{\delta}\right)$
Extension 1a: Each topic is a distinct treatment condition. This extension compares many treatment conditions to a single control condition	$\widehat{\bar{\delta}} = \text{avg}\left(\bar{y}^T\right) - \bar{y}^C$	$\widehat{\text{Var}}\left(\widehat{\bar{\delta}}\right) = \frac{1}{J}\,\text{var}\left(\bar{y}^T\right) + \frac{\text{var}\left(y^C\right)}{N^C}$
Extension 1b: Each topic is a distinct control condition. This extension compares a single treatment condition to many control conditions	$\widehat{\bar{\delta}} = \bar{y}^T - \text{avg}\left(\bar{y}^C\right)$	$\widehat{\text{Var}}\left(\widehat{\bar{\delta}}\right) = \frac{\text{var}\left(y^T\right)}{N^T} + \frac{1}{J}\,\text{var}\left(\bar{y}^C\right)$
Extension 2: Each topic is either a treatment condition or a control condition. This extension compares many treatment conditions to many control conditions	$\widehat{\bar{\delta}} = \text{avg}\left(\bar{y}^T\right) - \text{avg}\left(\bar{y}^C\right)$	$\widehat{\text{Var}}\left(\widehat{\bar{\delta}}\right) = \frac{1}{J^T}\,\text{var}\left(\bar{y}^T\right) + \frac{1}{J^C}\,\text{var}\left(\bar{y}^C\right)$
Single-Topic Design: Topic does not vary. This is a single treatment condition compared to a single control condition	$\widehat{\delta} = \bar{y}^T - \bar{y}^C$	$\widehat{\text{Var}}\left(\widehat{\delta}\right) = \frac{\text{var}\left(y^T\right)}{N^T} + \frac{\text{var}\left(y^C\right)}{N^C}$

4. A Hierarchical Model of the Heterogeneity

A premise of topic sampling is that treatment effects vary across topics; indeed, that is the motivation for including multiple topics in the study. This variation implies three or four quantities of interest: (1) the typical treatment effect across topics, (2) the amount of variation across topics (e.g., the SD), (3) the treatment effects for particular topics, and (4) descriptions of how the treatment effects vary with topic-level predictors. The nonparametric approach above works well for the typical treatment effect. However, the hierarchical model adds just a little structure, allowing researchers to estimate a much richer set of quantities of interest.

To motivate the hierarchical model, we assume that the treatment effects for particular topics δ_j are exchangeable.[4] Formally, this means that the treatment effects $\delta_1, \delta_2, \ldots, \delta_J$ have the same joint distribution as *any* permutation $\delta_{[1]}, \delta_{[2]}, \ldots, \delta_{[J]}$. A simple random sample of topics guarantees that exchangeability holds in the sense that the *j*th topic is selected at random. However, exchangeability has a subtler and useful interpretation. We suggest thinking of exchangeability as a structural prior (Gelman, Simpson, and Betancourt 2017). That is, we suggest thinking of the effects as "different but similar" or as draws from a distribution with a parameter that captures the similarity. Substantively, this means the label for the topic contains no information about its treatment effect; Gelman *et al.* (2004) write that "ignorance implies exchangeability" (121) and that "with no information available to distinguish [the parameters], we have no logical choice but to model [the parameters] exchangeably" (124).

Of course, researchers are not always ignorant of how the treatment effects vary across topics. This does not pose a major problem for the assumption of exchangeability, though. If researchers have knowledge of how the treatment effect δ_j varies across topics, then researchers can use topic-level predictors to explicitly model the topic-level variation and obtain exchangeability (conditional on predictors). In other words, they can model the expected variation across topics and treat the unexpected variation as exchangeable. For example, Clifford *et al.* (2023) model the effect of partisan cues using the public's awareness of parties' positions on the policies. Adding topic-level predictors changes the motivation and computation only a little, but it meaningfully changes the interpretation of the model parameters, so we discuss each separately. We first discuss models *without* topic-level predictors.

[4] Bernardo (1996) offers a brief and careful discussion of the concept of exchangeability. Gelman *et al.* (2004, pp. 121-124) offer a brief, intuitive, and substantively motivated discussion of exchangeability. Gelman (2005) and Feller and Gelman (2015) offer thorough, accessible discussions of the assumption of exchangeability in the context of randomized experiments.

4.1. Models without Topic-Level Predictors of the Treatment Effect

We begin in the scenario where each respondent is assigned to a single topic and then randomly assigned to treatment or control. We imagine a hierarchical normal–linear model $y_i = \alpha_{j[i]} + \delta_{j[i]} T_i + \epsilon_i$, where $\epsilon_i \sim N\left(0, \sigma_\epsilon^2\right)$. In short, we imagine separate regressions for each topic (but assume a constant variance across topics). Critically, we model the intercept and slope as batches of different-but-similar parameters, so that

$$\begin{pmatrix} \alpha_j \\ \delta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \overline{\alpha} \\ \overline{\delta} \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\delta \\ \rho\sigma_\alpha\sigma_\delta & \sigma_\delta^2 \end{pmatrix}\right).$$

Using R's popular mixed-effects syntax (Bates *et al.* 2015), we have the model `y ~ treatment + (1 + treatment | topic)`. The multivariate normal distribution implies that $\delta_j \sim N\left(\overline{\delta}, \sigma_\delta^2\right)$. This allows a much richer set of quantities of interest, without an overly restrictive parametric model. First, we obtain an estimate of $\overline{\delta}$, which is the "typical" treatment effect across topics. Second, we obtain an estimate of σ_δ, which is the SD of the treatment effects across topics. Third, we obtain an estimate of the treatment effect δ_j for each topic included in the study.

The model *without* topic-level predictors gives us a summary of the following form: "The treatment effects are about $[\overline{\delta}]$ give or take $[\sigma_\delta]$ or so" along with estimates for particular topics. This is conceptually equivalent to summarizing a small data set with an average and an SD.[5] But researchers should report the estimates of the treatment effects for particular topics to allow readers to assess the variation across topics in detail. Tappin (2023) provides an example of this approach (see Tappin's Figure 2 on p. 876).

4.2. Models with Topic-Level Predictors of the Treatment Effect

We now consider a model *including* topic-level predictors. Researchers should think carefully about topic-level predictors to measure and use in their model. These topic-level predictors improve their ability to (1) describe the variation in the effects and (2) precisely estimate the treatment effects for particular topics.[6]

To include topic-level predictors, we adopt the same setup as before, imagining the set of models $y_i = \alpha_{j[i]} + \delta_{j[i]} T_i + \epsilon_i$, where $\epsilon_i \sim N\left(0, \sigma_\epsilon^2\right)$. We make one small change to the model for α_j and δ_j. We add subscripts j to their means $\overline{\alpha}_j$ and $\overline{\delta}_j$ to signify that they vary systematically across topics, so that

$$\begin{pmatrix} \alpha_j \\ \delta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \overline{\alpha}_j \\ \overline{\delta}_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\delta \\ \rho\sigma_\alpha\sigma_\delta & \sigma_\delta^2 \end{pmatrix}\right).$$

Then, we model the means using the covariates in the usual linear way, so that

$$\overline{\alpha}_j = \gamma_0^{[\overline{\alpha}]} + \gamma_1^{[\overline{\alpha}]} z_{j1} + \gamma_2^{[\overline{\alpha}]} z_{j2} + \cdots + + \gamma_k^{[\overline{\alpha}]} z_{jk} = Z\gamma^{[\overline{\alpha}]}$$

and

$$\overline{\delta}_j = \gamma_0^{\left[\overline{\delta}\right]} + \gamma_1^{\left[\overline{\delta}\right]} z_{j1} + \gamma_2^{\left[\overline{\delta}\right]} z_{j2} + \cdots + \gamma_k^{\left[\overline{\delta}\right]} z_{jk} = Z\gamma^{\left[\overline{\delta}\right]}.$$

Adding topic-level predictors changes the model only a little but changes the interpretation substantially. The interpretation changes in two important ways:

[5] Other summaries of the heterogeneity are possible. The parameter σ_δ serves as an estimate of the SD of δ in the population of topics, but researchers can also compute summaries for only those topics included in the study, such as the standard deviation or relevant quantiles of the δ_j. These summaries might be of interest when researchers use a "diverse collection" of topics rather than a random sample or the distribution of treatment effects for topics is not normal. However, the substantive conclusions are not likely to change whether one uses the σ_δ or an alternative for most applications.

[6] "Good predictors" explain variation in the intercept α_j, the treatment effect δ_j, or both. Alternatively, we can think of these predictors as predicting variation in the average for the control group, the average for the treatment group, or the difference between the two.

1. First, the parameter $\overline{\delta}_j$ is no longer a single-number summary, but a *conditional* typical treatment effect that depends on the topic-level predictors. For a single topic-level predictor z_{j1}, we would have $\overline{\delta}_j = \gamma_0^{\left[\overline{\delta}\right]} + \gamma_1^{\left[\overline{\delta}\right]} z_{j1}$, for example.
2. Second, the parameter σ_δ is no longer the SD of the treatment effect, but the SD of the treatment effects around the *expected* treatment effect $\overline{\delta}_j$.

To emphasize this important shift in interpretation, we describe these new quantities as the "*conditional* typical treatment effect" and the "*conditional* SD of the treatment effects." The "conditional" here refers to "conditional on the topic-level predictors" and emphasizes the change in interpretation. Clifford *et al.* (2023) provide an example of this approach (see Clifford *et al.*'s Figure 2).

4.3. Extensions

As we discussed above, our baseline design assumes that each topic contains a treatment group and a control group. However, researchers can also use the hierarchical model for Extensions 1a, 1b, and 2 discussed earlier, in which each topic is considered as treatment *or* control. In this case, we define the typical treatment effect as $\overline{\delta} = \overbrace{\left[\frac{1}{\mathcal{J}_T}\sum_{j=1}^{\mathcal{J}_T}\overline{Y}_j^T\right]}^{\text{typical outcome among treatment topics}} - \overbrace{\left[\frac{1}{\mathcal{J}_C}\sum_{j=1}^{\mathcal{J}_C}\overline{Y}_j^C\right]}^{\text{typical outcome among control topics}}$. We can construct hierarchical models directly from this definition. When the treatment group contains many topics and the control group contains many other topics (Extension 2), we have the model $y_i = \mu_{j[i]}^C(1-T_i) + \mu_{j[i]}^T T_i + \epsilon_i$, where $\mu_j^C \sim N\left(\overline{\mu}^C, \sigma_{\mu^C}^2\right)$ and $\mu_j^T \sim N\left(\overline{\mu}^T, \sigma_{\mu^T}^2\right)$. Then, $\overline{\delta} = \overline{\mu}^T - \overline{\mu}^C$. When the topic varies only within the treatment group (Extension 1a), we have $y_i = \mu^C(1-T_i) + \mu_{j[i]}^T T_i + \epsilon_i$, where μ^C is fixed and $\mu_j^T \sim N\left(\overline{\mu}^T, \sigma_{\mu^T}^2\right)$, so that $\overline{\delta} = \overline{\mu}^T - \mu^C$. When the topic varies only within the control group (Extension 1b), we have $y_i = \mu_{j[i]}^C(1-T_i) + \mu^T T_i + \epsilon_i$, where $\mu_j^C \sim N\left(\overline{\mu}^C, \sigma_{\mu^C}^2\right)$ and μ^T is fixed, so that $\overline{\delta} = \mu^T - \overline{\mu}^C$.

If researchers assign each respondent to multiple topics, they can model respondent-level variation with another set of varying parameters. In most applications, a varying intercept for each respondent might work well. However, researchers can use a varying intercept and a varying treatment effect for each respondent if they assign respondents to several treatment topics and several control topics.

Additionally, researchers can extend the parametric hierarchical model in the usual ways if they wish. For example, researchers can use binary logistic regression, multinomial logistic regression, or ordinal logistic regression to model categorical outcomes. As another example, researchers can use spline-based smooths or Gaussian processes to estimate nonlinear relationships between topic-level moderators and treatment effects.

4.4. Estimation Methods

There are two commonly used tools to estimate hierarchical models: reduced-information maximum likelihood (REML) and full posterior simulation. REML can produce estimates quickly (less than a second, in many cases), but it cannot effectively propagate uncertainty in the variance parameters $\begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\delta \\ \rho\sigma_\alpha\sigma_\delta & \sigma_\delta^2 \end{pmatrix}$ into the estimates for particular topics. However, REML is sufficiently popular and useful (because of its speed) that we evaluate this estimator below. Full posterior simulation is somewhat slower, but we ultimately suggest that researchers use full posterior simulation to estimate these models. REML occasionally produces unrealistic estimates of the variance parameters, and full posterior simulation better propagates uncertainty into the estimates for particular topics.

4.5. Evaluating the Hierarchical Model Using the SDC Data

We offer two estimators: (1) a nonparametric estimator using design-based assumptions and (2) a parametric, hierarchical model. The hierarchical model allows researchers to characterize the heterogeneity across topics, rather than simply marginalize across topics. However, researchers might worry about the robustness of the hierarchical model to violations of the parametric assumptions. What happens when the errors are not drawn precisely from a normal distribution? What happens when the treatment effects for particular topics are not drawn precisely from a normal distribution? To alleviate these concerns and illustrate the advantages of the hierarchical model, we use samples from an enormous survey experiment in which the true values of the quantities of interest are approximately known.

As a real-world test of the hierarchical model's ability to characterize the heterogeneity across topics, we use the SDC megastudy (Voelkel *et al.* 2023).[7] This study randomly assigns 32,059 individuals to 25 different interventions intended to lower partisan animosity (and other outcomes). The large SDC study is designed to estimate and compare the effects of several "promising interventions" on a common set of outcomes; this is distinct from our motivation of estimating the typical effect across a collection of theoretically connected topics.[8] However, this large data set allows us to compare estimates from samples to the approximately known values using the full data set. We conceptualize each of the 25 interventions in the SDC study as a diverse collection of topics—a set of interventions that might reduce partisan animosity. We are interested in how well the hierarchical model estimates (1) the SD of treatment effects across topics and (2) the treatment effects for each particular topic. We show that there is some bias in the estimates of the SD, but the 90% CI works well. We also show that there is some bias in the estimates of the treatment effects δ_j for particular topics, but the hierarchical model produces a smaller RMSE than the unbiased difference in means. The 90% CI works well for all quantities of interest.

4.5.1. The Model Specification

In this study, the topic (i.e., the particular intervention) only varies within the treatment (Extension 1a). We use a normal model for partisan animosity given the treatment

$$PA_{i[j]} = \alpha + \delta_j T_{i[j]} + \epsilon_i, \text{ where } \epsilon_i \sim N\left(0, \sigma_\epsilon^2\right).$$

Then, we model the treatment effect as a linear function of two topic-level predictors: (1) the degree to which the treatment references partisan animosity and (2) the degree to which the treatment corrects misperceptions of out-partisans (as a whole).[9]

$$\overline{\delta}_j = \gamma_0^{\left[\overline{\delta}\right]} + \gamma_1^{\left[\overline{\delta}\right]} \text{References PA}_j + \gamma_2^{\left[\overline{\delta}\right]} \text{Corrects Misperceptions}_j.$$

In the original study, two coders score both measures on a scale from 1 to 5. We average these two scores and standardize the scores across interventions to have an average of zero and an SD of 0.5.

4.5.2. The Monte Carlo Simulation

To establish the "truth" in our simulation study, we fit the model to the full SDC data set and compute point estimates for (1) the average of the treatment effects for the 25 interventions, (2) the SD of the treatment effects across the 25 interventions, and (3) the 25 treatment effects for the particular interventions. To create each sample for our Monte Carlo simulation, we randomly sample 10 of the

[7] Data available at https://osf.io/jzbnt/?view_only=285791c6d9f648b79da4200dff4889c6.

[8] From a broad community of academics and practitioners, the authors "selected the 25 most promising interventions in collaboration with an expert panel of social scientists and practitioners, basing our selections on evaluations of each submissions' likelihood of significantly reducing one or more of the target variables, novelty in the field, and uniqueness among the selected interventions" (Voelkel *et al.* 2023, 8). They describe the process in detail in their supporting information.

[9] For details on the topic-level predictors, see Section 14 of the Supporting Information for Voelkel *et al.* (2023).

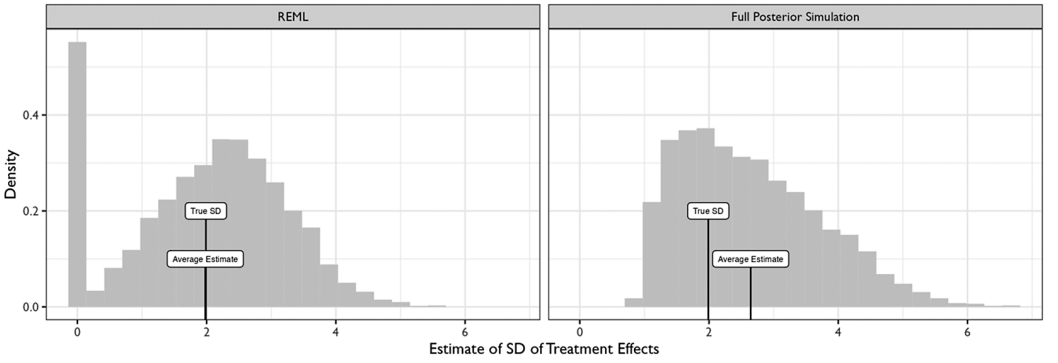

Figure 1. This figure shows the sampling distribution of the estimates of the SD of the treatment effects across topics. We compute this distribution by repeatedly taking small samples of 1,700 respondents and 10 topics from the SDC megastudy of 30,000+ respondents across 25 topics. Notice that the REML approach produces estimates of zero in some cases, while the full posterior simulation approach cannot rule out large SDs (e.g., greater than five) from only ten topics.

25 topics, and then, we randomly sample 100 respondents from each of those 10 topics. We randomly sample 700 respondents from the control group to mimic the roughly 7:1 ratio in the full study. We then fit the model to the sample of 1,700 respondents and compare the estimates using the sample to the estimates using the full SDC data set.

4.5.3. The Typical Effect

For the typical effect, both the point estimates and CIs work well in the simulation. The point estimates are approximately unbiased, as expected. More importantly, the simulations show that the CIs work well: The nonparametric 90% CI captures the true typical effect in 95% of the simulations, the REML 90% CI captures the true effect in 92% of simulations, and the full posterior simulation 90% CI captures the true effect in 94% of simulations.

4.5.4. The SD of the Treatment Effects

Figure 1 shows the sampling distribution for the (conditional) SD of the treatment effects for the REML and full posterior simulation estimates compared to the true SD of 1.99. Both estimates perform reasonably well. Notice that the REML estimate is approximately unbiased. However, REML occasionally returns a (problematic) estimate of zero; Chung *et al.* (2013) discuss this problem in detail. Thus, we recommend researchers use full posterior simulation rather than REML. For this simulation study, the full posterior simulation estimate is biased upward by about 30%. This bias occurs because the small number of topics (10, in this case) does not allow the likelihood to rule out large values of the SD. However, this bias shrinks as the number of topics grows.

The REML 90% CI captures the SD from the full sample about 85% of the time. The full posterior simulation 90% CI captures the SD from the full data set about 93% of the time—a slightly conservative interval. Thus, these data suggest that the hierarchical model offers a useful point estimate and CI for the (conditional) SD of the treatment effects.

4.5.5. The Treatment Effects for Particular Topics

We now turn to an evaluation of the point estimates of the treatment effects for particular topics. To evaluate the point estimates, we must keep in mind both bias and variance. An unbiased estimator is readily available; we could just use the difference in means from the control group and the particular treatment group. However, the hierarchical model produces estimates with a smaller root-mean-squared error (RMSE). The hierarchical model does introduce some bias, but the reduction in variance

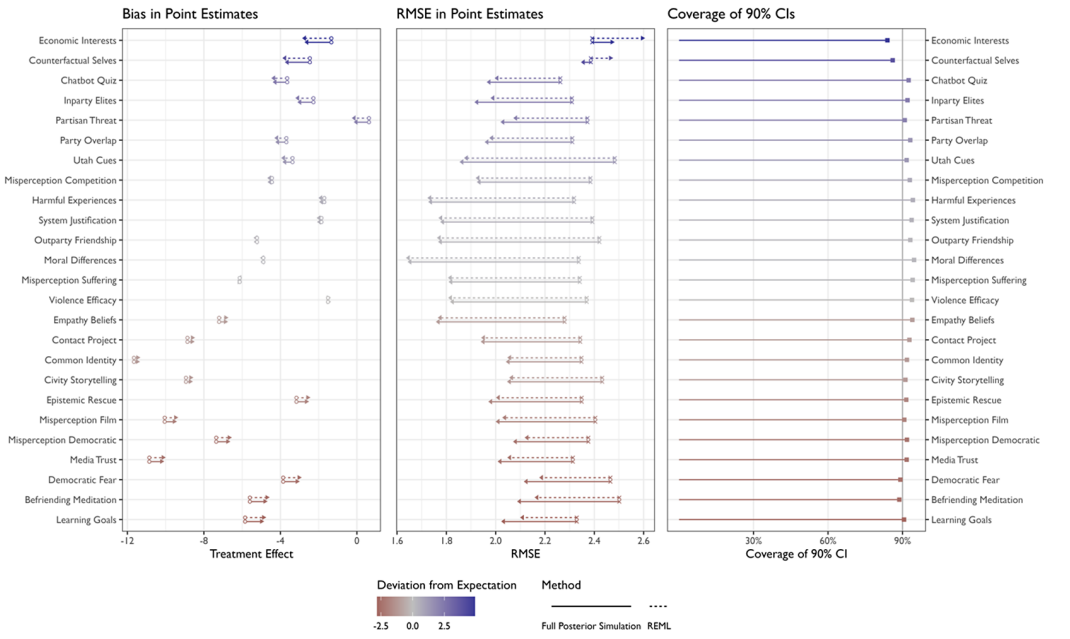

Figure 2. This figure shows the bias and RMSE of the point estimates and the coverage of the 90% CI from the hierarchical model. The topics are ordered by their deviation from expectation given topic-level predictors (top, treatment effect closer to zero than expected; bottom, treatment effect further from zero than expected). We compute the bias, RMSE, and coverage by repeatedly taking small samples of 1,700 respondents and 10 topics from the SDC megastudy of 30,000+ respondents across 25 topics. While the hierarchical model introduces some bias by pooling information across topics, it meaningfully reduces the RMSE and the 90% CIs work well. The **left panel** shows the bias for the point estimates of the treatment effects for particular topics. The hollow circles show the treatment effects from the full data set, and the arrowheads point to the expected value of the estimates. Thus, the length of the arrow shows the magnitude of the bias. The **middle panel** shows the RMSE. The x shows the RMSE of the difference in means, and the arrowhead shows the RMSE of the hierarchical model. Thus, the length of the arrow shows the reduction in RMSE when using the hierarchical model rather than the unbiased difference in means. The **right panel** shows the coverage of the 90% CI.

more than offsets this bias. Thus, we are interested in two summaries of the sampling distribution of the point estimates. First, what is the bias? Second, how does the RMSE of the hierarchical model compare to the RMSE for the difference in means?

The left panel of Figure 2 shows the bias for the point estimates of the treatment effects for particular topics. The hollow circles show the treatment effects from the full data set (which we treat as the truth in our simulations), and the arrowheads point to the expected value of the estimates. Thus, the length of the arrow shows the magnitude of the bias. The average absolute bias is only about 0.6 points on the 100-point partisan animosity scale. However, a 0.5-point bias is about 25% of the SD of the treatment effects across topics, so we consider it meaningful. The estimates for the interventions with unexpectedly large or small treatment effects (conditional on topic-level predictors) are the most biased. To show this, the color shows how unexpected the treatment effect is, or how far the treatment effect falls from the (conditional) typical effect. Red indicates that the intervention reduces partisan animosity more than typical, and blue indicates that the intervention reduces partisan animosity less than typical. For example, the Economic Interests intervention is much less effective at reducing partisan animosity than expected. Given the predictors—it does not correct a misperception, but it does clearly reference partisan animosity—the intervention should reduce partisan animosity by about 6.2 points on the 100-point scale. However, the reduction is actually about 1.3 points or about 2.5 SDs smaller than the (conditional) typical effect. This highlights the importance of topic-level predictors: If researchers have information that allows them to predict the bias—that is, those topics with larger or smaller treatment effects—then they should include those predictors in the model.

However, there is a tradeoff here between bias and variance. The hierarchical model not only introduces bias but also shrinks the variance of the estimates. This reduction in the variance usually more than offsets the errors due to bias. The middle panel of Figure 2 shows the RMSE. The x shows the RMSE of the difference in means, and the arrowhead shows the RMSE of the hierarchical model. Thus, the length of the arrow shows the reduction in RMSE when using the hierarchical model rather than the unbiased difference in means. For most topics, the RMSE is much improved. It shrinks by about 20 percent, on average, across the topics. For some topics, the RMSE does get worse. In these data, the RMSE gets substantially worse for one topic (Economic Interests). Again, this is the topic that differs substantially from the (conditional) typical effect (highlighting the importance of topic-level predictors).

The right panel of Figure 2 shows the coverage of the 90% CIs for each topic. We report these CIs only for the full posterior simulation estimates because REML cannot easily propagate uncertainty to the estimates for particular topics. The coverage is about 90% on average across the topics. For topics with effects that are unusually far from the (conditional) typical effect, the coverage can drift downward from 90%. For topics with effects that are unusually close to the (conditional) typical effect, the coverage can drift upward. For these data, the coverage ranges from 84% (Economic Interests) to 95% (moral differences).

5. Conclusion

Many experimental research programs offer general hypotheses that speak to a wide range of particular topics. In practice, many implemented experiments must use a *particular* topic—a particular test of the experimenter's claim—even if researchers' theory operates at a much more general level. As a research program advances, it should address the generalizability of the core hypotheses to a broad range of topics.

We do not suggest removing single-topic studies from our social science toolkit. On the contrary, they are essential tools. In the early stages of a research program, single-topic studies allow researchers to establish the plausibility of the core hypothesis. If researchers carefully select a topic with a large treatment effect, they can do this early work with a much smaller sample size (e.g., Clifford and Rainey 2023; Kuklinski *et al.* 2000). While findings from a single-topic study do not necessarily generalize to a broader collection of topics (Clifford and Rainey 2023), we are optimistic that researchers and readers can intelligently communicate and understand these limitations. For example, Bartels and Mutz (2009) limit their conclusions and highlight the usefulness of focusing on single topics (two, in their case):

> [Our study] is limited to only two controversial issues, two substantive arguments, and two institutions, which cannot claim to represent all potential persuasive contexts in which institutions render decisions. Moreover, these particular issues are much better known and understood by the public than many highly technical pieces of legislation decided by Congress or decisions made by the Court... (258)

However, researchers can use topic sampling to advance research programs past a handful of *ad hoc* particular topics to a large collection of topics of interest. Ideally, researchers can enumerate the population of topics and select a random sample. However, if the population cannot be enumerated, then a "diverse collection" of topics works well under stronger modeling assumptions. Clifford *et al.* (2023) suggest using about 25 to 50 topics and about 20% to 50% more respondents to obtain a precision comparable to a single-topic study. If the literature tends to use 1,000 respondents in single-topic studies, then researchers can perhaps use 1,200 or 1,500 respondents to make much more general—but similarly precise—claims about treatment effects in a broad population of topics.

We describe two complementary strategies to analyze the data from a topic-sampling experiment. We describe a nonparametric, unbiased estimator that allows researchers to estimate the typical

treatment effect across topics. This is analogous to a nonparametric difference-in-means test for a simple experiment with two conditions—it is unbiased under the assumptions of the design (random sampling of topics and randomization into treatment and control). We also describe a parametric, hierarchical model that allows researchers to effectively summarize the heterogeneity across topics. First, the hierarchical model has a scale parameter that researchers can estimate and interpret as the SD of the treatment effects across topics. This serves as a useful single-number summary of the heterogeneity—a give-or-take number around the typical effect. Second, the hierarchical model allows researchers to estimate the treatment effect for all the topics included in the study with surprising precision. Third, the hierarchical model allows researchers to describe the variation in the treatment effects for particular topics using topic-level predictors.

Concerns about the generalizability of experiments are not new—McDermott (2002) notes that concern about the generalizability of convenience samples has been a "near obsession" in political science (334). For example, recent work examines the generalizability of experiments using online convenience samples to nationally representative samples (Berinsky, Huber, and Lenz 2012; Coppock, Leeper, and Mullinix 2018) and of laboratory experiments to field experiments (Barabas and Jerit 2010; Coppock and Green 2015; Jerit, Barabas, and Clifford 2013). A growing literature focuses on the generalizability *across topics*. For example, research on incivility (Skytte 2022), fake news (Clemm von Hohenberg 2023), media (Wittenberg *et al.* 2021), partisan cues (Clifford *et al.* 2023; Tappin 2023), discrimination (Crabtree *et al.* 2022; Elder and Hayes 2023), and even placebos (Porter and Velez 2022) shows that substantive effects can meaningfully depend on the stimulus researchers choose. We applaud these substantive and methodological efforts. In this paper, we advance this work by offering a careful discussion of estimation. We suggest an unbiased estimator of the typical treatment effect and a hierarchical model to summarize the heterogeneity. Using the topic-sampling design suggested by Clifford *et al.* (2023) and the estimators described here, researchers can generalize beyond the treatment effect for a particular topic and estimate more general quantities of theoretical interest.

Data Availability Statement. All data and code to reproduce our results are available on Dataverse at https://doi.org/10.7910/DVN/YBV9Z8 (Rainey 2024).

References

Barabas, J., and J. Jerit. 2010. "Are Survey Experiments Externally Valid?" *American Political Science Review* 104 (2): 226–242.

Baribault, B., et al. 2018. "Metastudies for Robust Tests of Theory." *Proceedings of the National Academy of Sciences* 115 (11): 2607–2612.

Bartels, B. L., and D. C. Mutz. 2009. "Explaining Processes of Institutional Opinion Leadership." *Journal of Politics* 71 (1): 249–261.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48.

Berinsky, A. J., G. A. Huber, and G. S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20 (3): 351–368.

Bernardo, J. M. 1996. "The Concept of Exchangeability and Its Applications." *Far East Journal of Mathematical Sciences* 4: 111–122.

Bertrand, M., and S. Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013.

Brandt, M. J. and F. M. A. Wagemans. 2017. "From the Political Here and Now to Generalizable Knowledge." *Translational Issues in Psychological Science* 3 (3): 317–320.

Brutger, R., J. D. Kertzer, J. Renshon, D. Tingley, and C. M. Weiss. 2023. "Abstraction and Detail in Experimental Design." *American Journal of Political Science* 67 (4): 979–995. https://doi.org/10.1111/ajps.12710.

Chong, D., and J. N. Druckman. 2013. "Counterframing Effects." *Journal of Politics* 75 (1): 1–16.

Chung, Y., S. Rabe-Hesketh, V. Dorie, A. Gelman, and J. Liu. 2013. "A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models." *Psychometrika* 78: 685–709.

Clemm von Hohenberg, B. 2023. "Truth and Bias, Left and Right: Testing Ideological Asymmetries with a Realistic News Supply." *Public Opinion Quarterly* 87 (2): 267–292.

Clifford, S., T. J. Leeper, and C. Rainey. 2023. "Generalizing Survey Experiments Using Topic Sampling: An Application to Party Cues." *Forthcoming in Political Behavior.* https://doi.org/10.1007/s11109-023-09870-1.
Clifford, S., and C. Rainey. 2023. "The Limits (and Strengths) of Single-Topic Experiments." https://doi.org/10.31235/osf.io/zaykd.
Clifford, S., G. Sheagley, and S. Piston. 2021. "Increasing Precision without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* 115 (3): 1048–1065.
Coppock, A., and D. 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 (1): 113–131.
Coppock, A., T. J. Leeper, and K. J. Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates across Samples." *Proceedings of the National Academy of Sciences* 115 (49): 12441–12446.
Crabtree, C., S. M. Gaddis, J. B. Holbein, and E. Nergård Larsen. 2022. "Racially Distinctive Names Signal Both Race/Ethnicity and Social Class." *Sociological Science* 9: 454–472.
Elder, E. M., and M. Hayes. 2023. "Signaling Race, Ethnicity, and Gender with Names: Challenges and Recommendations." *Journal of Politics* 85 (2): 764–770.
Feller, A., and A. Gelman. 2015. "Hierarchical Models for Causal Effects." In *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, edited by R. Scott and S. Kosslyn, 1–16. John Wiley & Sons.
Gelman, A. 2005. "Analysis of Variance—Why It Is More Important Now Than Ever." *Annals of Statistics* 33 (1): 1–53.
Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.
Gelman, A., D. Simpson, and M. Betancourt. 2017. "The Prior Can Often Only Be Understood in the Context of the Likelihood." *Entropy* 19 (10): 1–13.
Grieco, J. M., C. Gelpi, J. Reifler, and P. D. Feaver. 2011. "Let's Get a Second Opinion: International Institutions and American Public Support for War." *International Studies Quarterly* 55 (2): 563–583.
Hainmueller, J., D. J. Hopkins, and T. Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22 (1): 1–30.
Jerit, J., J. Barabas, and S. Clifford. 2013. "Comparing Contemporaneous Laboratory and Field Experiments on Media Effects." *Public Opinion Quarterly* 77 (1): 256–282.
Kuklinski, J. H., P. J. Quirk, J. Jerit, D. Schwieder, and R. F. Rich. 2000. "Misinformation and the Currency of Democratic Citizenship." *Journal of Politics* 62 (3): 790–816.
Lindner, N. M., and B. A. Nosek. 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.
Mummolo, J., and E. Peterson. 2019. "Demand Effects in Survey Experiments: An Empirical Assessment." *American Political Science Review* 113 (2): 517–529.
Mutz, D. C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.
Porter, E., and Y. R. Velez. 2022. "Placebo Selection in Survey Experiments: An Agnostic Approach." *Political Analysis* 30 (4): 481–494.
Rainey, C. 2024. "Replication Data for: 'Estimators for Topic-Sampling Designs.'" Harvard Dataverse, V1. https://doi.org/10.7910/DVN/YBV9Z8.
Skytte, R. 2022. "Degrees of Disrespect: How Only Extreme and Rare Incivility Alienates the Base." *Journal of Politics* 84 (3): 1746–1759.
Tappin, B. M. 2023. "Estimating the Between-Issue Variation in Party Elite Cue Effects." *Public Opinion Quarterly* 86 (4): 862–885.
Tappin, B. M., A. J. Berinsky, and D. G. Rand. 2023. "Partisans' Receptivity to Persuasive Messaging Is Undiminished by Countervailing Party Leader Cues." *Nature Human Behaviour* 7 (4): 568–582.
Voelkel, J. G., et al. 2023. "Megastudy Identifying Effective Interventions to Strengthen Americans' Democratic Attitudes." OSF Preprints. March 20. https://doi.org/10.31219/osf.io/y79u5.
Wells, G. L., and P. D. Windschitl. 1999. "Stimulus Sampling and Social Psychological Experimentation." *Personality and Social Psychology Bulletin* 25 (9): 1115–1125.
Wittenberg, C., B. M. Tappin, A. J. Berinsky, and D. G. Rand. 2021. "The (Minimal) Persuasive Advantage of Political Video over Text." *Proceedings of the National Academy of Sciences* 118 (47): e2114388118.
Yarkoni, T. 2022. "The Generalizability Crisis." *Behavioral and Brain Sciences* 45: 1–78.

Cite this article: Clifford, S. and Rainey, C. (2024). Estimators for Topic-Sampling Designs. *Political Analysis*, 431–444. https://doi.org/10.1017/pan.2024.1