Political Analysis (2025), **33**, 164–170
doi:10.1017/pan.2024.20

PA

LETTER

The Limits (and Strengths) of Single-Topic Experiments

Scott Clifford[1] and Carlisle Rainey[2]

[1]Texas A&M University; [2]Florida State University

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

(Received 10 October 2023; revised 19 February 2024; accepted 30 March 2024; published online 24 January 2025)

Abstract

We examine the generalizability of single-topic studies, focusing on how often their confidence intervals capture the typical treatment effect from a larger population of possible studies. We show that the confidence intervals from these single-topic studies capture the typical effect from a population of topics at well below the nominal rate. For a plausible scenario, the confidence interval from a single-topic study might only be half as wide as an interval that captures the typical effect at the nominal rate. We highlight three important conclusions. First, we emphasize that researchers and readers must take care when generalizing the inferences from single-topic studies to a larger population of possible studies. Second, we demonstrate the critical importance of similarity across topics in drawing inferences and encourage researchers to consider designs that explicitly estimate and leverage similarity. Third, we emphasize that, despite their limitations, single-topic experiments have some important advantages.

Keywords: Experimental design; External validity; Heterogeneous treatment effects; Hierarchical models; Topic sampling

Edited by: Jeff Gill

1. Introduction

When researchers design an experiment, they must choose among many different experiments that could test the same general claim. Clifford, Leeper, and Rainey (2024) and Clifford and Rainey (2024) refer to these experiments as varying in their "topic," but "topic" can refer to substantive or ancillary details.[1] It is reasonable to suppose that each of these possible experiments has a different (average) treatment effect $\delta_j = E\left(y_{i[j]} \mid T_{i[j]} = 1\right) - E\left(y_{i[j]} \mid T_{i[j]} = 0\right)$, where $y_{i[j]}$ represents the potential outcome for respondent i under topic j and $T_{i[j]}$ represents the treatment indicator. However, researchers might want to generalize beyond any single topic that they might use in their experiment.[2]

Chong and Druckman (2013, 14) directly discuss the tension between their inferences about a single topic and inferences to their larger theoretical population:

> 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.

[1]For example, scholars studying the effects of a party cue on respondents' policy attitudes must focus on a particular policy. "Topic" suggests substantive details (e.g., allowing imported drugs from Canada; marijuana legalization), but can refer to ancillary details as well (e.g., the level of detail in a vignette experiment; Brutger *et al.* 2022).

[2]For example, recent research on incivility (Skytte 2022), fake news (Clemm von Hohenberg 2023), media (Wittenberg *et al.* 2021), partisan cues (Tappin 2023; Clifford, Leeper, and Rainey 2024), discrimination (Elder and Hayes 2023; Crabtree *et al.* 2022), and even placebos (Porter and Velez 2022) shows that substantive effects can meaningfully depend on researchers' decisions about the topic or design of the experiment.

This quote highlights a weakness of single-topic studies: the variation in treatment effects across topics might be large relative to the sampling variability in the estimate for the single topic. That is, the standard error from a single-topic study reflects the sampling error in estimating the treatment effect *for that topic*. The standard error does not include the error from choosing an unrepresentative topic, and this topic-level error might be large. Experimenters are sensitive to this issue. The quote from Chong and Druckman (2013) is not unrepresentative.[3] Thus, our goal with this paper is not to make experimentalists aware of this general issue—they are clearly well-aware and mindful of the problem. Instead, our goal is to use a formal framework to make the limitations of single-topic studies concrete and precise.

To generalize beyond a single topic, Clifford, Leeper, and Rainey (2024) and Clifford and Rainey (2024) suggest marginalizing across "topics" or across the collection of possible experiments that the researcher might use to test the same conceptual claim. They define a "typical" treatment effect $\overline{\delta}$ as the average of the treatment effects δ_j across topics. To estimate the typical treatment effect across topics, they suggest topic sampling: (1) take a random sample of topics from the larger population, (2) run parallel arms of the experiment for each topic, and (3) use a hierarchical model to pool the estimates and summarize the heterogeneity. Clifford, Leeper, and Rainey (2024) and Tappin (2023) offer recent examples of this basic approach and Clifford and Rainey (2024) describe the estimators in detail.[4]

The precise statement of the quantity of interest—the "typical" treatment effect across topics—allows a precise description of the slippage between studies of single topics and inferences about the larger collection. We explore this disconnect by focusing on the properties of confidence intervals (CI) from single-topic studies. We ask: *how often does the 95% CI from a single-topic study capture the typical treatment effect?*[5] We focus on two scenarios: (1) when the researcher chooses a topic at random and (2) when the researcher intentionally chooses an unrepresentative topic with a large treatment effect. By thinking carefully about the coverage of these intervals, we can reason precisely about the limits and strengths of single-topic studies.

2. A Randomly Chosen Topic

First, we consider the CIs for a *randomly chosen* topic. That is, in each repetition of the experiment, the researcher uses a new topic selected at random from the population. When there is *any* variation in the treatment effects across topics, the CIs from the single-topic study are too narrow to cover the typical treatment effect at the nominal rate.[6] As the variation across topics becomes large and/or the sampling variability in the single-topic estimate becomes small (e.g., large N), then the coverage (slowly) approaches 0%.[7]

Consider a stylized single-topic design. In this stylized design, the researcher selects N subjects and assigns half to treatment and half to control, then computes the difference in the two means to estimate

[3] Bartels and Mutz (2009, 258) write that their 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."

[4] For further emphasis from outside political science, see Wells and Windschitl (1999) on "stimulus sampling," Baribault *et al.* (2018) on "radical randomization," Almaatouq *et al.* (2024) on "integrative experimental design," and Yarkoni (2022) on "the generalizability crisis."

[5] The 95% confidence interval is a useful focus because it (1) has a clear evaluative criterion (i.e., 95% coverage), (2) is a common tool in practice, and (3) allows us to address variance estimation (i.e., the CI is too narrow in Figure 1) and bias (i.e., the CI is too far right in Figure 2) in a common framework. This approach corresponds exactly (or nearly so) to more traditional evaluative frameworks that focus on, say, the power function, but develops the intuition for the inferences more clearly and accessibly.

[6] This is a somewhat unfair critique of the single-topic design. The single-topic design is built to estimate the treatment effect *for a particular topic*. And while experimentalists are certainly aware of this limitation, we make the limitations concrete and precise. We formally describe the importance of the unaccounted variability across topics (in a statistical sense), and how that variability might limit the generalizability of a particular study of a single topic.

[7] The coverage is close to 0% only in extreme situations (e.g., a sample size of one million).

the treatment effect. Assuming equal variance σ_y^2, this difference has a standard error $\text{SE}_{\delta_j} = \sqrt{\frac{2\sigma_y^2}{\frac{1}{2}N}} = \sqrt{\frac{4\sigma_y^2}{N}} = \frac{2\sigma_y}{\sqrt{N}}$. Importantly, this standard error models the error in the estimated treatment effect due to randomly assigning respondents to treatment and control. This standard error does not include the error from choosing an unrepresentative topic.

Although it seems obvious that this standard error is for the estimate of the treatment effect for the particular topic, researchers might be tempted to expand their conclusions about the specific topic to the broader collection of possible topics, which might better represent the theoretical claim. For example, Bakker, Lelkes, and Malka (2020, 1073) use food irradiation and farm subsidies to determine which motive—bounded rationality or expressive utility—is "more salient for partisan cue-taking." They specifically note the limitations on generalizability, but their broader claim is more general than food irradiation and farm subsidies. Similarly, Nicholson (2012, 64) studies immigration and foreclosure policy and concludes that "out-party leaders play a more potent role in shaping partisan opinion." The author again addresses limitations on generalizability, but aims to generalize beyond immigration and foreclosure. Thus, using specific topics to test a broader theoretical claim makes it *tempting* to generalize beyond the specific topics. So, we ask: "under what conditions is it 'particularly bad' to generalize beyond the specific topic(s) of the experiment?" To answer this question, we use the framework from Clifford and Rainey (2024) to examine how often the single-topic CI captures the typical treatment effect in the broader population of topics.

Suppose that the particular treatment effect differs from the typical treatment effect by ψ_j and that ψ_j is a random variable with variance σ_ψ^2.[8] The parameter σ_ψ captures the similarity in treatment effects across topics. Treating ψ_j as random (e.g., the single topic is selected at random), the standard error of the difference-in-means becomes $\sqrt{\text{SE}_{\delta_j}^2 + \sigma_\psi^2}$. This implies that the standard error for the single-topic estimate needs to be $\sqrt{\frac{\sigma_\psi^2}{\text{SE}_{\delta_j}^2} + 1} = \sqrt{\frac{\sigma_\psi^2}{\frac{4\sigma_y^2}{N}} + 1}$ times larger to capture the typical effect across topics at the nominal rate.[9] Alternatively, the standard error for the single-topic study is only $\left(1/\sqrt{\frac{\sigma_\psi^2}{\frac{4\sigma_y^2}{N}} + 1}\right) \times 100\%$ as wide as the standard error for the typical treatment effect.

The party cue experiment in Clifford, Leeper, and Rainey (2024) allows us to select reasonable values of σ_y and σ_ψ. We use their data to estimate these parameters and obtain $\sigma_y = 1.87$ and $\sigma_\psi = 0.18$.[10] If we set $N = 1{,}000$, the deflation factor is $1/\sqrt{\frac{\sigma_\psi^2}{\frac{4\sigma_y^2}{N}} + 1} \approx 0.55$. Therefore, with a sample size of 1,000, the CI from a single-topic study is only about 55% as wide as it needs to be to capture the typical effect at the nominal rate.[11] For sample sizes ranging from 100 to 3,000, Panel A of Figure 1 shows that the deflation factor ranges from 90% to 35%.

Alternatively, we can compute the coverage of these intervals. How often does the CI from the single-topic study capture the typical treatment effect?[12] For the single-topic study with $\sigma_y = 1.87$, $\sigma_\psi = 0.18$,

[8] We might imagine two sources of this random variation. First, perhaps the researcher selects the topic at random from the population of topics. Second, we might imagine this particular treatment effect is like a realization of a random variable.

[9] Find c, such that $\sqrt{\text{SE}_{\delta_j}^2 + \sigma_\psi^2} = c\text{SE}_{\delta_j}$. Solving for c gives $c = \sqrt{\frac{\sigma_\psi^2}{\text{SE}_{\delta_j}^2} + 1} = \sqrt{\frac{\sigma_\psi^2}{\frac{4\sigma_y^2}{N}} + 1}$.

[10] In a similar experiment, Tappin (2023) estimates $\sigma_y = 1.8$ and $\sigma_\psi = 0.18$. See Table A8 on p. 18 of his supplementary information (osf.io/27szb). Tappin rescales his seven-point outcomes to range from zero to one, so the values in his table (0.3 and 0.03, respectively) need to be multiplied by six to match the seven-point scale of Clifford, Leeper, and Rainey (2024).

[11] For a sample size of only 100, the CIs need to be about 10% wider. In other words, the CIs will be quite wide with such a small sample, but *still* not wide enough to consistently capture the typical effect.

[12] If we assume that the coverage of the 95% CI for the treatment effect for a particular topic is $1 - 2\left(1 - \Phi(1.96)\right) = 95\%$, then this interval (for a randomly selected topic) will capture the typical treatment effect with probability $1 - 2\left(1 - \Phi\left(1.96/c\right)\right)$. For $N = 1{,}000$, $c \approx 1.8$ and $1 - 2\left(1 - \Phi\left(1.96/1.8\right)\right) = 72\%$.

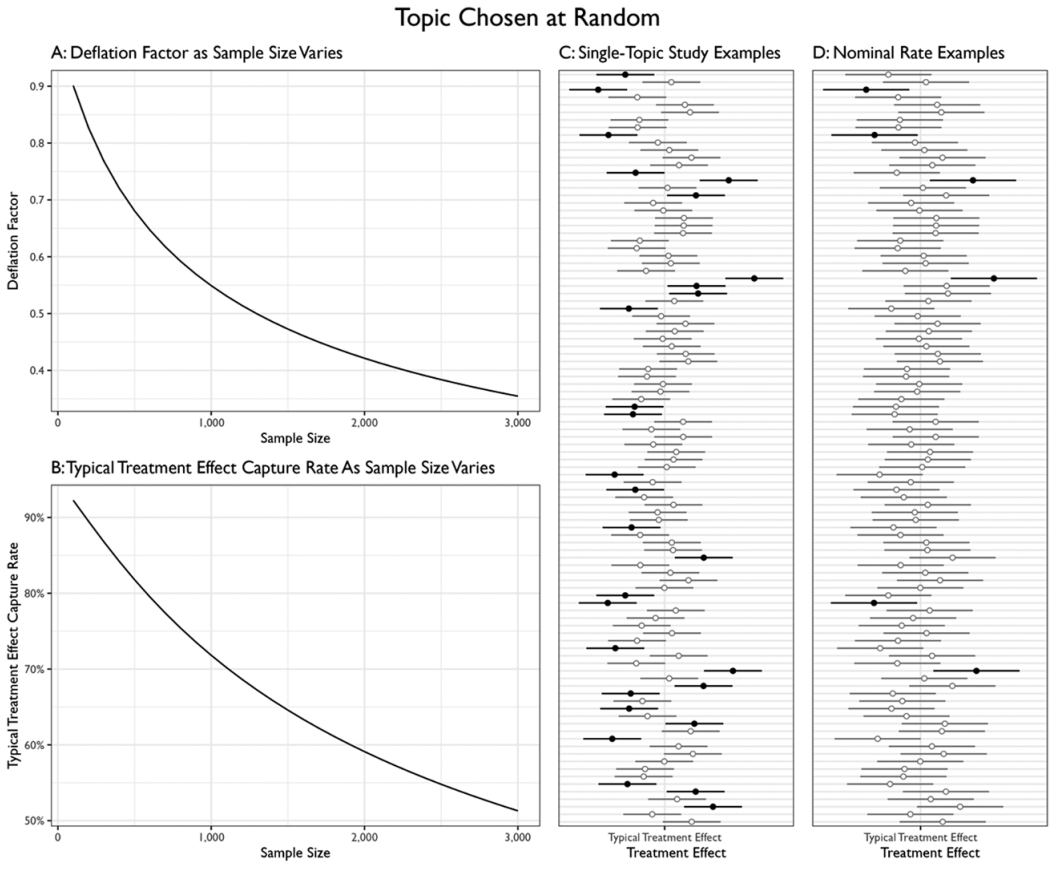

Figure 1. This figure shows how well the CI from a single-topic study captures the typical treatment effect in a larger population of possible experiments *when the topic is selected randomly from the population*. Panel A shows the deflation factor as the sample size varies. The deflation factor is the width of the single-topic CI compared to an interval that would capture the typical effect as the nominal rate. **Panel B** shows how often the single-topic CI captures the typical treatment effect as the sample size varies. **Panel C** shows simulated single-topic CIs for $N = 1{,}000$ and **Panel D** shows intervals that would capture the typical treatment effect at the nominal rate.

and $N = 1{,}000$, the 95% CI captures the typical treatment effect 72% of the time. This might seem surprisingly high. After all, the interval is only about half as wide as it needs to be to capture the typical effect 95% of the time. However, this interval fails to capture the typical treatment effect 28% of the time. If the researcher uses the single-topic study to infer the *direction* of the typical effect, the size of the test is 14% rather than the nominal 2.5% (one-tailed)—increasing the error rate by about 460%. The bottom panel of Figure 1 shows that the coverage ranges from 92% to 51% as the sample size ranges from 100 to 3,000.

To solidify the intuition for this process, Panels C and D of Figure 1 show many 95% CIs for single-topic studies with $N = 1{,}000$ and an alternative interval with nominal coverage. While the point estimates in Panel C equal the typical effect on average, the CIs are much too narrow to consistently capture the typical effect. In Panel D, we increase the width of the CIs (multiplying by c) to obtain the nominal coverage rate of 95%. These panels clearly show that the interval from the single-topic study is *much* too narrow. Many 95% CIs miss the typical effect and some miss *badly*.

Perhaps ironically, as estimates from a single-topic study become more precise, they become less likely to capture the typical treatment effect. One possible *incorrect* conclusion is that researchers should therefore prefer smaller samples sizes and wider CIs because they are more likely to capture the typical

effect; that's not correct. To study the typical effect, researchers must build on a careful collection of singletopic studies and generalize using appropriate tools (e.g., Clifford, Leeper, and Rainey 2024). So long as researchers remain mindful of the limitations, more precise estimates from single-topic studies are *helpful* in this effort to generalize.

3. An Intentionally Unrepresentative Topic

In addition to a randomly chosen topic, we consider the behavior of the 95% CI when using an intentionally unrepresentative topic. When initially testing an idea, researchers might use a topic that they expect creates especially large treatment effects. For example, in a study of partisan cues, Levendusky (2010, 119) selected issues for which respondents would have "weak prior beliefs" such that the "experimental manipulation—and not the respondent's pre-existing opinion—is the key source of information about the issue."[13] This can be a useful tool—it allows the researcher to obtain high statistical power without a large, expensive sample. But how often will the 95% CI from this single-topic study capture the typical treatment effect?

To analyze this situation, we assume that the researcher chooses a topic with a treatment effect that falls one standard deviation above average in the population of topics (or at the 84th percentile). Using the same values of $\sigma_y = 1.87$, $\sigma_\psi = 0.18$, and $N = 1{,}000$, only about 67% of the 95% CIs will capture the typical effect (almost all misses are *above* the typical effect). Perhaps more starkly, about one in three 95% CIs will *not* include the typical treatment effect. This means that the size of the test is 33%. If the researcher uses the single-topic study to infer that the typical effect is positive, the error rate under the null will be about 33% rather than the nominal 2.5% (one-tailed)—increasing the error rate by about 1,220% above the nominal rate. Panel A of Figure 2 shows that the capture rate ranges from 92% to 25% for sample sizes from 100 to 3,000. Panel B shows several simulated 95% CIs from this single-topic study to clarify the behavior of the CIs. These intervals are not too narrow, they are simply shifted above the typical effect (by design).

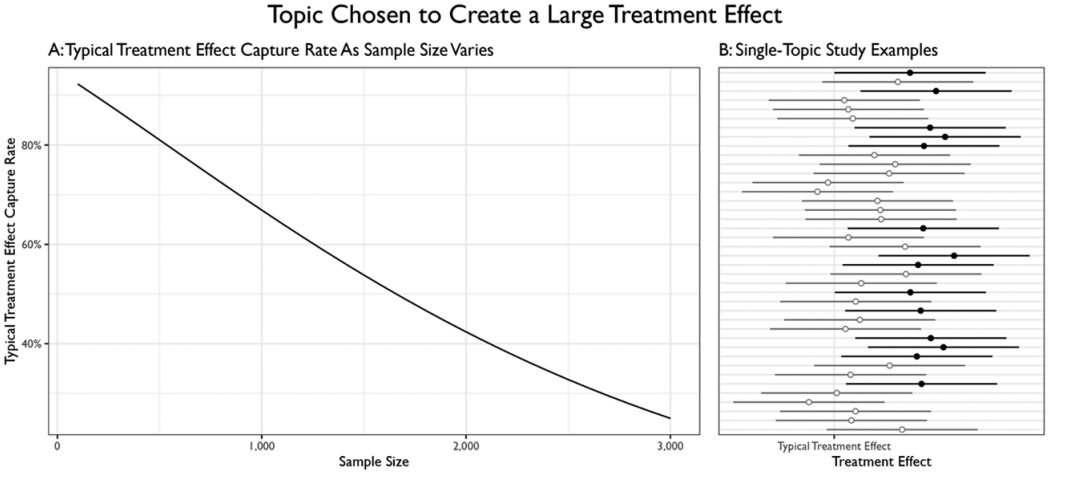

Figure 2. This figure shows how well the CI from a single-topic study captures the typical treatment effect in a larger population of possible experiments *when the researcher intentionally selected a topic with a large treatment effect (one standard* deviation above average or 84th percentile). **Panel A** shows how often the single-topic CI captures the typical treatment effect as the sample size varies. **Panel C** shows simulated single-topic CIs for *N = 1*,000.

[13] Similarly, in another study of partisan cues, Kam (2005, 167) selected the issue of food irradiation because it was "a relatively low-information, low-salience issue on which the parties have not taken clear stands."

4. Conclusion

In this research note, we consider the generalizability of single-topic studies. Researchers and readers must remember that single-topic studies are studies of *particular* topics and the estimates for that topic do not allow the researcher or reader to generalize to other topics without assumptions about the similarity of the topics. Experimentalists certainly appreciate that findings from one topic might not generalize to others, but we provide a concrete, *statistical* description of the problem that allows researchers and readers to properly limit their inferences and understand the risks of over-generalization.

We focus on how often the CIs from the single-topic studies capture the typical treatment effect and show that coverage is much lower than the nominal rate. Using hypothetical, but plausible parameters, we show that for a study with 1,000 respondents (1) the 95% CI for a randomly selected topic might capture the typical treatment effect about 71% of the time (missing high and low) and (2) the 95% CI for an intentionally chosen topic with a large treatment effect might capture the typical treatment effect 67% of the time (missing high). These are only illustrations for a plausible scenario and the properties will vary across contexts. However, our plausible scenario clearly illustrates the slippage between studies of single topics and inferences about other possible experiments.

We emphasize three conclusions. First, single-topic studies have limits, and we encourage researchers to remain mindful of those limits. Researchers should (continue to) take care when drawing inferences about a larger population of topics from a single-topic study, particularly regarding the size of an effect.

Second, there is, *of course*, a use for single-topic studies. They can be useful early in a research program to convincingly establish a particular effect because they allow researchers to obtain high statistical power without huge sample sizes. For many research problems, it seems reasonable to assume that treatment effects across topics will tend to have the same *direction.* If researchers are comfortable with this working assumption, then single-topic studies allow researchers to establish the plausibility of the theory relatively efficiently: they can select a topic with an especially large treatment effect and thus dramatically increase their statistical power. A common concern is pretreated respondents (e.g., Slothuus 2016). But careful selection of a single topic can allow researchers to avoid small effects due to pretreatment.[14] Of course, the generalizability of the treatment effect of the hand-picked topic—even the direction of the effect—remains an assumption in the absence of studies of other topics. Thus, single-topic studies are an important first step in the research process but should not be the last. With several careful studies of single topics to draw upon, researchers can improve generalizability by investing resources into systematically studying multiple topics (e.g., Clifford, Leeper, and Rainey 2024; Clifford and Rainey 2024) rather than pouring all their resources into a careful study of a single topic. Nonetheless, a collection of careful studies of single topics remains a valuable foundation—perhaps an *essential foundation*—for more generalizable work.

Third, this discussion highlights the statistical and substantive importance of the *similarity* of treatment effects across topics. When one conducts a single-topic study it is common and natural to speculate about how the estimated effects might generalize to other topics. This leads to researchers to think about the *variability* of effects across topics and leads naturally to the topic-sampling framework (Clifford, Leeper, and Rainey 2024; Clifford and Rainey 2024). This motivation highlights that knowledge of the variability across topics is critical for drawing inferences about the typical effect. But estimating the variability is not only statistically helpful, it is also substantively meaningful. The variability in treatment effects across topics is a substantive quantity of interest that researchers should consider further. The framework from Clifford and Rainey (2024) allows us to describe the impact of variability in treatment effects across topics on inferences. But, more importantly, their framework also allows researchers to easily design experiments to estimate the similarity and the typical treatment effect—what we assume is (or at least, might sometimes be) the ultimate quantities of interest.

[14] Of course, the subset of issues with little to no pretreatment may still vary in important ways and systematically differ from issues with more pretreatment.

Data Availability Statement. All code to reproduce our results is available on Dataverse at https://doi.org/10.7910/DVN/QX0BSK (Rainey 2024).

References

Almaatouq, A., et al. 2024. "Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences." Forthcoming in *Behavioral and Brain Sciences*. (47): 1–18.

Bakker, B. N., Y . Lelkes, and A . Malka. 2020. "Understanding Partisan Cue Receptivity: Tests of Predictions from the Bounded Rationality and Expressive Utility Perspectives." *The Journal of Politics* 82(3): 1061–1077.

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." *The Journal of Politics* 71(1): 249–261.

Brutger, R., J. D. Kertzer, J. Renshon, D. Tingley, and C. M. Weiss. 2022. "Abstraction and Detail in Experimental Design." *American Journal of Political Science* 67(4): 979–995.

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

Clifford, S., and C. Rainey. 2024. "Estimators for Topic-Sampling Designs." *Political Analysis* 32(4): 431–444.

Clifford, S., T. J. Leeper, and C. Rainey. 2024. "Generalizing Survey Experiments Using Topic Sampling: An Application to Party Cues." *Political Behavior* 46(2): 1233–1256.

Crabtree, C., S. M. Gaddis, J. B. Holbein, and E. N. 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." *The Journal of Politics* 85(2): 764–770.

Kam, C. D. 2005. "Who Toes the Party Line? Cues, Values, and Individual Differences." *Political Behavior* 27(2): 163–182.

Levendusky, M. S. 2010. "Clearer Cues, More Consistent Voters: A Benefit of Elite Polarization." *Political Behavior* 32(1): 111–131.

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

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: The Limits (and Strengths) of Single-Topic Experiments". Harvard Dataverse, V1. 10.7910/DVN/QX0BSK.

Skytte, R. 2022. "Degrees of Disrespect: How Only Extreme and Rare Incivility Alienates the Base." *The Journal of Politics* 84(3): 1746–1759.

Slothuus, R. 2016. "Assessing the Influence of Political Parties on Public Opinion: The Challenge from Pretreatment Effects." *Political Communication* 33(2): 302–327.

Tappin, B. M. 2023. "Estimating the Between-Issue Variation in Party Elite Cue Effects." *Public Opinion Quarterly* 86(4): 862–885.

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

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): 1–7.

Yarkoni, T. 2022. "The Generalizability Crisis." *Behavioral and Brain Sciences* 45(1): 1–78.

Cite this article: Clifford, S. and Rainey, C. (2025). The Limits (and Strengths) of Single-Topic Experiments. *Political Analysis*, 164–170. https://doi.org/10.1017/pan.2024.20