New Evidence and Design Considerations for Repeated Measure Experiments in Survey Research

Diana Jordan
Duke University

Trent Ollerenshaw
University of Houston

Andrew Trexler*
University of Wisconsin–Madison

October 15, 2025

Abstract

We re-examine recent influential claims that repeated measure experimental designs do not introduce bias and offer large precision gains in survey research (Clifford, Sheagley, and Piston 2021). We test these claims by experimentally varying the design of six classic political science experiments across three distinct large samples of U.S. adults (total $N = 13,163$). In contrast to the original study, we observe consistent attenuation of treatment effects in repeated measure designs. However, this average design effect is small enough, and the precision gains large enough, that we largely affirm the recommendation to employ repeated measure designs in many practical research applications. We additionally extend the literature on repeated measure designs by exploring how several design considerations affect the bias-precision trade-off, such as the use of within-subject versus between-groups designs, the relative separation of repeated measures within single surveys, and differences in respondent characteristics across sample types.

*Corresponding author. Please direct correspondence to atrexler@wisc.edu.

Political scientists increasingly leverage randomized experiments to estimate causal effects in human subjects research, particularly through surveys. A common experimental design, the between-groups "post-only" design, randomly assigns participants to treatment conditions and measures the outcome variable(s) only post-treatment. The average treatment effect (ATE) is estimated by comparing the outcome means across groups. However, this predominant design suffers from low precision when estimating treatment effects and may therefore miss small or heterogeneous treatment effects (Mutz 2011), while also risking substantial overestimates of effects (Gelman and Carlin 2014; Loken and Gelman 2017). Given the increasing evidence that low precision contributes to low replicability rates in social science research (Arel-Bundock et al. 2022; Gelman and Carlin 2014), improving experimental design is essential for advancing research in political science and related disciplines.

A common alternative experimental design is the "repeated measure" design, which measures outcomes both pre- and post-treatment. By measuring respondents' pre-treatment outcome levels, repeated measure designs can substantially increase precision in ATE estimates. Yet researchers have often been reluctant to implement repeated measure designs, especially within the same survey, due to concerns that pre-treatment measurement of outcomes may inadvertently bias ATE estimates by priming respondents to the treatment, inducing pressure to provide consistent responses or creating demand incentives. Lacking clear evidence about the degree of bias introduced versus precision gained, researchers have historically opted against repeated measure designs.

However, a recent influential study by Clifford, Sheagley, and Piston (2021, referred to as CSP hereafter) in the *American Political Science Review* experimentally manipulates design type, providing evidence that repeated measure designs enhance precision without biasing ATE estimates. CSP conclude that traditional concerns about repeated measure designs can be largely dismissed, recommending "that researchers use pre-post and within-subject designs whenever possible" (Clifford, Sheagley, and Piston 2021, 1062). These recommendations have gained traction in the social sciences: in the first four years since publication, CSP was cited by 118 peer-reviewed studies, of which 83 cite CSP specifically to justify using repeated measure designs.

The rapid adoption of repeated measure designs speaks to the importance of CSP's findings. Yet CSP's conclusions rest on just six experiments—a valuable but ultimately limited basis for such a broad shift in survey experimental practice, and thus one that merits large-scale replication. Further, researchers lack information on key design considerations that could impact the utility of repeated measure designs in some settings. For example, it remains unclear whether these designs are suitable for short surveys. CSP placed their pre- and post-treatment measures far apart, reflecting the intuition that placing them close together might increase bias by making the repetition more apparent. In this and other respects, best practices for implementing repeated measure designs remain underdeveloped.

In a large-scale replication and extension, we substantially expand the available evidence on repeated measure designs and address three key knowledge gaps. First, we assess the suitability of repeated measure designs for between-groups versus within-subject experiments. Second, we analyze how the proximity between repeated measures alters design effects, offering insights on the suitability of repeated measures designs when pre- and post-treatment measures are placed close together. Third, we conduct experiments on both probability- and non-probability-based samples with diverse respondent pools to assess how respondent characteristics like professionalization and attentiveness affect the bias-precision trade-off.

We experimentally manipulate the design of six published political science experiments, including three within-subject experiments and three between-groups experiments to allow for comparison across experiment types. We randomly vary the proximity of repeated measures in these experiments to evaluate how this design consideration affects bias and precision. We field all six experiments in omnibus surveys on three distinct online samples of U.S. adults ($N_j = 18$ studies, $N_i = 13,163$ respondents, $N_{ij} = 78,978$ total observations). These include a sample from the probability-based AmeriSpeak panel maintained by NORC ($n_i = 4,033$) and two non-probability samples (Lucid $n_i = 4,869$, Prolific $n_i = 4,261$).

These large samples provide excellent statistical power to detect small design effects and assess moderators.

Contrary to CSP's original findings, we observe a small but consistent attenuation of treatment effects in repeated measures designs relative to post-only designs. Despite this design effect, our findings largely affirm CSP's case for repeated measure designs, as the substantial precision gains often outweigh the weak attenuation in treatment effects to produce (in expectation) more accurate ATE estimates in many practical applications. Further, we provide robust evidence that repeated measure designs are suitable for both within-subject and between-groups experiments, across probability and non-probability samples with varying levels of respondent professionalization and attention, and in surveys where repeated measures must necessarily appear in close proximity. That said, we also find some evidence that asking attitude-recall questions or fielding multiple repeated measure experiments in one survey may exacerbate later design effects. In sum, while we identify some circumstances where well-powered post-only designs may be preferable, our findings reinforce the field's nascent shift toward repeated measures designs and the enhanced precision they offer.

Repeated Measure Designs in the Social Sciences

Survey experiments are widely used for social inquiry, with the "post-only," between-groups design being the most common in political science (Clifford, Sheagley, and Piston 2021). In this design, participants are randomly assigned and exposed to treatment or control stimuli, then outcomes are measured post-treatment and compared across conditions, with differences between the treatment groups' outcomes interpreted as the average treatment effect (ATE). Under a set of relatively weak assumptions—successful randomization, the stable unit treatment value assumption (SUTVA), no differential attrition—the post-only design provides unbiased estimates of the ATE.

A major downside of post-only designs is that the treatment effect is often imprecisely estimated. Treatment interventions in the social sciences typically explain only a small fraction of the variation in the outcome variable. Post-only designs therefore often have large residual errors, reducing statistical power—a critical consideration for experimental design (Rainey 2025). Statistical power (β) is the probability that a test rejects its null hypothesis in favor of a specified alternative hypothesis if it is true, a common goal of experiments. Power for a two-tailed test of a treatment effect (τ) can be expressed as

$$\beta = 1 - \Phi_{cdf}[\Phi^{-1}_{std}(1 - \frac{\alpha}{2}) \cdot SE_{\tau}; \mu = \tau, \sigma = SE_{\tau}] \tag{1}$$

where Φ_{cdf} is the cumulative density function of a normal distribution, Φ^{-1}_{std} is the inverse of the standard normal distribution, α is the chosen significance threshold of the test (for example, 0.05), and SE_{τ} is the standard error of the treatment effect τ, whose sampling distribution is assumed to be a normally distributed variable with a mean of $\mu = \tau$ and standard deviation of $\sigma = SE_{\tau}$. With an ordinary least squares (OLS) estimator, a standard approach for hypothesis testing in survey experiments, the standard error of the ATE ($SE_{\hat{\tau}}$) is estimated as

$$\widehat{SE}_{\hat{\tau}} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{N} (D_i - \bar{D})^2}} \tag{2}$$

where D_i, an indicator for assignment to treatment, and the root mean squared error $\hat{\sigma}$ is an asymptotic function of the residual errors $\hat{u}_i$ and the sample size N:

$$\hat{\sigma} \approx \sqrt{\frac{\sum_{i=1}^{N} \hat{u_i}^2}{N}} \tag{3}$$

The large residuals common to post-only designs thus reduce statistical power by producing large standard errors around the treatment effect, expanding the confidence interval around $\hat{\tau}$ and reducing the probability that this interval excludes the null value of the parameter of interest (e.g., $\tau = 0$), thus reducing the likelihood the null hypothesis can be rejected. Post-only designs therefore require large samples to reliably detect and precisely estimate treatment effects (Peters 2017).

Imprecision has myriad negative effects on scientific knowledge production. Imprecise studies risk failing to detect small treatment effects and variations in effects (Mutz 2011) and may overestimate effect sizes (Gelman and Carlin 2014; Loken and Gelman 2017). Structural incentives to publish "positive" findings meeting conventional significance thresholds can lead to published experiments with noisy data and brittle evidence propping up theories (Gerber, Green, and Nickerson 2001; Kühberger, Fritz, and Scherndl 2014). Statistical imprecision and underpowered experiments are increasingly recognized as major contributors to low replicability rates in social science research (Arel-Bundock et al. 2022; Gelman and Carlin 2014). Increasing precision in survey experiments is vital to enhancing the credibility of empirical social science.

Repeated measure designs offer improvements over post-only designs in terms of precision and power. Researchers have long recognized that the standard errors of estimated treatment effects can be reduced by adjusting for pre-treatment covariates, as this reduces the residual errors $\hat{u}_i$ by accounting for some additional variation in the outcome variable. This accordingly reduces $SE_{\hat{\tau}}$ by approximately $\sqrt{(1-\rho^2)}$ (Cox and McCullagh 1982; Bloom 1995), where ρ is the correlation between the outcome variable $\hat{Y}_i$ and the pre-treatment covariate $\hat{X}_i$—meaning that the stronger the correlation, the greater the reduction in $SE_{\hat{\tau}}$. Statistical power thus improves to:

$$\beta \approx 1 - \Phi_{cdf}[(\Phi_{std}^{-1}(1-\frac{\alpha}{2}) \cdot \sqrt{(1-\rho^2)} \cdot SE_{\tau}; \mu = \tau, \sigma = SE_{\tau}] \qquad [4]$$

The logic of repeated measure designs is that the pre-treatment covariate most likely to strongly correlate with the post-treatment outcome is an identical pre-treatment outcome measure. This design therefore measures outcomes both before and after exposure to treatment. By adjusting for respondents' pre-treatment outcome levels, this approach greatly enhances the precision of treatment effect estimates, substantially reducing the sample size required to achieve conventional levels of statistical power. Repeated measures designs come in two main types: between-groups, where respondents are randomized to either treatment or control stimuli, and within-subject, where all respondents receive both stimuli (List 2025;

Clifford, Sheagley, and Piston 2021).

Despite these advantages, researchers often worry that repeated measure designs may bias the ATE estimate. Repeated measure designs require an additional assumption beyond those in post-only designs: that pre-treatment measurement does not itself influence post-treatment outcomes differentially across treatment arms. Three concerns cast doubt on that assumption: *priming*, where pre-treatment measurement may lead respondents to focus on specific considerations (e.g., Klar, Leeper, and Robison 2020); *consistency pressures*, where respondents may feel pressure to provide post-treatment responses that align with their pre-treatment responses (e.g., Cialdini, Trost, and Newsom 1995; Tourangeau and Rasinski 1988); and *demand effects*, where respondents may adjust their post-treatment responses based on their perception of the study's purpose (e.g., Charness, Gneezy, and Kuhn 2012; Zizzo 2010; but see Mummolo and Peterson 2019).

Conventional wisdom thus suggests a trade-off between bias and precision when considering post-only or repeated measure designs. In practice, political science survey experiments have typically prioritized minimizing bias over addressing imprecision, defaulting to post-only designs (Clifford, Sheagley, and Piston 2021). To our knowledge, however, CSP is the only study to date that empirically tests the bias-precision trade-off for repeated measure designs. Their internal meta-analysis of six experiments found no significant differences in estimated ATEs between the two designs, but found that repeated measures designs substantially improve precision, allowing researchers to achieve more power with fewer participants. For instance, a 1,000 respondent two-arm post-only experiment has roughly 80 percent power to detect a treatment effect of 0.20 standard deviations, but a repeated measure design can achieve the same power with about 200 to 600 respondents, depending on the strength of the correlation between pre- and post-treatment measures. Given these precision gains and minimal evidence of bias, CSP argue that there is no meaningful bias-precision trade-off and recommend that researchers employ repeated measure designs as the default.

Contribution and Hypotheses

As of April 2025, CSP (2021) has been cited in 118 peer-reviewed studies, of which 83 are original studies citing CSP to justify a repeated measure design (see Appendix Table A.5.1 for the full list of studies). While most citations are from political science journals, the citations span 83 journals in a range of fields, including communication, criminology, economics, education, and environmental studies. The article's broad influence on experimental practice is already clear and likely to grow as disciplines become more critical of underpowered experiments (Arel-Bundock et al. 2022; Ioannidis, Stanley, and Doucouliagos 2017; Open Science Collaboration 2015).

CSP provides a valuable, overdue examination of the bias-precision trade-off in repeated measure designs. However, the empirical literature on this design remains limited. CSP's analysis is only well-powered to rule out large design effects—on the order of altering the ATE by 40 percent or more (Huber and Graham FC). Moderate design effects, which could be substantively meaningful in many contexts, require larger samples to detect.

Additionally, key questions about best practices for repeated measure designs remain unanswered. First, 41 percent of studies citing CSP used within-subject repeated measure designs rather than between-groups designs (see Appendix Table A.5.1), yet just one of CSP's experiments employed a within-subject design—a rather limited basis for a large shift in research practice. Because within-subject designs assign *all* participants to treatment (either before or after the control), they may alter the scope of consistency, demand, or priming effects relative to between-groups designs—meaning that design effects may differ in magnitude across the two approaches.

Our study substantially expands the evidence on design effects under within-subject designs by replicating three within-subject experiments in each of three samples, totaling nine studies with a meta-analytic sample over 43 times larger than the single study analyzed by CSP. Simultaneously, we replicate three of CSP's between-groups experiments on the same samples to further expand the evidence base for between-groups repeated measure designs. This allows us to rigorously test the preregistered[1] hypothesis that:

H1: Repeated-measure experimental designs do not bias estimated ATEs in either (a) between-group experiments or (b) within-subject experiments.

Second, we are not aware of any study to date that assesses how the proximity between repeated measures affects bias and precision. An intuitive hypothesis is that increasing the distance (i.e., adding more survey content) between repeated measures could reduce bias by mitigating priming, obscuring researcher intent, and enabling respondents to "forget" their pre-treatment responses, reducing pressure (or ability) to respond consistently. Indeed, given this intuition, many repeated measure experiments employ multi-wave panel surveys, allowing days or weeks of separation between measures. In single surveys, researchers (including CSP) commonly place their pre- and post-treatment questions at opposite ends of the survey. However, experimenters frequently work with very short surveys (or short modules in omnibus surveys), facing resource or logistical constraints that may require placing repeated measures close together. Further, close proximity may even be advantageous if it reduces random noise, strengthening the correlation between the repeated measures and increasing precision. We manipulate proximity between repeated measures in our surveys to test the following preregistered hypothesis:

H2: Repeated-measure experimental designs increase bias in estimated ATEs when measures are repeated measures are presented close together.

Third, CSP's six experiments used two student and four online non-probability samples. While their findings are important given the reliance on such convenience samples in experimental research (Jerit and Barabas 2023; Krupnikov and Levine 2014), it is unclear if the null design effect CSP observe is due to sampling design. Student samples differ from older adults on a variety of attitudinal and behavioral dimensions (Sears 1986). Probability-based sampling designs recruit respondents that are not only more representative of the target

[1] Anonymized preregistration materials are available here.

population, but also less professionalized and more attentive than opt-in non-probability panelists (Kennedy et al. 2016; MacInnis et al. 2018).

Differences in respondent characteristics may affect the relative strength of priming, consistency, or demand effects in repeated measure designs. For example, one of CSP's experiments ($N = 965$ students) revealed that many respondents whose responses changed pre-post also self-reported that their attitudes stayed the same. This may result from the unobtrusiveness of repeated measures—affirming their utility—but respondent inattentiveness may also play a role. Attentive respondents may be more likely to recognize repeated questions and alter their post-treatment responses accordingly. Similarly, professionalized respondents may be accustomed to experiments and react differently to repeated measures than less professionalized respondents. We explore these possibilities by fielding identical experiments on both probability and non-probability samples.

Data and Methods

We replicate six previously published survey experiments, summarized in Table 1, randomly manipulating each experimental design (post-only vs. repeated measure).[2] We briefly describe each experiment, with additional information provided in Appendix B.

In Study 1, we replicate a classic information treatment experiment on support for foreign aid (from Gilens 2001), in which treated respondents are informed that foreign aid spending represents about 1 percent of the U.S. federal budget. We expect this treatment to increase support. In Study 2, we replicate a party cues experiment from CSP on policy support for prescription drug imports from Canada, in which treated respondents are told that Democrats support and Republicans oppose this policy. Here, we analyze the second difference in support between Democrats and Republicans among those treated versus not treated. We expect the treatment to increase support among Democrats and decrease sup-

[2] This research was approved by the Institutional Review Board of [REDACTED] under protocol [REDACTED]. We further affirm that this research adheres to the American Political Science Association's Principles and Guidance for Human Subjects Research.

port among Republicans, widening the gap between the parties. In Study 3, we replicate a framing experiment from CSP on support for genetically modified organisms (GMOs), in which respondents are either treated with positively-framed information about GMOs (treatment) or negatively-framed information about GMOs (control). We expect the positively-framed treatment to increase support relative to the negatively-framed control. In Study 4, we replicate a classic question wording experiment on support for anti-poverty spending (from Smith 1987) that asks about support for spending on "welfare" or "assistance to the poor." We expect support to be higher when spending is described as assistance to the poor relative to welfare. In Study 5, we replicate a classic study on affirmative action from Wilson et al. (2008) which asks about support for affirmative action for women or for racial minorities. We expect support to be higher when the policy is aimed at women relative to racial minorities. In Study 6, we replicate a study from de Benedictis-Kessner and Hankinson (2019) on support for opening a new methadone clinic to address opioid addiction, in which the proposed clinic would be nearby (a quarter mile away) or further away (two miles away) from where the respondent lives. We expect that support will be higher in the latter condition. We thus define treatment and control (somewhat arbitrarily) such that the relevant ATE in each study is expected to be positive, to facilitate comparison across all six experiments.

These six studies were selected for their relative brevity (no more than three questions each), allowing us field more studies in a single survey, and because each found moderate to large treatment effects in the original studies, improving our ability to estimate design effects.[3] We purposively selected replication studies to cover diverse topics and treatments (e.g., informational treatments, party cues, framing effects) to provide breadth across areas of substantive inquiry (Clifford, Leeper, and Rainey 2024; Clifford and Rainey 2025). Four of our studies also appear in CSP's original paper[4] and we supplement these with two

[3] That is, while repeated measure designs are appropriate and even advantageous for studying small treatment effects because of increased power, we opted for studies with more powerful treatments to minimize risk that floor effects could cause a false negative in our estimation of design effects.

[4] Studies 1–4 described here correspond to studies 2, 5, 6, and 1 in CSP, respectively.

Table 1: Summary of Replicated Survey Experiments

	Topic	**Design**	**Treatment**	**Control**
1	Foreign Aid	Between-Groups	Foreign Aid 1% of Budget	No Information
2	Drug Imports	Between-Groups	DEM Favors, REP Opposes	No Party Cues
3	GMOs	Between-Groups	Pro: Prevents Blindness	Con: Uncertain Health Effects
4	Anti-poverty	Within-Subject	Assistance to the Poor	Welfare Spending
5	Affirmative Action	Within-Subject	Target Women	Target Racial Minorities
6	Opioid Clinic Policy	Within-Subject	Clinic 2 Miles Away	Clinic 1/4 Mile Away

experiments from Wilson et al. (2008, denoted Study 5) and de Benedictis-Kessner and Hankinson (2019, denoted Study 6) to increase the number of within-subject studies. We thus have three between-groups and three within-subject repeated measure designs for comparison against otherwise equivalent post-only designs.

Experimental Design

We fielded all six studies on three omnibus surveys and manipulated the experimental designs in a preregistered multi-stage randomization procedure (detailed in Appendix B.3). All respondents in each sample (combined $N_i = 13,163$ respondents) completed all six experiments (combined $N_{ij} = 78,978$ observations). Our randomization procedure independently assigned design conditions (post-only or repeated measures), treatment conditions (treatment or control stimuli), and the order of the experimental content for each respondent.

Of note, we hypothesized that design effects might be more pronounced when repeated measures are close together because respondents may be more likely to remember answering the same or similar questions, strengthening any priming, consistency, or demand effects. To test for this potential heterogeneity, our procedure has three noteworthy design features.

First, respondents were randomly assigned to four repeated measure designs and two post-only designs. The relatively larger share of repeated measure designs increases power for testing differences in design effects at various "distances" between repeated measures (defined as the number[5] of survey items separating the repeated measures). Second, we randomly assigned approximately half of respondents to complete two repeated measure experiments in very close succession, with just 0-3 units of separation between measures. This increases statistical power where we suspected we might find non-linear changes in design effects—i.e., when repeated measures are very close together—while still allowing for comparisons across distances by independently randomizing the content that could appear in between any given pair of repeated measures. Third, we included six wholly unrelated questions about the National Football League (NFL), randomized alongside the main content, to extend the right tail of the "distance" distribution and provide additional distractor items that could appear between repeated measures.[6] The realized distribution of distances is shown in Figure 1 (see Appendix B.3 for details and illustrative examples).[7]

Our randomization procedure thus provides us with unbiased estimates of design effects while maximizing statistical power where we expected (a priori) that it would matter most—that is, where repeated measures appear very close together in the survey. By comparing point estimates for the ATE under the post-only versus repeated measure design, we can identify design effects (i.e., bias) introduced from repeated measure designs (addressing H1). We can also identify precision gained from repeated measure designs by comparing standard

[5] A slight caveat: we follow the Time-sharing Experiments for the Social Sciences (TESS) guidelines for evaluating the "length" of survey items, such that our four longest items each count as two units of distance. Specifically, two items each for the GMO and opioid experiments include longer paragraphs that precede the outcome question, and each therefore counts as two units (paragraph and question). Appendix B.4 provides the exact value of "TESS units" assigned to each item. In our surveys, each repeated measure experiment was separated by between 0 and 20 TESS units of distance. This variable is highly correlated with the amount of time elapsed (measured in milliseconds) between repeated measures. Excluding outlier observations in which the elapsed time between measures is more than thrice the median completion time for the entire survey, these correlations range from 0.68 to 0.74 across samples.

[6] For this purpose, NORC bundled our AmeriSpeak survey with six questions about the NFL fielded by an uninvolved researcher. We maintained these NFL items in the non-probability samples. See B.4 for details.

[7] Figure 1 shows the pooled distances; the distributions are very similar for each sample and experiment.

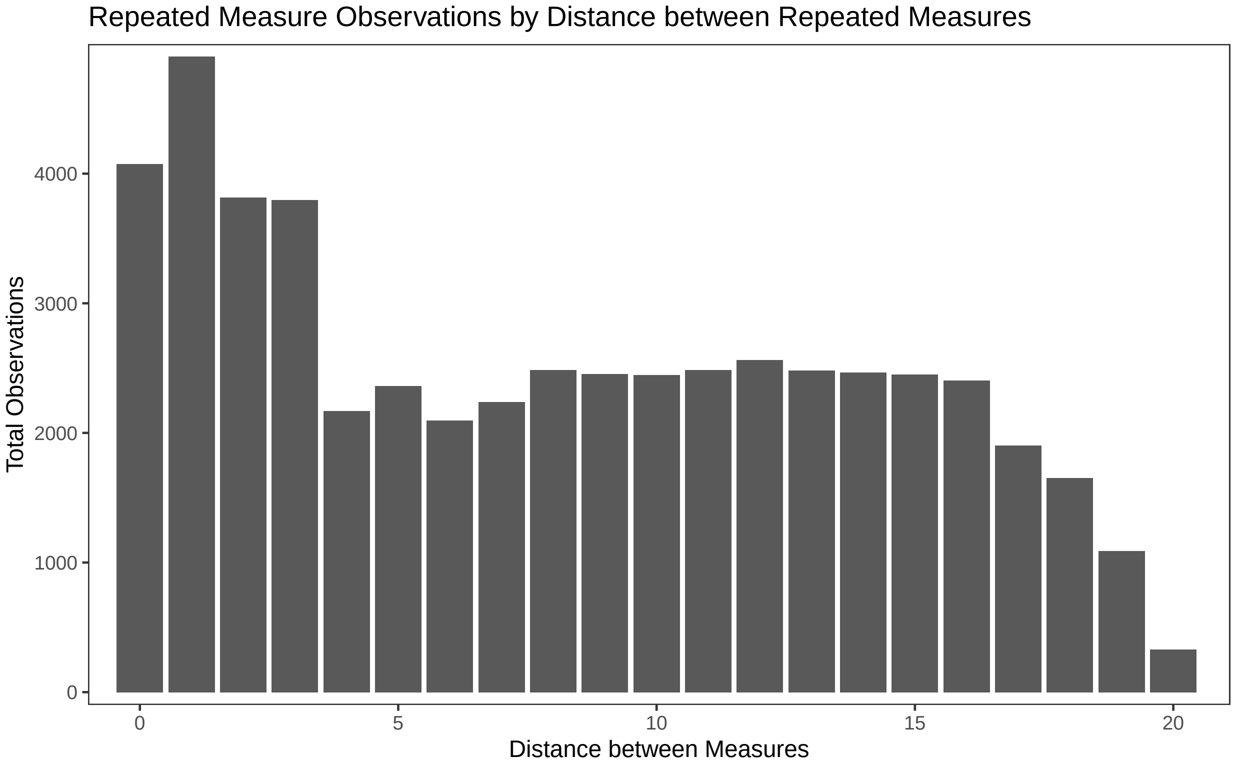

Figure 1: Histogram of distances between repeated measures. Figure shows the observed distances (counts of survey items) separating the pre- and post-treatment measures for observations in the repeated measure design setting. Data includes pooled observations from all experiments in all samples.

errors for the post-only and repeated measure designs (using bootstrapped regressions with equivalent sample sizes to account for the 2:1 oversampling of repeated measure designs). And by oversampling scenarios in which repeated measures appear in close proximity, we can test whether this proximity moderates design effects (addressing H2).

Sampling Approach

We fielded our experiments on three samples with concurrent omnibus surveys from June 27th through July 15th, 2024. Building on CSP's original studies, which drew samples from undergraduate pools or opt-in online panels, we obtained one sample from a probability-based online panel (NORC's AmeriSpeak panel) in addition to two non-probability samples recruited via quota sampling on Prolific and Lucid. These vendors are often used for political science research and offer substantial diversity in terms of respondent professionalization, respondent attentiveness, and sample representativeness on observables (Stagnaro et al. 2024). Table 2 summarizes key information for each sample; for further information, see Appendix B.

Table 2: Sample and Median Respondent Characteristics

	Survey Vendor	**Sampling Method**	**Median R: Survey Duration**	**Median R: Surveys per Month**	**Median R: Panel Memberships**
1	**AmeriSpeak** ($N_i = 4{,}029$)	Probability	6.1 min	2	1
2	**Prolific** ($N_i = 4{,}261$)	Non-probability	7.2 min	30	2
3	**Lucid** ($N_i = 4{,}869$)	Non-probability	7.3 min	15	4

Two key respondent characteristics vary across our three samples. The first is respondent professionalization, which refers to survey respondents' familiarity with and frequency of survey-taking. Most Americans take few surveys regularly; however, a small minority of Americans take many surveys frequently for income or entertainment (Hillygus, Jackson, and Young 2014). These professionalized respondents constitute an out-sized share of non-probability panels like Prolific and Lucid because high-propensity respondents can voluntarily opt into such panels and take surveys on demand. In contrast, members of probability-based panels like AmeriSpeak can only join if randomly sampled, and organizations that manage such panels invite panelists to take surveys relatively infrequently. Indeed, our AmeriSpeak respondents are much less professionalized than our Prolific and Lucid respondents in terms of the number of recent surveys taken and unique survey panel memberships (Table 2).

Professionalization may cause respondents to react differently to repeated measures, though in what direction remains uncertain. On the one hand, professionalized respondents may be inured to peculiarities of survey experiments like repeated measures, dampening design effects. Alternatively, professionalized respondents may be more likely to recognize that questions before and after experimental stimuli are testing for opinion change, heightening demand effects or consistency pressures. How respondent professionalization influences design effects is theoretically unclear and empirically untested.

The second relevant dimension is response quality, defined here as respondent attention and effort. A perennial issue in survey research is that respondents do not always pay close attention or put much effort into their responses, introducing statistical noise and possibly bias (Berinsky et al. 2021). Response quality issues are acute in self-administered surveys where there is no interviewer to induce attention and effort (e.g., Cannell, Miller, and Oksenberg 1981; Chang and Krosnick 2009; Lerner and Tetlock 1999). Because online surveys typically provide monetary incentives, some participants engage in extreme satisficing or speeding to maximize hourly earnings (Hillygus and LaChapelle 2022), and may use generative AI and other automated tools to do so (Veselovsky et al. 2023; Veselovsky, Ribeiro, and West 2023). In repeated measure designs, less attentive and effortful respondents may still be subject to issues like priming, consistency, and demand effects, but their disengagement might reduce the likelihood or strength of these biases.

To address response quality, some vendors engage in extensive panel management, such as requiring panelists to pass quality filters (e.g., consistency checks, attention checks), while other vendors leave quality control to researchers. Consequently, non-probability samples can vary considerably in respondent attention and effort; some recent evidence suggests that Lucid performs relatively poorly and Prolific performs relatively well on these metrics (Stagnaro et al. 2024). On our Prolific and Lucid surveys, we included six preregistered quality checks (see Appendix B) and drop respondents that failed at least two from our main analyses.[8] Prolific respondents failed 0.115 checks on average; this falls to 0.081 in the

[8] NORC discourages attention checks, out of concern that AmeriSpeak respondents are unaccustomed to the practice and may discontinue participation, so we preregistered a more limited procedure for dropping AmeriSpeak respondents (see Appendix B.2). We assume that retained AmeriSpeak respondents are sufficiently high quality, given the quality metrics we have (e.g., completion times) and NORC's rigorous recruitment and management for AmeriSpeak panelists.

analysis sample after we exclude 38 respondents who failed at least two (as preregistered). Lucid respondents failed an average of 0.684 checks, which falls to 0.279 after we exclude 681 who failed at least two. Lucid respondents are thus less attentive and effortful than Prolific respondents on average, variation that we exploit to test whether these characteristics affect the performance of repeated measure designs.

In summary, our study expands and advances the evidentiary basis for repeated measure designs. We replicate four studies from CSP and two additional within-subject experiments from the political science literature in each of three large samples to test whether repeated measure designs introduce design effects (i.e., attenuation or exaggeration of the ATE), totaling 18 studies with a combined $N_{ij} = 78,978$. This represents a nearly tenfold increase over CSP's pooled samples. Our large samples not only provide power to detect small design effects, but also enable us to test for potential heterogeneity in design effects across several critical design considerations: experiment type (between-groups or within-subject), proximity of repeated measures, and vendor sampling designs and consequent respondent characteristics. Our study thus provides both well-powered tests and novel insights into how various design considerations affect the utility of repeated measure experiments.

Results

We first summarize the results of each experiment under each design and report the estimated design effect. Next, we report our overall findings on the design effect of repeated measures through a series of internal meta-analyses of the 18 experiments. We then test for potential heterogeneity in design effects along several key dimensions.

Summary of Experimental Results

For each experiment, we report the observed ATE for both post-only and repeated measures designs. To facilitate comparison across experiments, we rescale all outcome variables to range from 0 (most opposed) and 1 (most supportive). For the between-groups experiments (Studies 1–3) we compare the difference in ATEs by estimating separate ordinary least squares (OLS) regressions for each design. These regressions model the post-treatment outcome as a function of a binary treatment indicator,[9] with the pre-treatment outcome included as a covariate in the repeated measures design.[10] We then combine these regressions via seemingly unrelated regression estimation, which allows us to conduct a linear combination test for equivalence of ATEs across the two designs.

For the within-subject experiments (Studies 4–6), we compare the difference in ATEs using random effects models. These models regress the dependent variable on an indicator for treatment interacted with an indicator for repeated measure assignment. This approach explicitly acknowledges the nested nature of the data by clustering standard errors at the respondent level (as some respondents contribute two observations), capturing individual-level variation in the outcome (the "random" effects) that is unrelated to the explanatory variables. The coefficient on the interaction term estimates the difference in ATEs between the designs.

As preregistered, we follow prior authors' inclusion of specific pre-treatment covariates (e.g., partisanship, ideology) in the model for each experiment, as noted below. We report the results of each experiment separately for each of the three samples (AmeriSpeak, Prolific, and Lucid). A summary of the results is provided in Table 3, which we briefly detail below.[11]

Study 1: Foreign Aid

In this between-groups experiment, we regress support for foreign aid spending on a treatment indicator for receiving information that foreign aid spending is about 1% of the federal budget. Following CSP, we include partisanship and ideology as covariates. All three samples replicate CSP's finding (and that of Gilens 2001) that the informational treatment

[9] For Study 2, we interact the treatment indicator with an indicator for Democratic party identification. The coefficient of interest is on the interaction term. We exclude respondents who do not lean toward either party from this analysis.

[10] Blair et al. (2019) show that, relative to using difference scores as the outcome, adjusting for the pre-treatment measure reduces noise when adjustment bias is minimal, as in randomized experiments (Lin 2013).

[11] Note that observations for within-subject repeated measure models reflect two observations per repeated measures respondent and one observation per post-only respondent (less non-response).

Table 3: Summary of Experimental Results

		Post-Only		*Repeated*		*Design Effect*		
Experiment	Sample	Est. ATE	Obs.	Est. ATE	Obs.	Estimate	SE	Δ in ATE
Fgn. Aid	AmeriSpeak	0.089^{***}	1,272	0.065^{***}	2,662	-0.023	0.015	-26.1%
	Prolific	0.111^{***}	1,433	0.068^{***}	2,828	-0.043^{**}	0.014	-38.6%
	Lucid	0.063^{***}	1,616	0.056^{***}	3,252	-0.008	0.016	-12.0%
Drug Imp.	AmeriSpeak	0.125^{***}	1,360	0.055^{***}	2,580	-0.070^{*}	0.029	-56.0%
	Prolific	0.096^{***}	1,261	0.072^{***}	2,485	-0.024	0.032	-25.1%
	Lucid	0.110^{***}	1,375	0.077^{***}	2,699	-0.033	0.032	-30.4%
GMOs	AmeriSpeak	0.162^{***}	1,283	0.129^{***}	2,660	-0.033^{*}	0.017	-20.2%
	Prolific	0.180^{***}	1,396	0.162^{***}	2,865	-0.017	0.016	-9.7%
	Lucid	0.144^{***}	1,650	0.124^{***}	3,218	-0.021	0.016	-14.2%
Anti-pov.	AmeriSpeak	0.202^{***}	1,351	0.159^{***}	5,172	-0.044^{*}	0.020	-21.3%
	Prolific	0.165^{***}	1,481	0.110^{***}	5,560	-0.055^{***}	0.017	-33.1%
	Lucid	0.169^{***}	1,637	0.135^{***}	6,457	$-0.033^{\dagger}$	0.019	-20.0%
Afrm. Act.	AmeriSpeak	0.095^{***}	1,310	0.079^{***}	5,430	-0.017	0.022	-17.3%
	Prolific	0.094^{***}	1,420	0.053^{***}	5,682	$-0.040^{\dagger}$	0.022	-43.2%
	Lucid	0.094^{***}	1,593	0.079^{***}	6,546	-0.014	0.020	-15.0%
Clinic	AmeriSpeak	0.113^{***}	1,361	0.101^{***}	5,328	-0.012	0.018	-10.8%
	Prolific	0.071^{***}	1,361	0.126^{***}	5,800	0.055^{**}	0.019	$+76.5\%$
	Lucid	0.047^{**}	1,624	0.050^{***}	6,489	0.004	0.016	$+7.6\%$

†p<0.10; *p<0.05; **p<0.01; ***p<0.001

Note: Table displays the estimated ATE under each design in each experiment in each sample, followed by the repeated measure design's estimated design effect and percentage change from the ATE of the post-only design.

increases support for foreign aid in both the post-only and repeated measure designs. As with CSP's study, we find that the repeated measure design attenuates this treatment effect in the Prolific sample ($p = 0.002$). The design effects are negative but not significant in the AmeriSpeak ($p = 0.127$) and Lucid ($p = 0.630$) samples.

Study 2: Prescription Drug Imports

In this between-groups experiment, we regress support for prescription drug imports on a treatment indicator for receiving a signal about typical party positions, interacted with an indicator for Democratic versus Republican partisanship (excluding pure independents). The interaction coefficient provides a measure of polarization in attitudes between parties. All three samples replicate CSP's finding that party cues increase attitudinal polarization in both the post-only and repeated measure designs. We find an attenuation of this treatment effect in the repeated measure design in the AmeriSpeak sample ($p = 0.017$); the estimated design effects are negative but not significant in the Prolific ($p = 0.450$) and Lucid ($p = 0.301$) samples.

Study 3: GMOs

In this between-groups experiment, we regress support for GMOs on an indicator for receiving a pro-GMO frame (versus an anti-GMO frame). Following CSP, we include partisanship and ideology as covariates. All three samples replicate CSP's finding that the positive frame increases support for GMOs in both the post-only and repeated measure designs. We find an attenuation of this treatment effect in the repeated measure design in the AmeriSpeak sample ($p = 0.049$). The design effects are negative but not significant in the Prolific ($p = 0.273$) and Lucid ($p = 0.210$) samples.

Study 4: Anti-poverty

In this within-subject experiment, we regress support for anti-poverty spending on an indicator for whether these efforts are described as "assistance to the poor" (1) versus "welfare" (0), interacted with an indicator for the two-question repeated measures condition (1) versus the single-question post-only condition (0). Following CSP, we include partisanship and ideology as covariates. All three samples replicate CSP's finding (and that of Smith 1987) that support for anti-poverty spending is greater when characterized as "assistance to the poor," in both the post-only and repeated measure designs. We find attenuated treatment effects in the repeated measure designs the AmeriSpeak ($p = 0.025$), the Prolific sample ($p = 0.001$), and (at the 0.10 level) the Lucid sample ($p = 0.071$).

Study 5: Affirmative Action

In this within-subject experiment, we regress support for affirmative action on an indicator for whether the policies apply to women (1) or racial minorities (0), interacted with an indicator for assignment to the repeated measures (1) versus post-only condition (0). All three samples replicate the finding of Wilson et al. (2008) that support is greater for affirmative action for women relative to racial minorities, in both the post-only and repeated measure designs. We find an attenuated treatment effect at the 0.10 level (consistent with the original study's claim that asking both items induces consistency; see Wilson et al. 2008) in the Prolific sample ($p = 0.071$). The estimated design effects are negative but not significant in the AmeriSpeak ($p = 0.453$) and Lucid ($p = 0.483$) samples.

Study 6: Opioid Clinic

In this within-subject experiment, we regress support for opening a new methadone clinic on an indicator for whether the proposed clinic is located a quarter mile away (1) versus two miles away (0), interacted with an indicator for assignment to the repeated measures (1) versus post-only condition (0). All three samples replicate the finding of de Benedictis-Kessner and Hankinson (2019) that support is greater when the proposed clinic is located further away, in both the post-only and repeated measure designs. In contrast to the other five studies, we find a significant *positive* design effect (exaggerating the treatment effect) from the repeated measure design in the Prolific sample ($p = 0.004$), but no significant design effects in the AmeriSpeak (negative estimate, $p = 0.481$) and Lucid (positive estimate, $p = 0.821$) samples.

Repeated Measure Designs Cause (Slight) Attenuation of Treatment Effect Estimates

Across six experiments replicated thrice each in large samples, nearly every estimated design effect is negative. These design effects are often large, as shown in the final column of Table 3. We observe a median 20.1 percent reduction in the ATE from repeated measure designs relative to post-only designs across the 18 experiments. This consistent pattern suggests repeated measure designs attenuate treatment effects, but that the design effects are too small to reliably detect in individual experiments. We therefore conduct

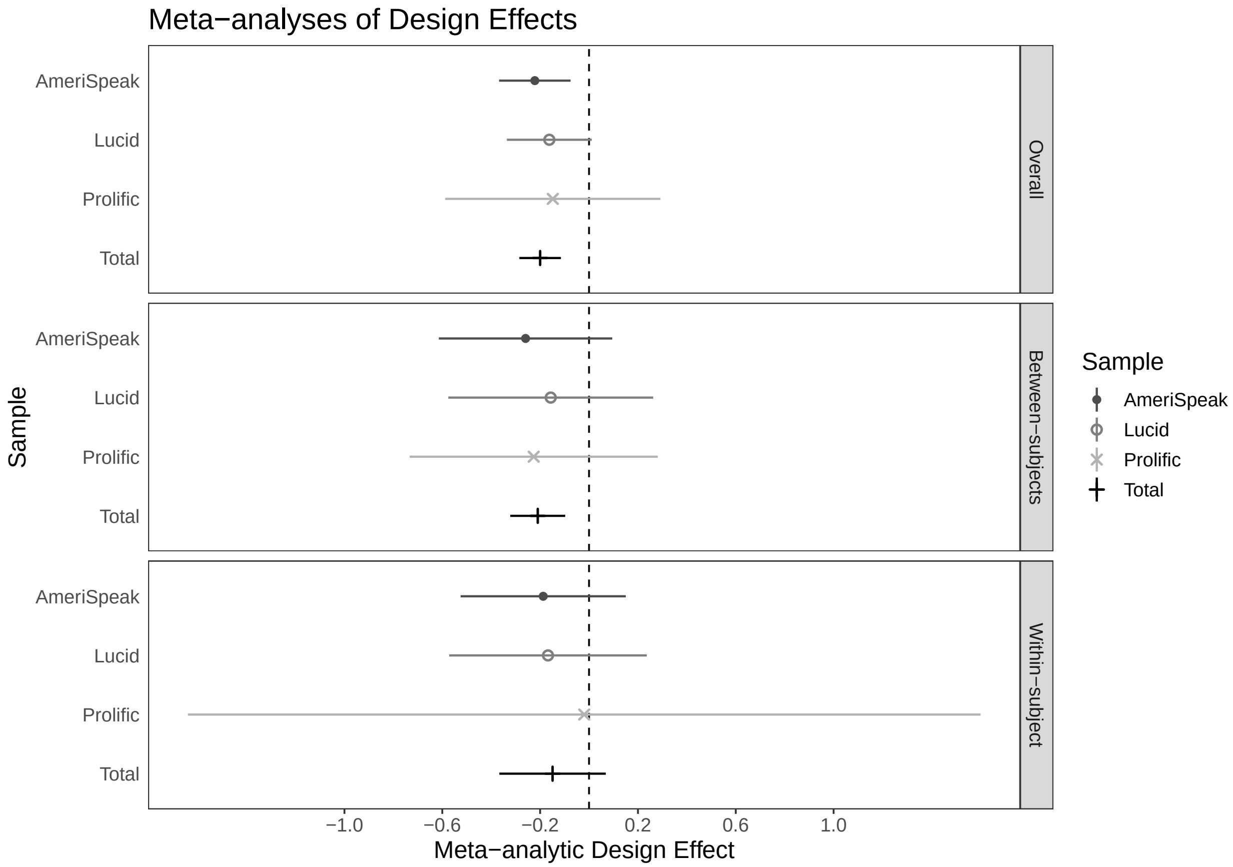

Figure 2: Internal Meta-Analyses. Figure displays estimated design effects from internal meta-analyses of experiments within each sample and across all three samples.

preregistered internal meta-analyses, rescaling the design effect and standard error in each experiment as proportional change from the post-only design's ATE (that is, a 20.1 percent attenuation is a design effect of -0.201).[12] We then meta-analyze all six experiments, the three between-groups experiments, and the three within-subject experiments, each set both within and across samples. The results are shown in Figure 2 and provided in tabular form in Appendix A.1.

When analyzing all six experiments together, as shown in the top panel of Figure 2, we find a meta-analytic design effect of -0.222 in the AmeriSpeak sample ($p = 0.011$), -0.162 in the Lucid sample ($p = 0.061$), and -0.148 in the Prolific sample ($p = 0.426$). Meta-analyzing all 18 experiments, we find a precisely estimated meta-analytic design effect of

[12] This is similar to the approach taken by Sheagley and Clifford (2025) and is primarily intended to ease interpretation of the resulting analysis.

-0.200 ($p < 0.001$, 95 percent CI = $[-0.285, -0.115]$). That is, the estimated attenuation of the ATE when using repeated measure designs is 20.0 percent on average.

This typical attenuation effect is most consistent for between-groups experiments in our data. While we do not find find a statistically significant design effect for either type of experiment in any one sample,[13] the between-groups estimate across samples is statistically significant (estimate -0.210, $p = 0.003$). The within-subject estimate across samples is smaller and not statistically significant (estimate -0.149, $p = 0.153$). This is due to a clear outlier in the Prolific sample, in which we observe a large positive design effect in the opioid clinic experiment. A meta-analysis of the within-subject experiments across samples but excluding this outlier is quite similar to the between-groups experiments (design effect estimate -0.227, $p = 0.003$).

Repeated Measure Designs Increase Statistical Power

Although we find evidence of treatment effect attenuation in repeated measure designs, these designs may still be preferable due to large precision gains. Since our experimental design assigns respondents to complete twice as many repeated measure experiments as post-only experiments, directly comparing standard errors would artificially privilege the precision of the repeated measure design, due simply to differential assignment. To address this, we re-estimate each ATE via a bootstrapping procedure that uses samples of identical size for both designs. Specifically, for each experiment in each sample, we estimate the respective models for the post-only and repeated measure designs 1,000 times, each time substituting a randomly drawn sample (with replacement) equal to the maximum number of unique observations in the post-only setting for that experiment in that sample. From these 1,000 estimated models, we then calculate pooled standard errors using Rubin's rule. In effect, this procedure estimates the relative precision across experimental designs for samples of equal size. Table 4 shows the bootstrapped ATE and standard errors under each design, as

[13] The meta-analysis of between-groups experiments in the AmeriSpeak sample is the slight exception here, which detects a design effect significant at the 0.10 level (estimate -0.259, $p = 0.088$).

well as the percentage change in standard error and root mean squared error (RMSE) that the repeated measure design provides.

As Table 4 shows, we find that repeated measure designs provide large gains to precision. We observe a median 49.4 percent reduction in standard error across all 18 experiments, with a minimum reduction of 31.1 percent. Similarly, we observe a median 41.0 percent reduction in the RMSE, with a minimum reduction of 28.6 percent. The consistently large reductions in standard errors confirm that repeated measure designs offer significant improvement in statistical precision. As we document in the Discussion, these major precision gains can often outweigh the disadvantage of slight attenuation to provide improved accuracy in many research settings.

Minimal Moderation by Distance Between Repeated Measures

Researchers regularly place pre- and post-treatment measures at opposite ends of a survey—or even on separate waves in panel surveys—to minimize the probability that respondents will recall being previously asked the same question and alter their post-treatment response. Given our finding that repeated measure designs slightly attenuate treatment effects, a reasonable concern is that a short survey or module might induce greater bias in the estimated ATE due to the proximity of the repeated measures. Our experimental design randomly varies the distance between pre- and post-treatment measures, allowing us to test how distance impacts design effects. We estimate a series of ATEs at each discrete distance between pre- and post-treatment measures (in counts of survey items, ranging from 0 to 19) for each experiment in each sample and standardize these ATEs relative to the post-only ATE observed for each study (see the third column of Table 3).[14] We then regress the standardized ATEs on the distance variable. Figure 3 shows the standardized ATEs and associated 95 percent confidence intervals for each experiment at each degree of separation.

[14] Although we observe distances up to 20 items between repeated measures, we have very few observations at the largest possible distance for each experiment (between 4 and 52 observations per experiment) and therefore exclude these imprecise estimates from our analysis.

Table 4: Bootstrapped Experimental Results

		Post-Only			*Repeated Measure*			*Increased Precision*	
Experiment	Sample	Est. ATE	Std. Err.	RMSE	Est. ATE	Std. Err.	RMSE	Δ in SE	Δ in RMSE
Foreign Aid	AmeriSpeak	0.089***	0.014	0.250	0.065***	0.008	0.147	−41.0%	−41.3%
	Prolific	0.111***	0.013	0.248	0.068***	0.007	0.136	−43.4%	−45.2%
	Lucid	0.063***	0.014	0.284	0.056***	0.010	0.200	−31.1%	−29.6%
Drug Imports	AmeriSpeak	0.125***	0.027	0.252	0.055***	0.016	0.148	−41.4%	−41.4%
	Prolific	0.096***	0.030	0.262	0.072***	0.015	0.125	−47.7%	−52.3%
	Lucid	0.110***	0.029	0.269	0.077***	0.020	0.179	−33.5%	−33.3%
GMOs	AmeriSpeak	0.162***	0.016	0.272	0.129***	0.009	0.175	−39.7%	−35.8%
	Prolific	0.180***	0.015	0.275	0.162***	0.008	0.162	−43.8%	−41.0%
	Lucid	0.144***	0.015	0.297	0.124***	0.010	0.212	−31.5%	−28.6%
Anti-poverty	AmeriSpeak	0.202***	0.018	0.343	0.159***	0.009	0.231	−51.0%	−32.6%
	Prolific	0.165***	0.016	0.314	0.110***	0.007	0.194	−55.6%	−38.1%
	Lucid	0.169***	0.018	0.354	0.135***	0.008	0.241	−53.1%	−31.9%
Affirm. Action	AmeriSpeak	0.095***	0.020	0.378	0.079***	0.009	0.216	−57.9%	−42.9%
	Prolific	0.094***	0.023	0.409	0.053***	0.007	0.191	−68.5%	−53.2%
	Lucid	0.094***	0.019	0.383	0.079***	0.008	0.244	−56.2%	−36.1%
Opioid Clinic	AmeriSpeak	0.113***	0.018	0.331	0.101***	0.006	0.164	−66.4%	−50.4%
	Prolific	0.071***	0.018	0.338	0.126***	0.006	0.168	−65.3%	−50.4%
	Lucid	0.047**	0.016	0.309	0.050***	0.006	0.171	−59.5%	−44.7%

$^{\dagger}p<0.10$; $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Note: Table displays the estimated ATE and bootstrapped standard error under each design in each experiment in each sample, each estimated 1,000 times with a number of sampled respondents equal to the number of observations in the respective post-only design (see column four of Table 3).

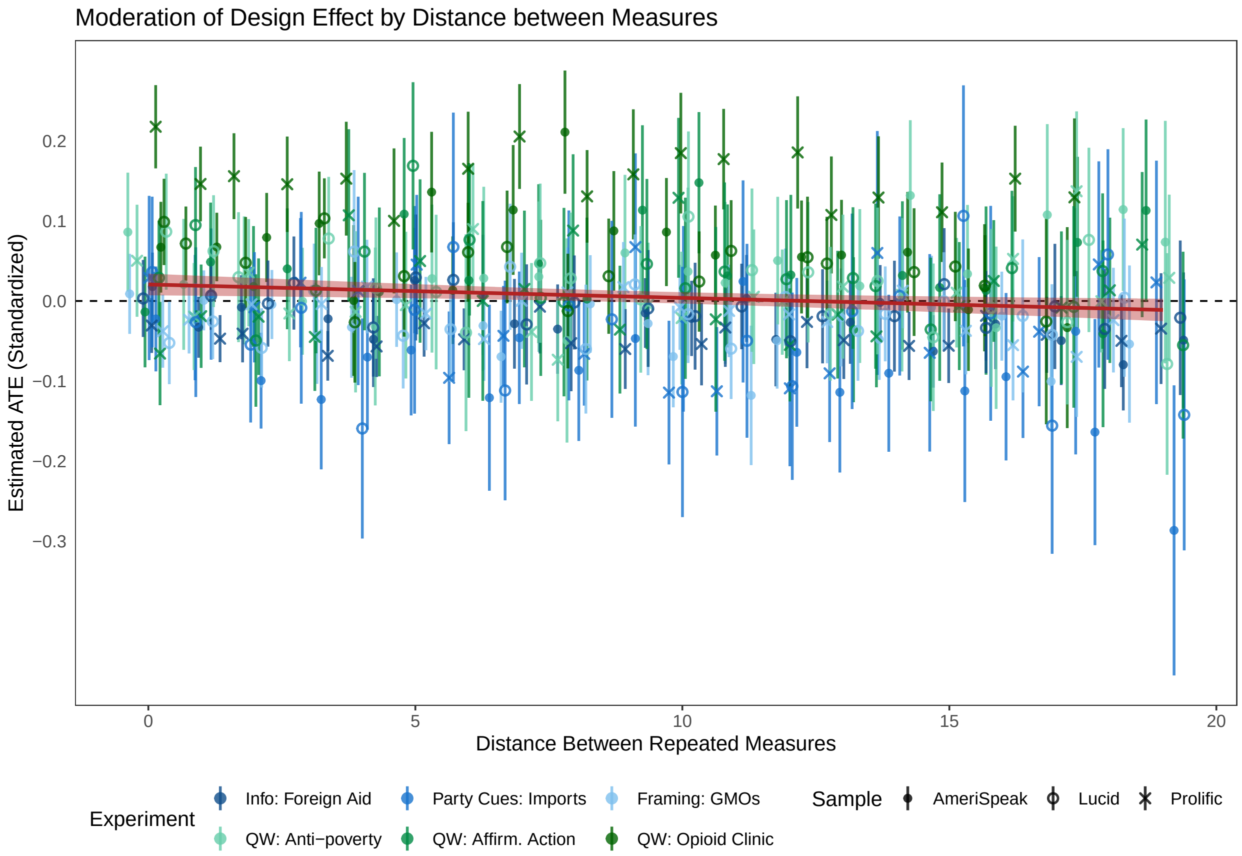

Figure 3: Estimated design effect by distance between repeated measures. Figure displays the estimated ATE at each distance between pre- and post-treatment measures (in counts of survey items, x-jittered for visual clarity) in each experiment in each sample, standardized to the respective observed post-only ATE. The red line indicates the fitted values from a linear regression on these ATE point estimates on distance; the shaded areas indicate 95 percent confidence intervals.

The red line indicates the predicted values from this linear regression, and the shaded areas show the corresponding 95 percent confidence interval.[15]

We find that the effect of distance between repeated measures is detectable but substantively small, as the regression line in Figure 3 suggests. Each additional item separating the pre- and post-treatment measures is estimated to attenuate the repeated measure ATE

[15] These analyses deviate from our preanalysis plan of using spline regressions to assess non-linear effects of distance on the design effect. We estimated spline regressions (interacting the treatment indicator with indicator variables for each discrete TESS distance observed) for each experiment in each sample, and found few significant interactions at all and no consistent pattern across the studies—that is, no clear evidence of non-linearity, as the point estimates in Figure 3 also suggest. We therefore opted for this alternate analysis for ease of presentation and interpretation.

by -0.002 ($p = 0.010$) on average, or by about 1.4 percent of the mean ATE in our data. Including fixed effects for the sample and experiment gives a similar but more precise result: the estimated attenuation in the expected ATE is -0.001 ($p = 0.009$) on average, or about 1.2 percent of the mean ATE in our data.[16] The slight influence of distance on the overall design effect suggests that repeated measure designs are about as well suited to close placement as to separating the measures by several minutes.

Some Moderation by Repeated Exposures to Repeated Measure Designs

An important feature of our design that distinguishes our approach from that of Clifford, Sheagley, and Piston (2021) is the inclusion of four repeated measure experiments in a single survey administered to each respondent. We also include several questions that ask respondents to recall their previous attitude on the pre-treatment measure and assess how much their opinion had changed—an unusual question that may in itself exacerbate consistency or demand pressures on subsequent repeated measure experiments. These features of our implementation may alter the design effect as respondents progress through the survey, in a way that unrelated distractor content would not.

We conduct an exploratory analysis to test whether design effects differ between respondents' first to fourth exposures to repeated measure designs. We estimate a meta-analytic design effect for (only) the first repeated measure experiment each participant encounters in their individual survey experience (pooling across experiments and samples, $k = 18$), and then estimate separate meta-analytic design effects for the second, third, and fourth repeated measure experiment encountered. These results are shown in Table 5. We find that treatment effect attenuation increases as the respondent encounters more repeated measure experiments in the survey, from a meta-analytic mean of -0.131 ($p = 0.018$) in the first exposure to -0.266 ($p < 0.001$) in the fourth exposure. The difference between these two design effect estimates is statistically significant, as shown in a fixed-effect model specification

[16] This estimated effect is sufficiently small that it may best be considered negligible (Rainey 2014).

reported in Appendix Table A.1.2 ($p = 0.049$), suggesting that fielding multiple repeated measure experiments in a single survey may exacerbate the design effect for experiments later in the survey. Notably, however, focusing strictly on design effects in the first repeated measures experiment respondents encountered—equivalent to CSP's approach of separate surveys—we still find significant attenuation.

Table 5: Repeated Measure Results by Order of Repeated Measure Design Encountered

Repeated Measure Experiment Encountered	Design Effect Estimate	Std. Error	95% CI	p-value
First (N = 38,628)	-0.131^{*}	0.050	$[-0.236, -0.025]$	0.018
Second (N = 38,664)	-0.206^{**}	0.065	$[-0.344, -0.068]$	0.006
Third (N = 38,673)	-0.183^{**}	0.048	$[-0.285, -0.082]$	0.001
Fourth (N = 38,665)	-0.266^{***}	0.041	$[-0.353, -0.180]$	< 0.001

†p<0.10; *p<0.05; **p<0.01; ***p<0.001

Note: Table displays the results of internal meta-analyses ($k = 18$ for each) of the repeated measure design effect, subset by the order of repeated measure experiments in the survey for each individual respondent. The total number of respondents included in each meta-analysis is provided in parentheses.

Notably, much of this increase appears to be a consequence of the attitude-recall questions in our surveys. In Appendix Table A.1.3, we estimate separate meta-analytic design effects for repeated measure experiments later in the survey flow (that is, the second, third, or fourth repeated measure experiment) but *prior* to any attitude-recall questions, versus those that appeared following an attitude-recall question. Later repeated measure experiments that appeared before any attitude-recall questions exhibit a similar design effect size (estimate -0.145, $p = 0.017$) to the earliest repeated measure experiment for each respondent (estimate -0.131, $p = 0.018$). In contrast, repeated measure experiments with post-treatment measurement after at least one attitude-recall question (i.e., on an earlier experiment) show a larger design effect (estimate -0.249, $p < 0.001$). We thus identify a small but meaningful design effect of singular repeated measure experiments, but fielding multiple repeated measure experiments in a single survey or (especially) using attitude-recall questions may exacerbate the design effect on later repeated measure experiments.

No Moderation by Respondent Professionalization

Non-probability samples are commonly used for experiments because of their convenience and relatively low costs (Jerit and Barabas 2023). Many online non-probability panelists take surveys frequently for income or enjoyment, making them professionalized and prone to satisficing (Hillygus, Jackson, and Young 2014; Hillygus and LaChapelle 2022). In contrast, NORC's probability-based AmeriSpeak panel restricts participation frequency to maintain respondent quality and avoid excessive professionalization, meaning that respondent attention may be higher in this sample.

Differences between probability and non-probability panels could affect the design effect of repeated measure designs. Greater respondent attention could increase recall of a pre-treatment measure or response, potentially elevating priming, consistency, or demand pressures on post-treatment responses. Respondent satisficing and speeding could reduce recall of a pre-treatment measure (and possibly exposure to treatment; see Hillygus, Jackson, and Young 2014) to potentially suppress these effects. Higher professionalization may inure respondents to repeated measures, improving their effectiveness. Conversely, professionalization may enable respondents to connect repeated questions to experimental stimuli, exacerbating consistency pressures or altering demand effects.

While we find no significant sample-level differences in design effects (as Figure 2 shows), our measures of respondent professionalization allow us to conduct exploratory analyses of how within-sample variation in respondent characteristics impacts the design effect of repeated measures. At the end of each survey, we asked respondents how many other online surveys they had completed in the past 30 days, as well as how many online survey companies they had completed surveys for in the past 30 days (active panel memberships). As expected, our Prolific and Lucid respondents are much more professionalized than the AmeriSpeak panelists: the median AmeriSpeak respondent reported completing just 2 surveys for 1 panel in the past 30 days, whereas the median Prolific and Lucid respondent reported completing

40 surveys for 2 panels and 17 surveys for 4 panels, respectively.[17] Within each sample, we then split respondents at the median on each dimension of professionalization, re-analyze each experiment using the subsample for each group, and the meta-analyze the estimated design effects (reported in Appendix A.2).

As shown in Appendix Figure A.2.1, the design effects are similar above and below the medians of each professionalization measure. That is, our data suggests that respondent professionalization does not substantially exacerbate or mitigate the design effect of repeated measures. This result, using individual-level professionalization measures, helps explain why the design effects are similar across our three samples despite large differences in respondent professionalization, and comports with recent scholarship suggesting that non-probability samples are suitable for experimental research (Jerit and Barabas 2023; Coppock, Leeper, and Mullinix 2018).

Respondent Attention and Perceived Attitude Change

Another way to assess the impact of respondent attention in repeated measures designs is to analyze how well respondents can recall their previous (pre-treatment) attitude after exposure to treatment. In their pre-post study on GMOs, CSP asked whether respondents' support for GMOs had changed since earlier in the survey—that is, since the pre-treatment measurement. CSP found that 40.5 percent of respondents provided different answers on the two measures and that, of these, 58.8 percent (mis)-reported that their attitudes had remained stable. CSP concluded that respondents may struggle to provide consistent responses in repeated measures experiments, even if feeling pressure to do so, simply because many cannot recall their earlier responses. This, CSP argue, reduces the risk of design effects in repeated measure experiments.[18]

For our three between-groups experiments, we followed the post-treatment measure with

[17] For these analyses, we preregistered excluding respondents who reported completing more than 1,000 surveys in the past 30 days or over 100 active panel memberships, as these responses are likely not genuine.

[18] We suggest that failure to recognize attitude change may also indicate a ceiling on design effects from priming and even demand pressures, as it signals fuzziness about the pre-treatment question itself.

a similar recall question for respondents assigned to the repeated measure condition (total $n_{ij} = 26,333$ across all samples, offering an analysis sample 27 times larger than CSP's previous single study).[19] Specifically, we asked whether the respondent's preferences about the relevant issue had changed since being asked it earlier in the survey; respondents could indicate whether their support had decreased, increased, or stayed about the same.[20] We distinguish respondents into the three groups based on observed pre-post change (less supportive, no change, or more supportive) and likewise group them by self-reported perceived change (less supportive, about the same, or more supportive).

We report the rates of perceived versus observed change in Appendix A.3. Like CSP, we find that few respondents who provided different responses between pre- and post-treatment measures were able to correctly identify that change (39.1 percent of observations). However, in an exploratory analysis described in Appendix A.3, we find no evidence that respondents' ability to accurately perceive (or accurately self-report) their direction of change has any impact on the magnitude of the design effect. This analysis suggests that prior attitude recall may not be the most critical factor in producing the slight average attenuation bias we find in repeated measure designs.

Discussion

Our study provides critical new evidence on the merits of repeated measure designs for experimental research. Like CSP's landmark studies, we find that repeated measure designs consistently offer enormous improvements in statistical precision over traditional post-only designs, observing a 49.4 median reduction in the ATE's standard errors across 18 studies. Unlike CSP, however, we find a small but consistent design effect in the repeated measure setting, observing a median 20.1 percent attenuation of the ATE relative to the post-only design. Figure 4 summarizes the balance of our evidence on this fundamental

[19] Because the within-subject experiments ask about plausibly different quantities (or at least quantities perceived to be different) in the two measurements, attitude change is not conceptually well-grounded in that setting.

[20] See Appendix B.4 for exact language.

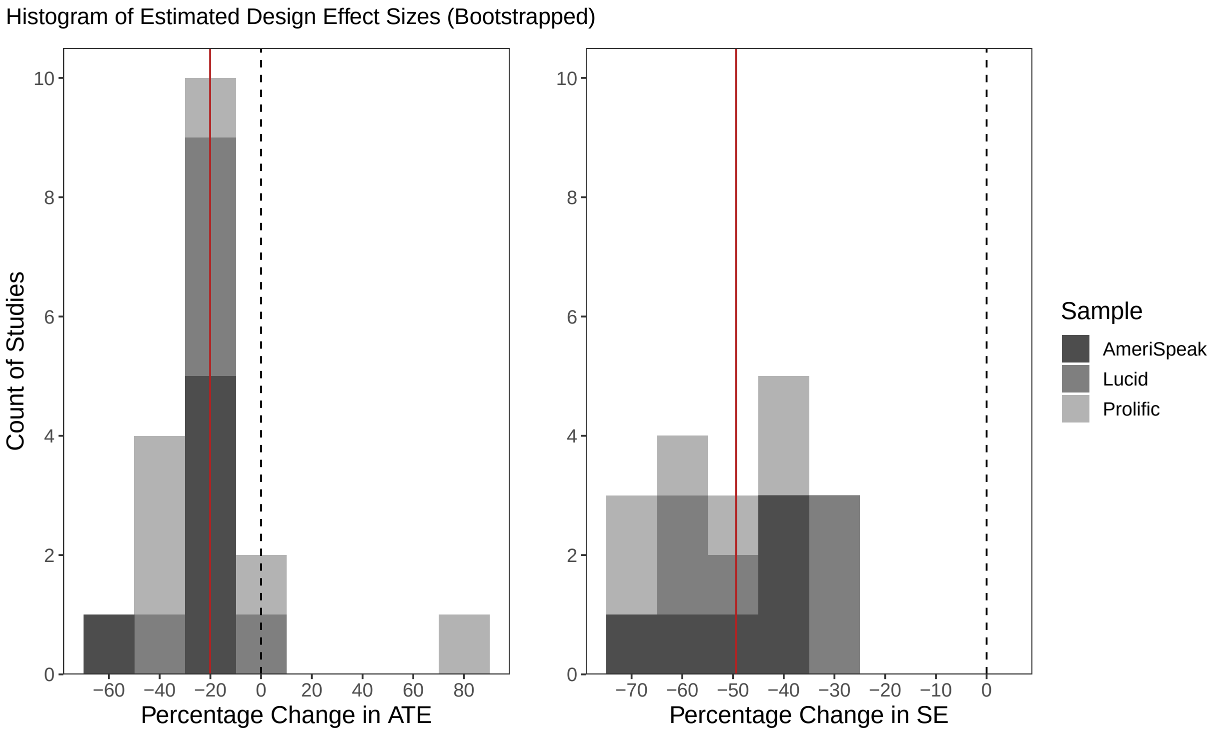

Figure 4: Histogram of design effects. Figure displays a histogram of observed design effects in terms of percentage change in estimated ATE (left panel) and standard error (right panel) in bootstrapped models with equal sample size across designs. The solid red line in each panel indicates the median percentage change in each statistic across all 18 experiments.

trade-off. Based on these findings, we provide practical recommendations for researchers as they consider whether and how to implement repeated measure designs.

Repeated Measure Designs versus Post-only Designs

Our experiments provide evidence that repeated measure designs consistently attenuate treatment effects, which may lead some readers to conclude that post-only designs should be preferred to repeated measure designs. However, many survey experiments primarily aim to identify whether a given treatment shifts the outcome in a hypothesized direction (versus a null effect). For this purpose, statistical power is an essential consideration—and repeated measure designs clearly dominate in this regard, as Figure 4 shows. Power is determined by the ratio of the treatment effect and its standard error (Rainey 2025); thus, although repeated measure designs slightly attenuate the treatment effect (the numerator), the resulting slight loss of power is often outweighed by power gained from shrinking the standard error (the denominator). In many applied settings, this power trade-off favors repeated measure designs, which are better suited to reliably detect true (albeit attenuated) treatment effects.

Should researchers who seek to estimate a treatment effect's precise magnitude (not simply its presence or direction) prefer post-only designs, given our findings of attenuated ATEs? Even here, we contend that repeated measure designs are often superior. Although post-only designs are unbiased in expectation, their imprecision in finite samples will cause estimates to vary widely around the true ATE. Our results suggest that repeated measure designs are so much more precise that even their slightly attenuated ATE estimates will fall closer to the true ATE in many circumstances.

To illustrate these tradeoffs, we simulate 1,000 two-arm (treatment and control) experiments per design at each of several sample sizes, true ATEs (expressed as Cohen's d), and true attenuation in the estimated ATE from the repeated measure design effect.[21] From these simulated experiments, we estimate the statistical power of the test (defined as one minus the observed proportion of false negatives, shown in Figure 5) and the mean absolute error in the estimated ATE (versus the true ATE) under each design (Figure 6). Each figure shows how the respective statistic changes as the sample size increases (within-panel), the true effect size increases (across columns), and the design effect attenuation of the estimated ATE increases (across rows). The means for each statistic is shown by the solid grey line for post-only experiments and by the dashed black line for repeated measure experiments. Figure 5 shows that repeated measure designs always offer superior power for hypothesis testing, with the largest gains when the true ATE or the sample size is smaller; the magnitude of the design effect has a comparatively small impact on power. For smaller samples

[21] Simulation details are provided in Appendix A.4. We assume a correlation between repeated measures of 0.8 and a corresponding reduction in the ATE standard error of 0.4, which is slightly conservative relative to the reductions we observe in our actual experiments.

and smaller true effect sizes, Figure 6 also shows that the repeated measure design provides a more accurate estimate of the true ATE than a post-only design, despite attenuation of the ATE from the design effect. Only when the sample size, true effect size, and true attenuation are large does the post-only design outperform repeated measures in terms of fidelity in expectation to the true ATE.[22]

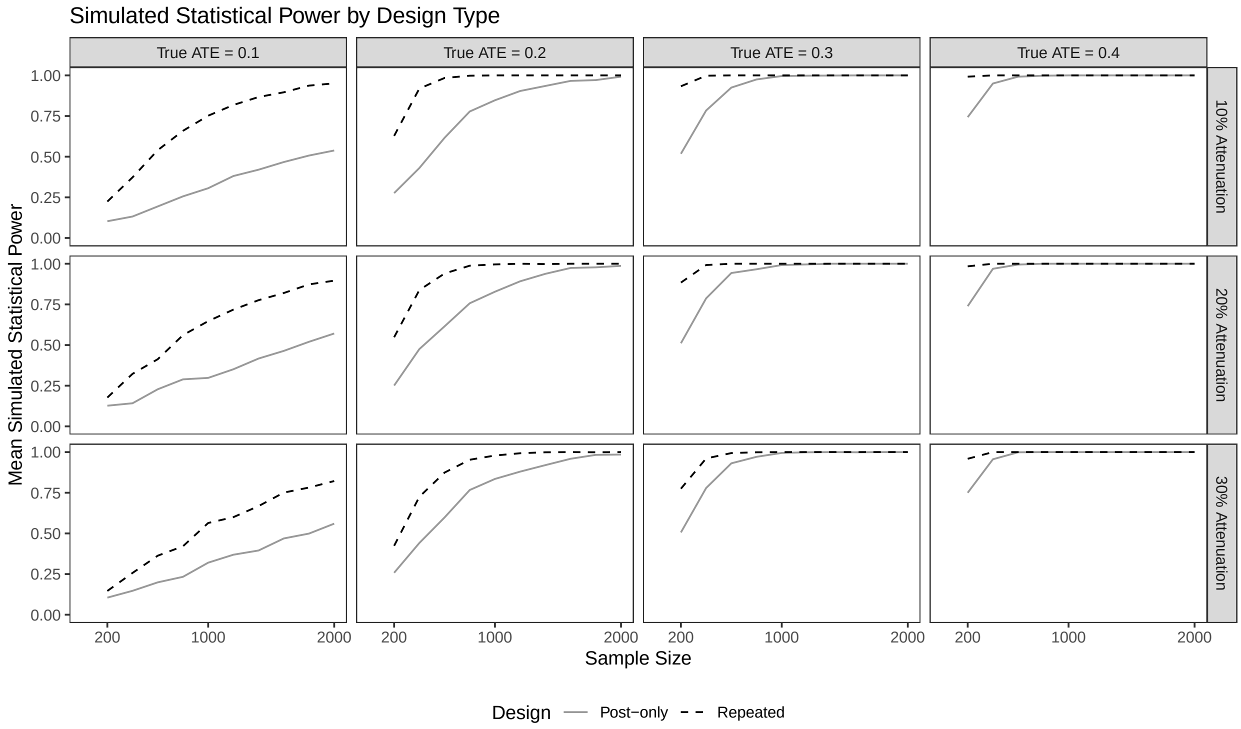

Figure 5: Statistical power in simulated experiments. Figure displays mean statistical power in simulated experiments at varying sample sizes, true effect sizes (Cohen's d), and true design effect (attenuation) of a repeated measure experiment.

Post-only designs may therefore still be preferable for researchers with access to a large sample and reason to expect (a priori) a strong treatment effect or strong design effects. Although a repeated measure design will usually offer an improvement in statistical power, this benefit declines as sample and true effect size increase (see Figure 5). In these circumstances, the expected attenuation of the ATE can be large enough relative to the precision gains to

[22] Appendix A.4 also examines the false discovery rate and coverage rate for the true ATE. When the true ATE is zero, the two designs perform similarly on both metrics. When the true ATE is non-zero, repeated measure designs have a lower false discovery rate for smaller samples and true effect sizes, while the coverage rate drops as the sample, true ATE, and degree of attenuation increase.

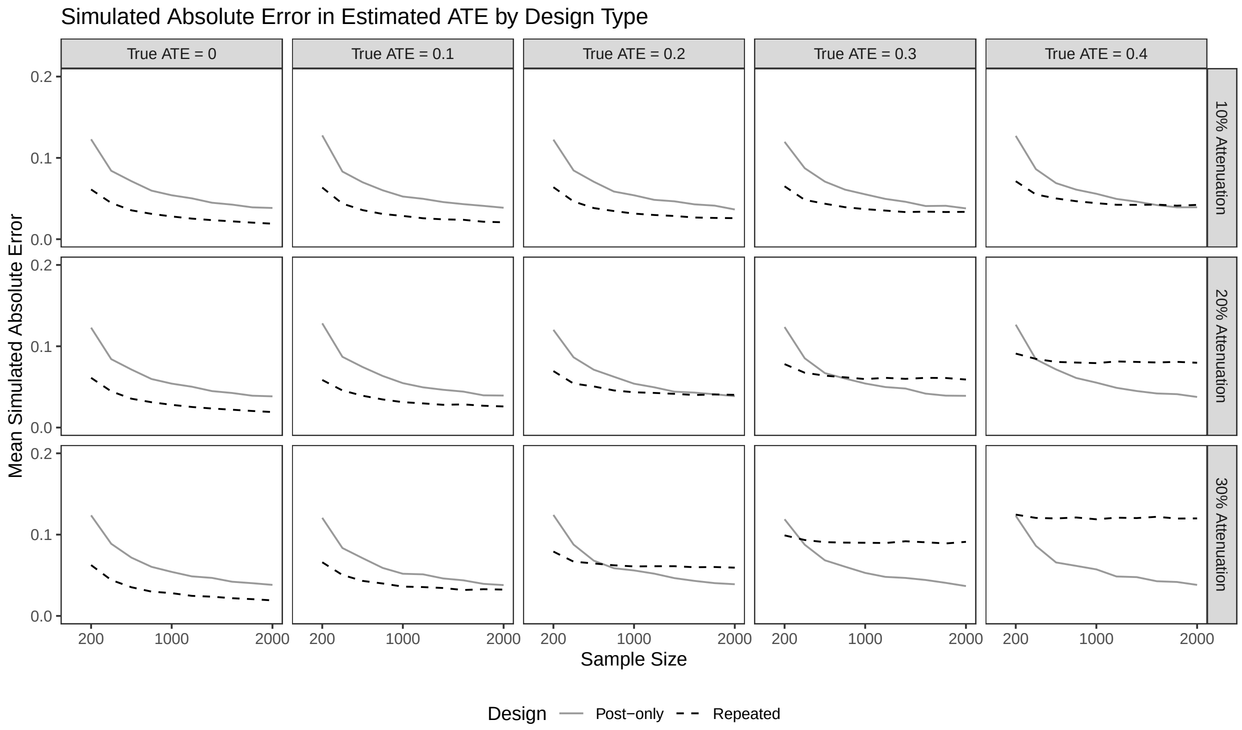

Figure 6: Absolute error in simulated experiments. Figure displays mean absolute error between the estimated and true ATE in simulated experiments at varying sample sizes, true effect sizes (Cohen's d), and true design effect (attenuation) of a repeated measure experiment.

make a repeated measure design less accurate in expectation than a post-only design.

A related concern pertains to experiments on socially sensitive topics.[23] Here, pre-treatment outcome measurement may substantially heighten social desirability biases that could induce respondents to falsify post-treatment responses (either towards consistency or towards responsiveness to the treatment, depending on the experiment), potentially putting the estimated ATE in greater jeopardy.[24] Repeatedly probing respondents about a very sensitive topic may also cause them emotional distress and increase attrition, raising ethical and practical concerns with repeated measure designs. While some of our experiments could be considered sensitive (e.g., on affirmative action and opioid clinics), none are on topics as sensitive as (say) illegal drug use or participation in violence, which are often examined

[23] We note that CSP also acknowledge this potential exception to their general recommendation (2021, 1062).

[24] That said, recent scholarship suggests that the sensitivity of items for the respondent is often overestimated (Kramon and Weghorst 2019).

in list experiments (e.g., Aronow et al. 2015; García-Sánchez and Queirolo 2021; Redlawsk, Tolbert, and Franko 2010; Walsh and Braithwaite 2008). Thus, we lack robust evidence on the risks of repeated measures for experiments on very sensitive topics, and researchers should proceed cautiously when using them in such contexts.

As social science moves towards larger samples to address concerns about under-powered research (e.g., Arel-Bundock et al. 2022), a post-only design is also a reasonable choice for high-powered experiments when the aim is to minimize absolute error, rather than to minimize the detectable effect. Researchers might particularly prefer a large-N post-only design to accurately estimate the magnitude of a treatment effect as a quantity of interest for downstream analyses. A meaningfully attenuated estimate of (e.g.) an online learning intervention could, for example, risk introducing substantial bias in downstream cost-benefit analyses for evaluating widespread policy implementation.

That said, post-only designs present considerable opportunity costs. Given arbitrarily large but nevertheless finite resources, experimenters could use repeated measure designs to increase the number of treatment arms or field multiple distinct experiments with smaller samples for roughly the same costs (and statistical power) as a single, larger post-only experiment.[25] This approach enables researchers to extend the range of interventions tested and improve cost effectiveness, expanding the scope of empirical inquiry.

In sum, though we note some general circumstances in which experimenters should still consider a post-only design, we largely concur with CSP that repeated measure designs are a better default than post-only designs. Applied social science often prioritizes identifying directional effects over estimating precise magnitudes, and like other disciplines political science continues to struggle with under-powered research (Arel-Bundock et al. 2022). We

[25] An alternate approach would be to use a mixed design strategy by randomly assigning participants to different designs for same experiment, as we do here. In theory, this would allow the experimenter to estimate the design effect for their specific setting and intervention. In practice, however, we view this as a risky approach. As we document, design effects are identifiable but small across a range of settings and interventions, meaning that excellent power for both design types is required to precisely estimate them. Absent sufficiently large samples, a mixed design strategy is thus likely to yield a very noisy estimate of the design effect, which may lead researchers to falsely conclude that the design effect is zero (or very large) in their particular context.

encourage the use of power calculations at the design stage to compare post-only and repeated measure designs under different assumptions about treatment effect size and precision (for helpful guidance on these comparisons, see Rainey 2025). Nonetheless, because treatment effects in the behavioral sciences tend to be small (Amsalem and Zoizner 2020; Funder and Ozer 2019; Gignac and Szodorai 2016; Hummel and Maedche 2019; Walter et al. 2020) and are rarely known to the experimenter a priori, repeated measure designs are generally the conservative choice. Absent a strong, justified expectation of a large treatment effect and access to a large sample, researchers are likely better served by repeated measure designs.

Between-groups versus Within-subject Experiments

One of this study's aims was to expand the evidence base on within-subject repeated measure designs. CSP's useful initial evidence comes from one experiment on anti-poverty spending ($N = 900$ students). We analyze three within-subject experiments (on anti-poverty, affirmative action, and opioid policy) across three samples for a total $n_{ij} = 39,489$, providing robust evidence on the utility of this type of repeated measure design. While we find that within-subject experiments are susceptible to some slight attenuation bias, we find that this bias is (if anything) smaller than for between-groups repeated measure experiments, and the precision gains are perhaps greater. In our bootstrapped analyses of equivalent sample sizes between designs (see Table 4), we observe a median 17.3 percent attenuation of the ATE for the within-subject experiments versus 25.1 percent for the between-groups experiments; we also observe a 57.9 percent median reduction in the standard error versus a 41.0 percent reduction. We recommend repeated measure designs for both types of experiments.

Probability versus Non-probability Samples

Fielding all of our experiments on three samples simultaneously allows us to assess the suitability of repeated measure designs for diverse sampling designs and respondent characteristics. We observe large differences in respondent characteristics between the three samples, such as higher professionalization in the non-probability samples and cross-sample variation in respondents' ability to recall their pre-treatment responses (see Appendix A.3). Nevertheless, we find no consistent differences in design effects across samples (see Figure 2). We further find no evidence that respondent professionalization alters design effects within each sample (see Appendix A.2). These results thus support repeated measure designs for both probability and non-probability general population samples.

Brief Survey Modules

In experimental survey research, repeated measures are commonly placed as far apart as possible to enable respondents to "forget" their pre-treatment responses, thus minimizing the risk of bias to the ATE. In our experiments, we randomly varied the proximity of pre- and post-treatment measures and find that distance between repeated measures alters the design effect only slightly, such that the attenuation bias increases marginally when the measures are placed further apart, as shown in Figure 3. Substantively, placing pre- and post-treatment measurements very close together appears to have similar results as placing them several minutes apart.[26] While our evidence offers no compelling reason to avoid distractor content between repeated measures, our findings offer reassurance that researchers can use repeated measure designs even when constrained to very limited survey space that precludes providing much separation.[27]

That said, our exploratory analyses regarding differences in the design effect from iterative exposure to multiple repeated measure designs in a single survey (see Appendix A.1) offers an important caution about repeated measure designs in omnibus surveys. We find that the attenuation in treatment effects from a repeated measure design increases slightly as respondents encounter more repeated measure experiments in our surveys, and that this increase is particularly pronounced for experiments fielded after an attitude-recall question.

[26] Note that our surveys are somewhat short overall—between 5 and 10 minutes for a majority of respondents. Our results cannot speak to the relative design effects of placing measures far apart on much longer surveys, or on separate surveys completed days or weeks apart.

[27] See also Sheagley and Clifford (2025) for similar findings regarding the placement of moderators.

When combining multiple studies in a single survey—as is common practice today through collaborative data collection efforts like TESS or the Cooperative Election Studies—repeated measure experiments placed in later modules may risk more severe attenuation of treatment effects, especially if attitude-recall questions are used in earlier modules.

Concluding Remarks

Considering the sum of our evidence, we offer three final remarks. First, we note that our study says little about the relative prevalence of priming, consistency, or demand effects. While one or more of these conventional concerns may contribute to the slight attenuation in repeated measure designs, there is likely heterogeneity in the relative strength of each across individuals, and some may even be operating in opposing directions to produce net attenuation on average. We encourage future research to better disentangle this knot.

Second, our 18 studies cover a lot of ground but necessarily leave much unexplored. For example, experiments on highly sensitive topics subject to strong social desirability bias may suffer from larger design effects than we identify here. Our omnibus surveys are relatively short and exclusively use the self-administered web survey mode. Repeated measure designs in other survey experimental contexts such as face-to-face or phone interviews, where the interviewers' presence may alter respondent reactions to repeated measurements (Lavrakas, Kelly, and McClain 2019), may face additional challenges that we cannot examine here—particularly for sensitive topics. Nevertheless, because experimental interventions and self-administered web surveys like ours are quite common in contemporary experimental research (Clifford, Sheagley, and Piston 2021; Jerit and Barabas 2023), we believe that our evidence provides useful insights for many experimental research contexts.

Finally, we return to the broad shift in design practice that has followed CSP's evidence-backed suggestion that "the default should shift away from the post-only design and toward repeated measure designs" (Clifford, Sheagley, and Piston 2021, 1063). Through our large-scale replications and extensions, our contribution should be viewed as a qualified endorsement of this new standard for experimental design. There remain some circumstances in which research aims can reasonably justify a traditional post-only design, but in our view these cases are not the modal enterprise in the discipline today. Our accumulated evidence suggests that the burden of justifying an experimental design should weigh more heavily on the use of post-only over repeated measure designs, rather than the historical reverse.

Acknowledgments: This research was generously supported by the Time-sharing Experiments for the Social Sciences (TESS) as part of a Special Competition on Replications, by the Rapoport Family Foundation, and by Bass Connections at Duke University. The authors thank NORC staff for their assistance with data collection, particularly Dan Costanzo and Alyssa Kahle. The authors thank Scott Clifford, Gustavo Diaz, Don Green, Jon Green, Sunshine Hillygus, Christopher Johnston, Molly Offer-Westort, Carlisle Rainey, Geoff Sheagley, Mallory SoRelle, Anton Strezhnev, Matthew Tyler, and two anonymous TESS reviewers, as well as participants at PolMeth 2024, TexMeth 2025, MPSA 2025, APSA 2025, the UT Austin Junior Methodologist Workshop, the Worldview Lab at Duke University, and the Graduate Research Workshop at the Sanford School of Public Policy for helpful comments that improved the research. Authorship for this work is in alphabetical order.

References

Amsalem, Eran and Alon Zoizner. 2020. "Real, but Limited: A Meta-analytic Assessment of Framing Effects in the Political Domain." *British Journal of Political Science* 52(1):221–237.

Arel-Bundock, Vincent, Ryan C. Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and Tom D. Stanley. 2022. Quantitative Political Science Research Is Greatly Underpowered. OSF Preprints.
URL: *https://osf.io/7vy2f*

Aronow, Peter M, Alexander Coppock, Forrest W Crawford, and Donald P Green. 2015. "Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence." *Journal of Survey Statistics and Methodology* 3(1):43–66.

Berinsky, Adam J., Michele F. Margolis, Michael W. Sances, and Christopher Warshaw. 2021. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?" *Political Science Research and Methods* 9(2):430–437.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Use change scores or control for pre-treatment outcomes? Depends on the true data generating process." *DeclareDesign* (blog):January 15, 2019.
URL: *https://declaredesign.org/blog/posts/use-change-scores-or-control.html*

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19(5):547–556.

Cannell, Charles F, Peter V Miller, and Lois Oksenberg. 1981. "Research on Interviewing Techniques." *Sociological Methodology* 12:389–437.

Chang, Linchiat and Jon A Krosnick. 2009. "National Surveys Via RDD Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality." *Public Opinion Quarterly* 73(4):641–678.

Charness, Gary, Uri Gneezy, and Michael A Kuhn. 2012. "Experimental Methods: Between-Subject and Within-subject Design." *Journal of Economic Behavior & Organization* 81(1):1–8.

Cialdini, Robert B, Melanie R Trost, and Jason T Newsom. 1995. "Preference for Consistency: The Development of a Valid Measure and the Discovery of Surprising Behavioral Implications." *Journal of Personality and Social Psychology* 69(2):318.

Clifford, Scott, Geoffrey Sheagley, and Spencer Piston. 2021. "Improving Precision without Altering Treatment Effects: Repeated Measure Designs in Survey Experiments." *American Political Science Review* 115(3):1048–1065.

Clifford, Scott and Carlisle Rainey. 2025. "The Limits (and Strengths) of Single-Topic Experiments." *Political Analysis* pp. 1–7.

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

Coppock, Alexander, Thomas Leeper, and Kevin Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates Across Samples." *Proceedings of the National Academy of Sciences* 115:12441–12446.

Cox, David R. and Peter McCullagh. 1982. "A biometrics invited paper with discussion. some aspects of analysis of covariance." *Biometrics* p. 541–561.

de Benedictis-Kessner, Justin and Michael Hankinson. 2019. "Concentrated Burdens: How Self-Interest and Partisanship Shape Opinion on Opioid Treatment Policy." *American Political Science Review* 113(4):1078–1084.

Feldman, Stanley. 1989. "Measuring Issue Preferences: The Problem of Response Instability." *Political Analysis* 1(1):25–60.

Funder, David C. and Daniel J. Ozer. 2019. "Evaluating Effect Size in Psychological Research: Sense and Nonsense." *Advances in Methods and Practices in Psychological Science* 2(2):156–168.

García-Sánchez, Miguel and Rosario Queirolo. 2021. "A Tale of Two Countries: The Effectiveness of List Experiments to Measure Drug Consumption in Opposite Contexts." *International Journal of Public Opinion Research* 33(2):255–272.

Gelman, Andrew and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9(6):641–651.

Gerber, Alan S, Donald P Green, and David Nickerson. 2001. "Testing for Publication Bias in Political Science." *Political Analysis* 9(4):385–392.

Gignac, Gilles E. and Eva T. Szodorai. 2016. "Effect Size Guidelines for Individual Differences Researchers." *Personality and Individual Differences* 102:74–78.

Gilens, Martin. 2001. "Political Ignorance and Collective Policy Preferences." *American Political Science Review* 95(2):379–396.

Hillygus, D Sunshine and Tina LaChapelle. 2022. "Diagnosing Survey Response Quality." *Handbook on Politics and Public Opinion* pp. 10–25.

Hillygus, D Sunshine, Natalie Jackson, and McKenzie Young. 2014. "Professional Respondents in Nonprobability Online Panels." *Online Panel Research: Data Quality Perspective, A* pp. 219–237.

Huber, Gregory and Matthew H. Graham. FC. Designing Survey Experiments. In *Handbook of Experimental Methodology*, ed. Leeat Yariv and Erik Snowberg. Elsevier.

Hummel, Dennis and Alexander Maedche. 2019. "How Effective is Nudging? A Quantitative Review on the Effect Sizes and Limits of Empirical Nudging Studies." *Journal of Behavioral and Experimental Economics* 80:47–58.

Ioannidis, John P.A., T.D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127(605):F236–F265.

Jerit, Jennifer and Jason Barabas. 2023. "Are Nonprobability Surveys Fit for Purpose?" *Public Opinion Quarterly* 87(3):816–840.

Kennedy, Courtney, Andrew Mercer, Scott Keeter, Nick Hatley, Kyley McGeeney, and Alejandra Gimenez. 2016. Evaluating Online Nonprobability Surveys. Pew Research Center.
URL: *https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/*

Klar, Samara, Thomas Leeper, and Joshua Robison. 2020. "Studying Identities with Experiments: Weighing the Risk of Posttreatment Bias Against Priming Effects." *Journal of Experimental Political Science* 7(1):56–60.

Kramon, Eric and Keith Weghorst. 2019. "(Mis) measuring sensitive attitudes with the list experiment: Solutions to list experiment breakdown in Kenya." *Public Opinion Quarterly* 83(S1):236–263.

Krupnikov, Yanna and Adam Seth Levine. 2014. "Cross-Sample Comparisons and External Validity." *Journal of Experimental Political Science* 1(1):59–80.

Kühberger, Anton, Astrid Fritz, and Thomas Scherndl. 2014. "Publication Bias in Psychology: A Diagnosis Based on the Correlation Between Effect Size and Sample Size." *PloS One* 9(9):e105825.

Lavrakas, Paul J., Jenny Kelly, and Colleen McClain. 2019. Investigating Interviewer Effects and Confounds in Survey-Based Experimentation. In *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment*, ed. Paul Lavrakas, Michael Traugott, Courtney Kennedy, Allyson Holbrook, Edith de Leeuw, and Brady West. Wiley.

Lerner, Jennifer S and Philip E Tetlock. 1999. "Accounting for the Effects of Accountability." *Psychological Bulletin* 125(2):255.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustment to Experimental Data: Re-examining Freedman's Critique." *The Annals of Applied Statistics* 7(1):295–318.

List, John A. 2025. The Experimentalist Looks Within: Toward an Understanding of Within-Subject Experimental Designs. NBER.
URL: *https://www.nber.org/papers/w33456*

Loken, Eric and Andrew Gelman. 2017. "Measurement Error and the Replication Crisis." *Science* 355(6325):584–585.

MacInnis, Bo, Jon A Krosnick, Annabell S Ho, and Mu-Jung Cho. 2018. "The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension." *Public Opinion Quarterly* 82(4):707–744.

Miratrix, Luke W., Jasjeet S. Sekhon, Alexander G. Theodoridis, and Luis F. Campos. 2018. "Worth Weighting? How to Think About and Use Weights in Survey Experiments." *Political Analysis* 26(3):275–291.

Mummolo, Jonathan and Erik Peterson. 2019. "Demand Effects in Survey Experiments: An Empirical Assessment." *American Political Science Review* 113(2):517–529.

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

Open Science Collaboration. 2015. "Estimating the Reproducability of Psychological Science." *Science* 349(6251):aac4716.

Peters, Gjalt-Jorn. 2017. "Why Most Experiments in Psychology Failed: Sample Sizes Required for Randomization to Generate Equivalent Groups as a Partial Solution to the Replication Crisis.".
URL: *osf.io/preprints/38vfn*

Rainey, Carlisle. 2014. "Arguing for a Neglible Effect." *American Journal of Political Science* 58(4):773–1091.

Rainey, Carlisle. 2025. Power Rules: Practical Statistical Power Calculations. Technical report OSF Preprints.
URL: *https://osf.io/preprints/osf/5am9q*

Redlawsk, David P, Caroline J Tolbert, and William Franko. 2010. "Voters, Emotions, and Race in 2008: Obama as the First Black President." *Political Research Quarterly* 63(4):875–889.

Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51(3):515–530.

Sheagley, Geoff and Scott Clifford. 2025. "No Evidence that Measuring Moderators Alters Treatment Effects." *American Journal of Political Science* 69(1):49–63.

Smith, Tom W. 1987. "That Which We Call Welfare Would Smell Sweeter: An Analysis of the Impact of Question Wording on Response Patterns." *Public Opinion Quarterly* 51(1):75–83.

Stagnaro, Michael N, James Druckman, Adam J Berinsky, Antonio A Arechar, Robb Willer, and David G Rand. 2024. "Representativeness versus Response Quality: Assessing Nine Opt-In Online Survey Samples.".
URL: *osf.io/preprints/psyarxiv/h9j2d*

Tourangeau, Roger and Kenneth A Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103(3):299.

Veselovsky, Veniamin, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023. "Prevalence and Prevention of Large Language Model Use in Crowd Work." *ArXiv* ArXiv preprint.
URL: *https://arxiv.org/pdf/2310.15683*

Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West. 2023. "Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks." *ArXiv* ArXiv preprint.
URL: *https://arxiv.org/html/2306.07899*

Walsh, Jeffrey A and Jeremy Braithwaite. 2008. "Self-reported Alcohol Consumption and Sexual Behavior in Males and Females: Using the Unmatched-Count Technique to Examine Reporting Practices of Socially Sensitive Subjects in a Sample of University Students." *Journal of Alcohol and Drug Education* pp. 49–72.

Walter, Nathan, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. "Fact-Checking: A Meta-analysis of What Works and for Whom." *Political Communication* 37(3):350–375.

Wilson, David C., David W. Moore, Patrick F. McKay, and Derek R. Avery. 2008. "Affirmative Action Programs for Women and Minorities: Expressed Support Affected by Question Order." *Public Opinion Quarterly* 72(3):514–522.

Zaller, John and Stanley Feldman. 1992. "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 36(3):579–616.

Zizzo, Daniel John. 2010. "Experimenter Demand Effects in Economic Experiments." *Experimental Economics* 13:75–98.

Online Appendix to

New Evidence and Design Considerations for Repeated Measure Experiments in Survey Research

Contents

A Supplemental Results

A.1 Internal Meta-analyses

In Table A.1.1, we report the tabular results of the main internal meta-analyses of design effects that are shown in Figure 2 of the main text.

Table A.1.1: Estimated Meta-analytic Design Effects by Design Type and Sample

Experiment Type	Sample	k	Estimate	Std. Error	95% CI	p-value
Both Types	AmeriSpeak	6	-0.222^{*}	0.057	[−0.368,−0.076]	0.011
	Prolific	6	−0.148	0.171	[−0.588, 0.292]	0.426
	Lucid	6	$-0.162^{\dagger}$	0.068	[−0.336, 0.011]	0.061
	Total	18	-0.200^{***}	0.040	[−0.285,−0.115]	< 0.001
Between-groups	AmeriSpeak	3	$-0.259^{\dagger}$	0.082	[−0.614, 0.095]	0.088
	Prolific	3	−0.226	0.118	[−0.733, 0.281]	0.195
	Lucid	3	−0.157	0.097	[−0.575, 0.262]	0.249
	Total	9	-0.210^{**}	0.049	[−0.322,−0.097]	0.003
Within-subject	AmeriSpeak	3	−0.187	0.079	[−0.525, 0.150]	0.140
	Prolific	3	−0.020	0.377	[−1.640, 1.601]	0.963
	Lucid	3	−0.169	0.094	[−0.572, 0.236]	0.216
	(Total - Outlier)	8	-0.227^{**}	0.051	[−0.348,−0.107]	0.003
	Total	9	−0.149	0.094	[−0.367, 0.069]	0.153

†p<0.10; *p<0.05; **p<0.01; ***p<0.001

Note: Table displays the results of internal meta-analyses of k studies by design type and sample. Design effect estimates are expressed as the proportional change in the post-only design ATE.

Note that our experimental design combines multiple experiments into a single omnibus survey, a common practice in contemporary social science. Because the order of these experiments is randomized, we can assess whether the design effect of repeated measure designs (relative to the post-only design) changes as respondents are iteratively exposed to more repeated measure design experiments (from a first to fourth encounter within a single survey). Table 5 in the main text reports internal meta-analyses using all 18 experiments, subset by respondents' first through fourth repeated measure design encountered in the survey. Table A.1.2 reports a combined fixed-effects meta-analysis that estimates differences in the design effect (relative to the post-only design) between the first and subsequent repeated measure design experiments.

Table A.1.2: Internal Meta-analysis with Repeated Measure Experiment Order Fixed Effects

Fixed Effect	Design Effect Estimate	Std. Error	95% CI	p-value
Second Encountered	−0.090	0.074	[−0.237, 0.058]	0.229
Third Encountered	−0.057	0.074	[−0.206, 0.091]	0.442
Fourth Encountered	−0.149*	0.074	[−0.297,−0.000]	0.049
Constant (First Encountered)	−0.123*	0.053	[−0.228,−0.019]	0.022

†p<0.10; *p<0.05; **p<0.01; ***p<0.001

Note: Table displays the results of an internal meta-analysis ($k = 18$) of the repeated measure design effect with a fixed effect for the order of repeated measure design experiment encountered. The first repeated measure design experiment encountered is held as the reference category.

Table A.1.3: Internal Meta-analysis with Repeated Measure Experiment Relative to First Attitude Recall Question

Subset	Design Effect Estimate	Std. Error	95% CI	p-value
First RM Experiment (Pre-recall)	−0.131*	0.050	[−0.236,−0.025]	0.018
Subsequent Pre-recall RM Experiments	−0.145*	0.055	[−0.260,−0.030]	0.017
All Pre-recall RM Experiments	−0.132*	0.050	[−0.238,−0.026]	0.017
All Post-recall RM Experiments	−0.249***	0.049	[−0.351,−0.146]	< 0.001

†p<0.10; *p<0.05; **p<0.01; ***p<0.001

Note: Table displays the results of internal meta-analyses ($k = 18$ for each) of the repeated measure design effect as a function of each respondent's first encounter with an attitude recall (perceived change) question.

A.2 Respondent Professionalization

In Figure A.2.1, we report the results of internal meta-analyses of design effects conditional on degree of respondent professionalization (above or below within-sample median). These results are also provided in tabular format in Table A.2.1. We operationalize respondent professionalization two ways, using the self-reported counts of surveys completed in the past 30 days (survey count) or the self-reported number of survey companies the respondent has completed surveys for in the past 30 days (panel memberships). We observe no substantive differences between respondents who are more or less professionalized in each sample; the estimated design effects are uniformly negative (from −0.009 to −0.043) and rarely differ from each other significantly.

Table A.2.1: Estimated Meta-analytic Design Effects by Professionalization

Quantile	Sample	k	Estimate	Std. Error	95% CI	p-value
Below Median (Survey Count)	AmeriSpeak	6	-0.043^{**}	0.011	[−0.071,−0.016]	0.010
	Prolific	6	−0.009	0.014	[−0.046, 0.027]	0.538
	Lucid	6	−0.018	0.011	[−0.046, 0.010]	0.153
Above Median (Survey Count)	AmeriSpeak	6	−0.020	0.012	[−0.052, 0.011]	0.157
	Prolific	6	−0.033	0.024	[−0.094, 0.027]	0.218
	Lucid	6	−0.011	0.013	[−0.045, 0.023]	0.444
Below Median (Panel Memberships)	AmeriSpeak	6	-0.033^{*}	0.009	[−0.057,−0.010]	0.015
	Prolific	6	$-0.027^{\dagger}$	0.013	[−0.061, 0.007]	0.096
	Lucid	6	−0.014	0.010	[−0.040, 0.012]	0.225
Above Median (Panel Memberships)	AmeriSpeak	6	−0.028	0.016	[−0.068, 0.012]	0.127
	Prolific	6	−0.018	0.028	[−0.090, 0.054]	0.553
	Lucid	6	−0.014	0.017	[−0.056, 0.029]	0.447

†p<0.10; *p<0.05; **p<0.01; ***p<0.001

Note: Table displays the results of internal meta-analyses (with k studies) of design effects by respondent professionalization (within-sample), operationalized as the count of surveys completed in the past 30 days or the count of active panel memberships in the past 30 days.

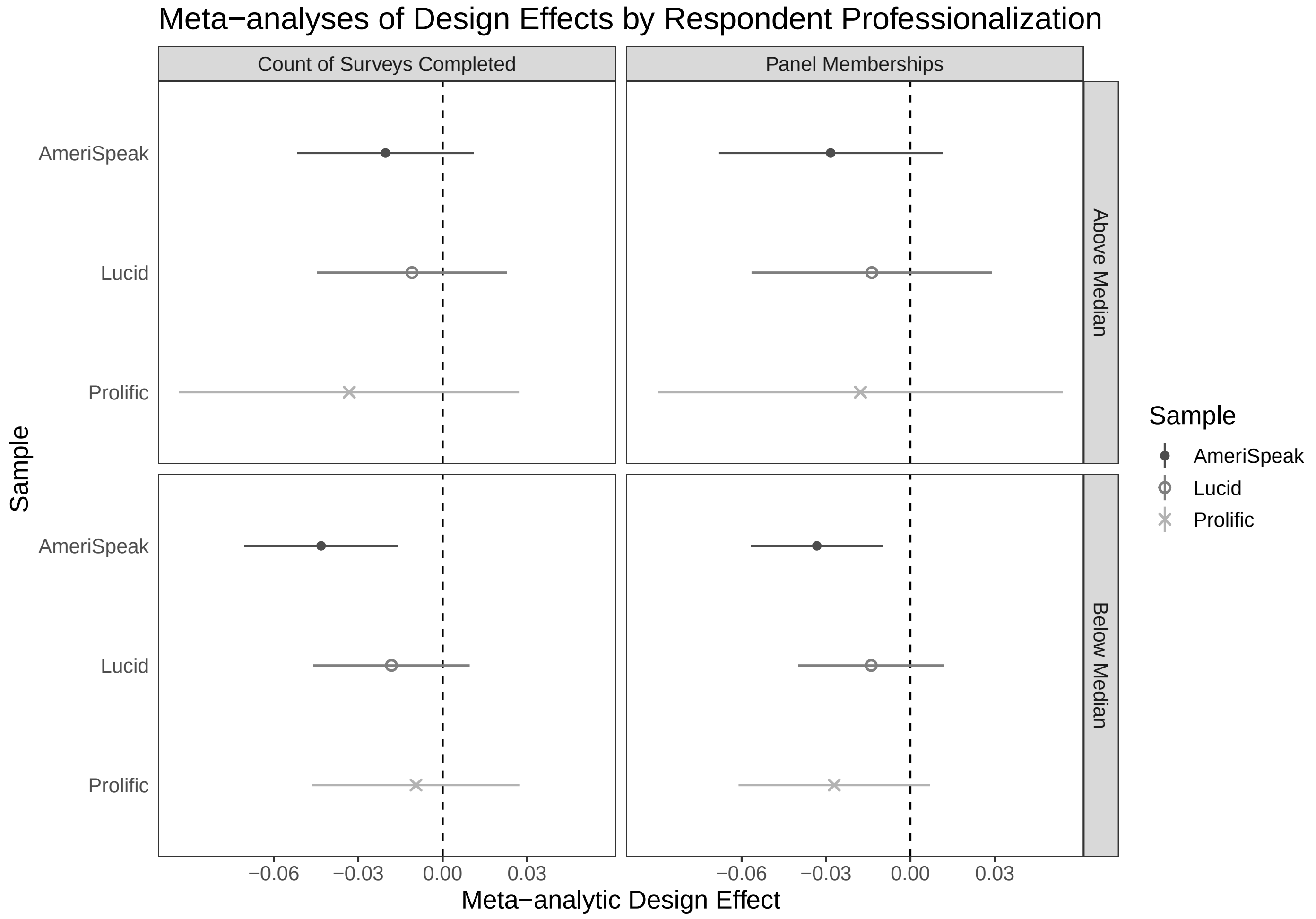

Figure A.2.1: Meta-Analytic design effect by respondent professionalization. The top (bottom) row of panels indicates the design effect among above (below) median respondents within each sample on each professionalization measure (columns).

A.3 Perceived Attitude Change

Figure A.3.1 shows the percentage of repeated measure experiment observations by observed stability or change versus self-reported perceived stability or change. We find that respondents in most repeated measure observations (69.9 percent) provide the same response both pre- and post-treatment, as shown in the center column of Figure A.3.1, and most of these (81.0 percent) self-report that their attitudes stayed about the same. Among the remaining 30.1 percent of participants whose observed responses did change (left and right columns), only 39.1 percent accurately perceived that change, with half (50.0 percent) incorrectly reporting no change and the remaining 10.9 percent reporting a change in the opposite direction from their actual change.

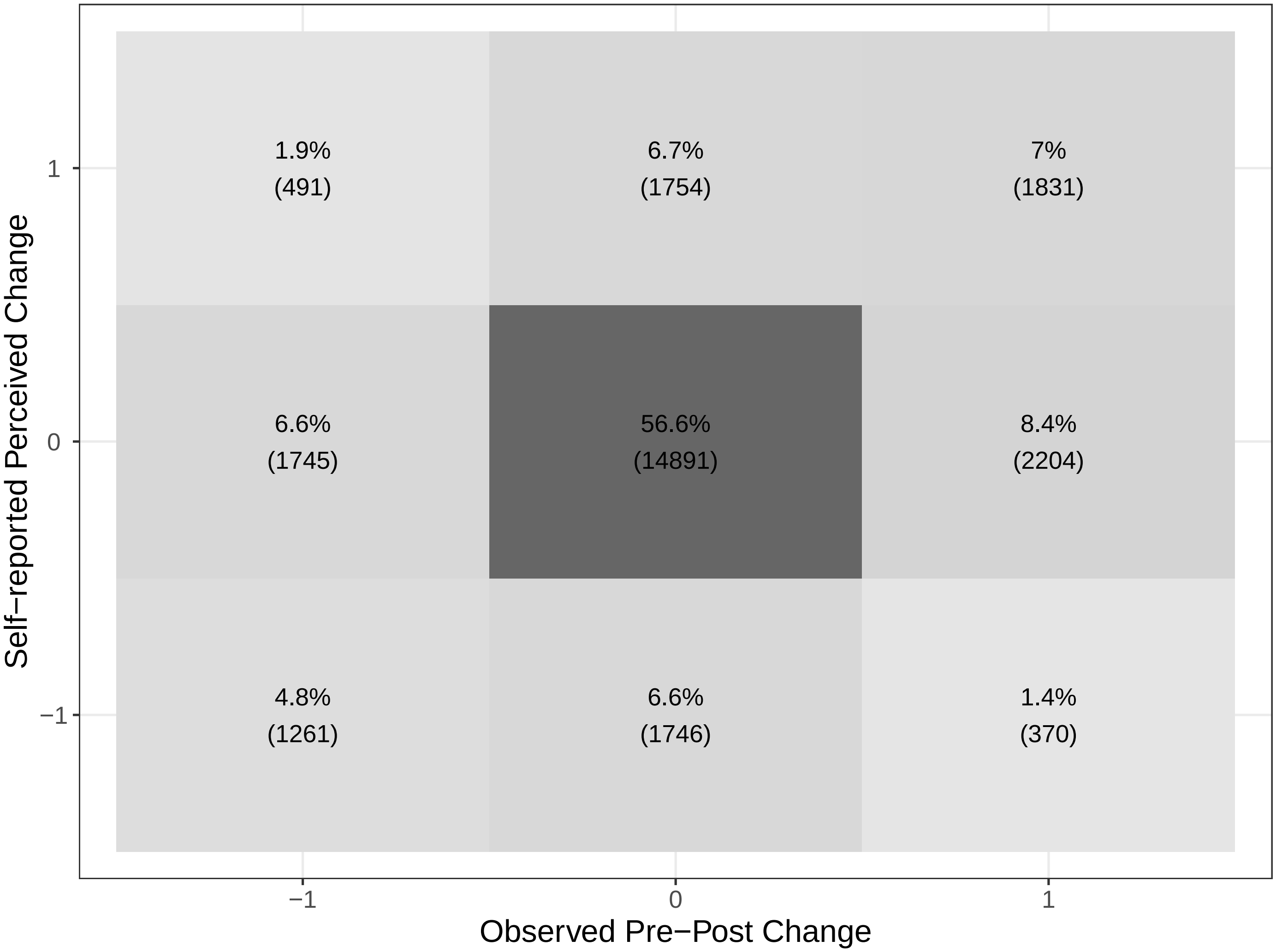

Figure A.3.1: Respondent perception of attitude change. Figure shows a contingency table of pooled respondents (all between-groups repeated measure observations in all samples) by actual observed pre-post change in responses (columns) and self-reported perceived pre-post change (rows). Frequency counts for each cell are shown in parentheses.

Figure A.3.2 offers an exploratory report of differences in perceived attitude change across samples and experimental context, with a contingency table of perceived versus observed change for each between-groups experiment in each sample. We find some differences in overall accuracy by sample: Lucid respondents accurately perceived their level of change in 60.1 percent of observations, whereas the overall accuracy rate is 68.1 percent in the AmeriSpeak sample and 78.1 percent in the Prolific sample. In part, this appears to be because Prolific respondents were more stable in their attitudes; in every experiment, a higher percentage of Prolific respondents were both stable in their observed pre-post responses and self-reported no change in attitude than for either of the other two samples. These results align with recent evidence that Prolific respondents tend to be more attentive than Lucid respondents, but may react differently to some treatment (Stagnaro et al. 2024).

We also observe some slight heterogeneity among the three between-groups experiments. The (cross-sample) accuracy rate is highest for the foreign aid experiment at 70.8 percent and slightly lower in the drug imports experiment at 68.9 percent. The lowest accuracy rate is in the GMO framing experiment 65.5 percent, which is perhaps to be expected given that all respondents in that experiment received either a positive or negative framing (no pure control) and were thus more likely to change their responses post-treatment. Indeed, we see that only 24.2 and 24.8 percent of respondents in the foreign aid and drug imports experiments (respectively) actually moved, whereas 41.1 percent of respondents did so in the GMO experiment.

Does the accuracy of respondents' self-perceptions of their attitude change (or lack thereof) relate to the design effect of repeated measure experiments? We conduct an exploratory analysis by separating the between-groups repeated measure observations into two subsamples, based on the first such repeated measure each respondent encountered in the survey flow: those who accurately perceived their level of attitude change in that experiment (that is, the three cross-diagonal cells from the bottom left to top right in Figure A.3.1) versus those who did not (all other cells).[28] We then re-estimate the design effect (as proportional change in the post-only design ATE in each sample) with each subsample for each subsequent between-groups repeated measure experiment that each respondent encountered, and finally meta-analyze these estimated design effects for accurate versus inaccurate respondents. We condition on the first between-groups repeated measure experiment to analyze subsequent repeated

[28] We caution that our interpretation of the center-top and center-bottom cells as "inaccurate" is on less firm ground, in that a respondent's views may have shifted slightly but not by enough to merit a change in response on a coarse close-ended scale.

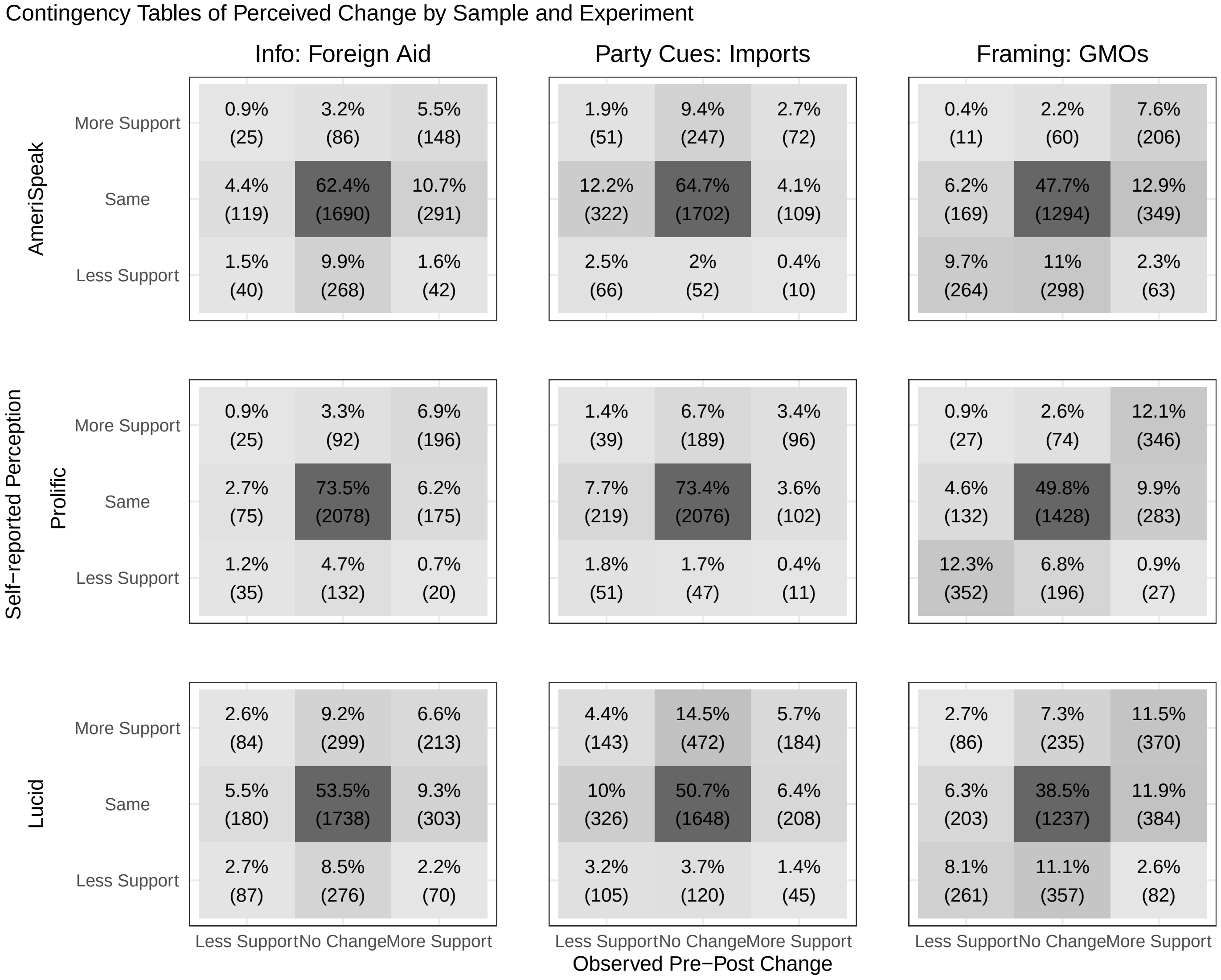

Figure A.3.2: Respondent perception of attitude change by sample and experiment. Figure shows contingency tables of respondents in each between-groups repeated measure experiment (panel columns) in each sample (panel rows) by actual observed pre-post change in responses (within-panel columns) and self-reported perceived pre-post change (within-panel rows). Frequency counts for each cell are shown in parentheses.

measure experiments (exclusively) to ensure that the conditioning variable is independent of treatment assignment in the analyzed experiments.[29] The results are reported in Table A.3.1.

We find little evidence of a relationship between self-reported recall accuracy and the design effect,

[29] Though this restriction substantially reduces the sample size (and power) of these analyses, using accuracy within each experiment itself would confound any potential effect of accuracy with a mechanical effect from treatment assignment. The recall questions that allow us to distinguish accurate versus inaccurate perceptions of change are post-treatment. Treatment itself predicts inaccuracy (33.9 percent among treated observations, 29.3 percent among control observations) because treated respondents are more likely to provide a different response relative to their pre-treatment observation, whereas most control respondents can accurately satisfice by self-reporting no change. Propensity to accurately report one's own level of change in each respective condition may vary by unobserved respondent characteristics; through this mechanism, analyzing the design effect conditional on accuracy directly within a given experiment may partially de-randomize the assignment to treatment.

Table A.3.1: Repeated Measure Results by Accuracy in Perceived Attitude Change

		Accurate Perception			*Inaccurate Perception*			*Design Effect vs. Post-only* (Δ in ATE)	
Experiment	Sample	Est. ATE	SE	Obs.	Est. ATE	SE	Obs.	Accurate	Inaccurate
Fgn. Aid	AmeriSpeak	0.052^{***}	0.009	905	0.056^{***}	0.015	428	-0.036^{*}	-0.033
Fgn. Aid	Prolific	0.071^{***}	0.008	1,032	0.061^{***}	0.017	338	-0.041^{**}	-0.050^{*}
Fgn. Aid	Lucid	0.062^{***}	0.012	942	0.039^{*}	0.018	675	-0.001	-0.023
Drug Imp.	AmeriSpeak	0.016	0.019	866	0.068^{*}	0.035	444	-0.109^{***}	-0.056
Drug Imp.	Prolific	0.031^{*}	0.014	964	0.115^{**}	0.037	270	$-0.065^{\dagger}$	0.019
Drug Imp.	Lucid	0.066^{**}	0.024	819	0.040	0.033	552	-0.044	-0.070
GMOs	AmeriSpeak	0.098^{***}	0.011	906	0.122^{***}	0.019	403	-0.064^{***}	-0.040
GMOs	Prolific	0.157^{***}	0.009	1,212	0.151^{***}	0.022	280	-0.022	-0.029
GMOs	Lucid	0.153^{***}	0.012	973	0.114^{***}	0.019	634	0.009	-0.030

†p<0.10; *p<0.05; **p<0.01; ***p<0.001

Note: Table displays the results of each experiment under the repeated measure design, conditional on whether the respondent accurately reported the direction of change in their attitude pre-post (or attitude stability) on a prior between-groups repeated measure experiment. The final two columns report the respective design effects (vs. post-only).

as shown in Table A.3.1. The attenuation bias is stronger among respondents who accurately perceived their level of pre-post change in four studies but stronger among those who were inaccurate in five, and the difference in ATEs between these two subsamples is statistically significant in none (meta-analytic estimated difference -0.001, $p = 0.990$).[30] Across all nine between-groups experiments, the meta-analytic design effect (expressed as proportional to the respective post-only ATE) among the recall-accurate respondents is -0.433 ($p = 0.008$) and a very similar -0.302 ($p = 0.004$) among recall-inaccurate respondents.

[30] The design effect (relative to the post-only design) is statistically significant more frequently among recall-accurate respondents, but this is primarily a function of power, as recall-accurate respondents outnumber recall-inaccurate respondents by more than two to one.

A.4 Simulation Results

To simulate the impact of the design effect of repeated measure designs on several considerations of experimental design, we simulate 1,000 repeated measure experiments and 1,000 post-only experiments at each of several sample sizes (from $N = 200$ to $N = 2000$, in increments of 200),[31], true effect sizes (from 0 to 0.4, in increments of 0.1) expressed in terms of Cohen's d, and design effect attenuation assumptions (10 percent, 20 percent, or 30 percent).[32]

For each experiment, we assume that each observation of the outcome variable is a random within-individual draw from a true value plus random error (i.e., Feldman 1989; Zaller and Feldman 1992). The true value for each individual is a random draw from a normal distribution with a mean of zero and a standard deviation of 1. The pre-treatment outcome measure (repeated measure experiments only) is the true value plus error, which is a random draw from a normal distribution with a mean of zero and a standard deviation of 0.4. The post-treatment outcome measure is again the true value plus a random draw for error (again from a normal distribution with a mean of zero and a standard deviation of 0.4). We randomly assign each individual to treatment or control with a random draw from a binomial distribution ($p = 0.5$). The post-treatment outcome value for those assigned to the treatment condition is the true (baseline) value of their inherent attitude plus random error, plus a homogeneous treatment effect (equal to the defined true effect) that is attenuated by the defined value in the repeated measure experiments.

After drawing these random values for each individual up to the defined sample size, we regress the post-treatment outcome variable on the treatment indicator, plus the pre-treatment outcome variable for the repeated measure experiments only. We then record the coefficient on the treatment indicator and its standard error in each experiment. Note that the choice of a standard deviation of 0.4 for the response error distribution provides a correlation between repeated measures of about 0.8 and a corresponding reduction in the standard error of the treatment coefficient of about 40 percent, which is in line with our observed data from the field, but slightly conservative.

We simulate 1,000 experiments for each design at each value of the sample size, true effect, and design effect parameters, or 300,000 experiments in total. For each combination of parameters, we

[31] Note that we simulate two-arm experiments, meaning our simulated sample sizes correspond to 100 to 1,000 respondents per cell in expectation, in increments of 100.

[32] Note that the mean observed attenuation in our fielded experiments is around 20 percent; the 10 percent and 30 percent attenuation rates reflect the possibility that our observed rate may be an over-estimate or under-estimate, respectively, in some experimental settings.

then calculate the statistical power of the test (as 1 minus the proportion of false negatives, shown in Figure 5 in the main text), the mean absolute error between the treatment coefficient and the true treatment effect (shown in Figure 6 in the main text), the mean false discovery rate (the proportion of experiments in which the treatment coefficient is significant but negative when the true effect is positive, or the confidence does not include zero when the true effect is zero), and the mean coverage rate (the proportion of experiments in which the confidence interval includes the true effect parameter) for each design type.

The results on the latter two metrics are shown in Figures A.4.1, and A.4.2. The false discovery rate is equivalent under both designs in most circumstances, but superior (lower) under a repeated measure design when the sample size and true effect size are both small. Finally, the coverage rate is similar for both designs when the true effect is zero or the true effect size and attenuation are both small, but otherwise the coverage rate under a repeated measure design quickly declines as the sample size, true effect size, and degree of attenuation increase.

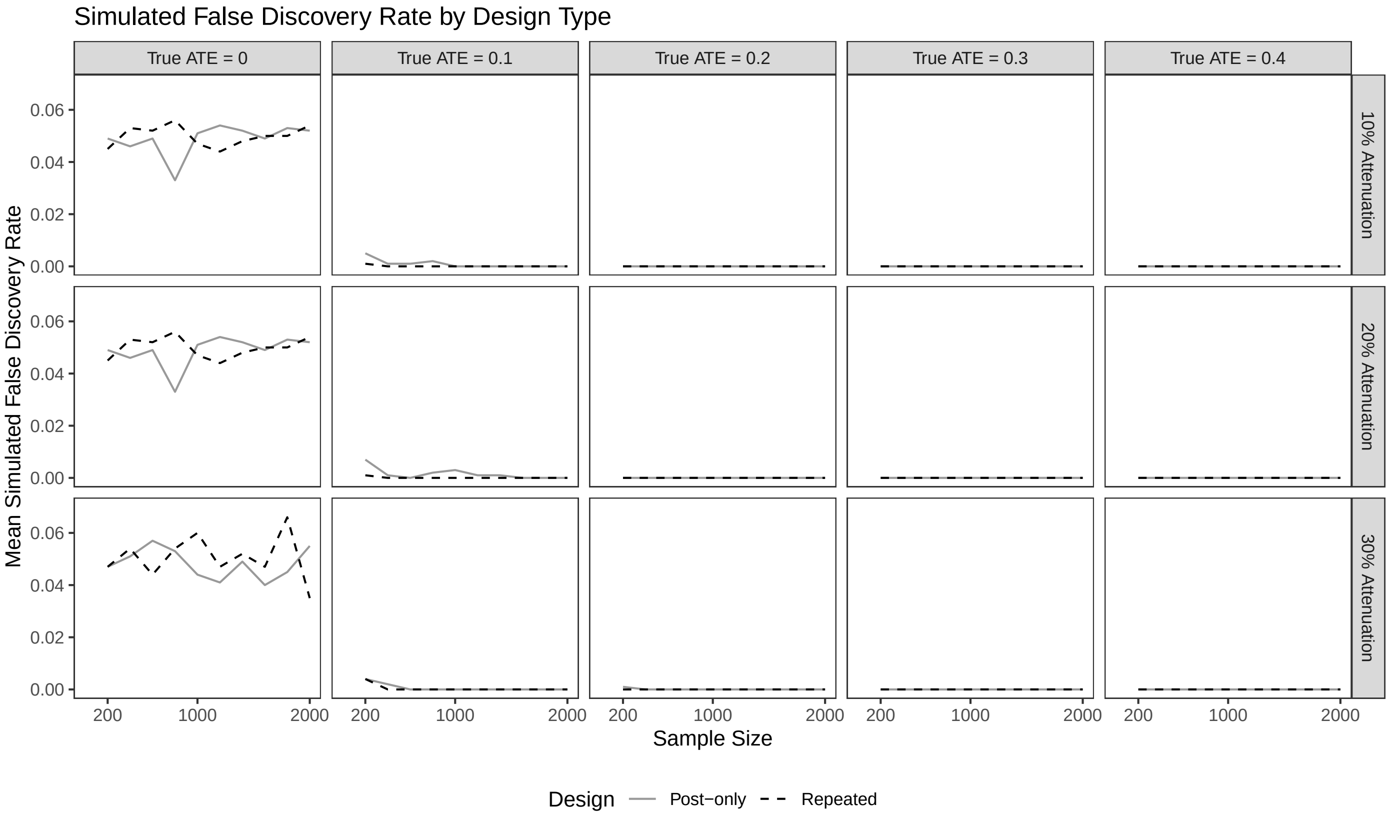

Figure A.4.1: False discovery rate in simulated experiments. Figure displays mean false discovery rate in simulated experiments at varying sample sizes, true effect sizes (Cohen's d), and true design effect (attenuation) of a repeated measure experiment.

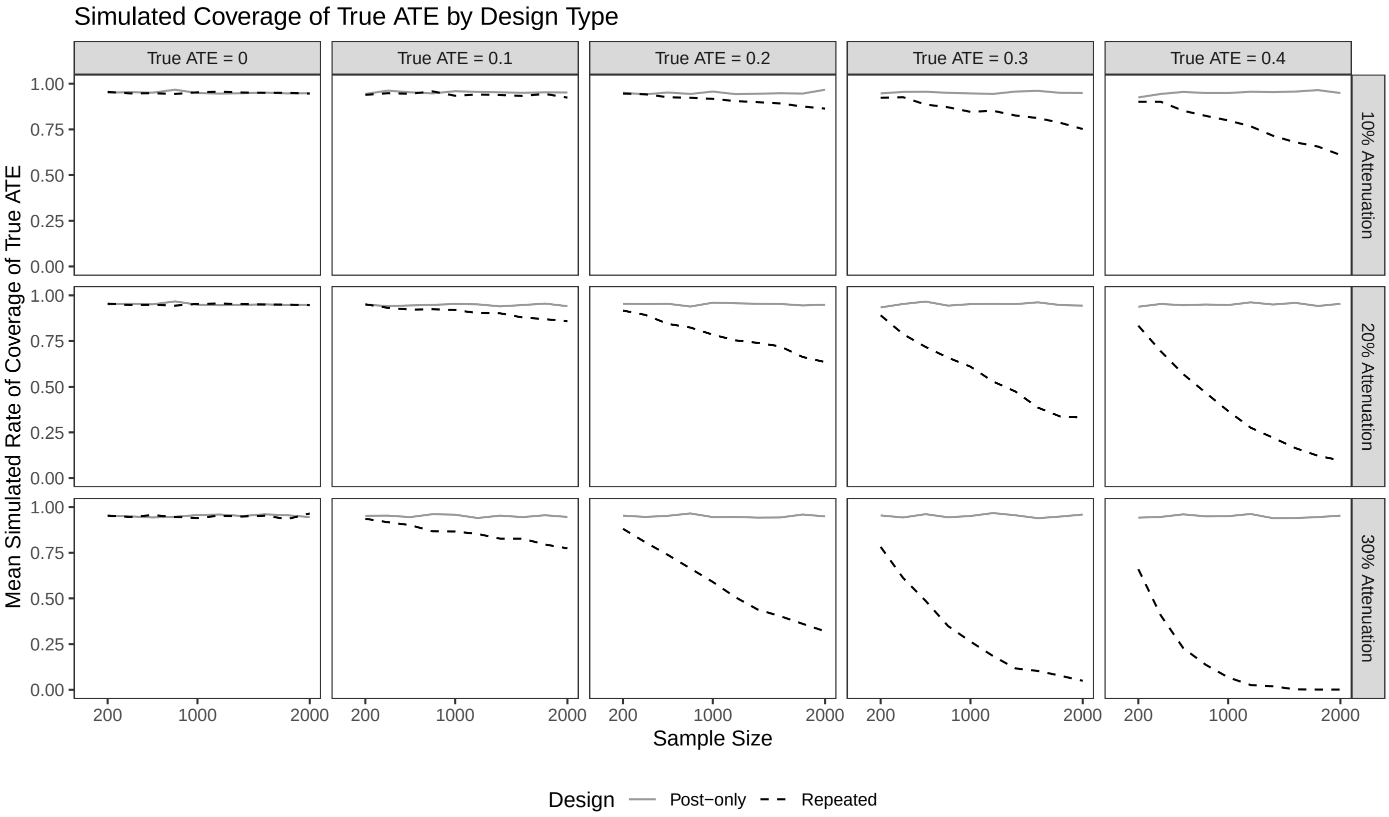

Figure A.4.2: Coverage rate simulated experiments. Figure displays mean rate of coverage of the true ATE in simulated experiments at varying sample sizes, true effect sizes (Cohen's d), and true design effect (attenuation) of a repeated measure experiment.

A.5 Evidence of the Influence of Clifford, Sheagley, and Piston (2021)

In the first four years since publication (through April 2025), Clifford, Sheagley, and Piston (2021) has garnered citations from 118 original peer-reviewed published studies spanning 83 different journals. Of the 118, 83 are original studies referencing CSP to justify using repeated measure designs. In Table A.5.1 below, the pre-post designs have bolded titles and within-subject designs do not.

Table A.5.1: Citations Referencing CSP to Justify Repeated Measure Designs

No.	Title	Journal
1	**Belief change in times of crisis: Providing facts about COVID-19-induced inequalities closes the partisan divide but fuels intra-partisan polarization about inequality**	Social Science Research
2	Can a constitutional monarch influence democratic preferences? Japanese emperor and the regulation of public expression	Social Science Quarterly
3	**Career adaptability of interpreting students: A case study of its development and interactions with interpreter competences in three Chinese universities**	Frontiers in Psychology
4	Correcting the Misinformed: The Effectiveness of Fact-checking Messages in Changing False Beliefs	Political Communication
5	**Depression and suicidality as evolved credible signals of need in social conflicts**	Evolution and Human Behavior
6	**Descriptive, injunctive, or the synergy of both? Experimenting normative information on behavioral changes under the COVID-19 pandemic**	Frontiers in Psychology
7	Equating silence with violence: When White Americans feel threatened by anti-racist messages	Journal of Experimental Social Psychology
8	Estimating the between-issue variation in party elite cue effects	Public Opinion Quarterly
9	Mass support for proposals to reshape policing depends on the implications for crime and safety	Criminology & Public Policy
10	Partisan news versus party cues: The effect of cross-cutting party and partisan network cues on polarization and persuasion	Research & Politics
11	**Reliable Sources? Correcting Misinformation in Polarized Media Environments**	American Politics Research
12	**The influence of unknown media on public opinion: Evidence from local and foreign news sources**	American Political Science Review
13	Unilateral Inaction: Congressional Gridlock, Interbranch Conflict, and Public Evaluations of Executive Power	Legislative Studies Quarterly
14	**When Journalists Run for Office: The Effects of Journalist-Candidates on Citizens' Populist Attitudes and Voting Intentions**	International Journal of Communication
15	A randomized experiment evaluating survey mode effects for video interviewing	Political Science Research and Methods
16	Biased expectations? An experimental test of which party selectors are more likely to stereotype ethnic minority aspirants as less favorable than ethnic majority aspirants	Politics, Groups, and Identities
17	**Does Evidence Matter? The Impact of Evidence Regarding Aid Effectiveness on Attitudes Towards Aid**	The European Journal of Development Research
18	**Does moral rhetoric fuel or reduce divides between parties and non-copartisan voters?**	Electoral Studies
19	**Can <3's Change Minds? Social Media Endorsements and Policy Preferences**	Social Media + Society
20	Equality, Reciprocity, or Need? Bolstering Welfare Policy Support for Marginalized Groups with Distributive Fairness	American Political Science Review
21	**How interactive visualizations compare to ethical frameworks as stand-alone ethics learning tools for health researchers and professionals**	AJOB Empirical Bioethics
22	**Latino-Targeted Misinformation and the Power of Factual Corrections**	Journal of Politics
23	**Media stereotypes, prejudice, and preference-based reinforcement: toward the dynamic of self-reinforcing effects by integrating audience selectivity**	Journal of Communication

No.	Title	Journal
24	**Polarizing political polls: How visualization design choices can shape public opinion and increase political polarization**	IEEE Transactions on Visualization and Computer Graphics
25	Politicized Battles: How Vacancies and Partisanship Influence Support for the Supreme Court	American Politics Research
26	**Public support for phasing out carbon-intensive technologies: the end of the road for conventional cars in Germany?**	Climate Policy
27	Public Support for Professional Legislatures	State Politics & Policy Quarterly
28	**Shifting partisan public opinion towards Community Choice Aggregation through outreach and awareness**	PLOS One
29	**Test score equating of multiple-choice mathematics items: techniques from characteristic curve of modern psychometric theory**	Discover Education
30	**The Holocaust, the Socialization of Victimhood and Outgroup Political Attitudes in Israel**	Comparative Political Studies
31	**The Most Important Election of Our Lifetime: Focalism and Political Participation**	Political Psychology
32	Women Experts and Gender Bias in Political Media	Public Opinion Quarterly
33	**Anger expressions and coercive credibility in international crises**	American Journal of Political Science
34	**Banklash: How Media Coverage of Bank Scandals Moves Mass Preferences on Financial Regulation**	American Journal of Political Science
35	**Beyond Changing Minds: Raising the Issue Importance of Expanding Legal Immigration**	Perspectives on Politics
36	**Beyond partisan filters: Can underreported news reduce issue polarization?**	PLOS One
37	Bureaucracy and Cyber Coercion	International Studies Quarterly
38	Changes in Perceptions of Border Security Influence Desired Levels of Immigration	Journal of Conflict Resolution
39	**Confronting Racism of Omission: Experimental Evidence of the Impact of Information about Ethnic and Racial Inequality in the United States and the Netherlands**	Du Bois Review: Social Science Research on Race
40	**Correcting Myopia: Effect of Information Provision on Support for Preparedness Policy**	Political Research Quarterly
41	**Critical Race Theory and Asymmetric Mobilization**	Political Behavior
42	**Digital Cloning of the Dead: Exploring the Optimal Default Rule**	Asian Journal of Law and Economics
43	**Does informing citizens about the non-meritocratic nature of inequality bolster support for a universal basic income? Evidence from a population-based survey experiment**	European Societies
44	**Does the prospect of further sovereignty loss fuel Euroscepticism? A population-based survey experiment**	European Societies
45	**Explaining the educational gradient in trust in politicians: a video-vignette survey experiment**	West European Politics
46	**Filling the EU information deficit mitigates negative EU attitudes among the least knowledgeable. Evidence from a population-based survey experiment**	Journal of European Integration
47	Frontline employees' responses to citizens' communication of administrative burdens	Public Administrative Review
48	**Going negative when spoiled for choice? Destabilizing and boomerang effects of negative political messaging in multiparty systems with multimember districts**	Political Research Exchange
49	Imagined otherness fuels blatant dehumanization of outgroups	Communications Psychology
50	Moral Rhetoric, Extreme Positions, and Perceptions of Candidate Sincerity	Political Behavior
51	**Partisan Poll Watchers and Americans' Perceptions of Electoral Fairness**	Public Opinion Quarterly
52	**Paying for growth or goods: Tax morale among property owners in Lagos**	Journal of Experimental Political Science
53	Public Reactions to Communication of Uncertainty: How Long-Term Benefits Can Outweigh Short-Term Costs	Public Opinion Quarterly
54	Role model stories can increase health professionals' interest and perceived responsibility to engage in climate and sustainability actions	The Journal of Climate Change and Health
55	Scientific supremacy: How do genetic narratives relate to racism?	Politics and the Life Sciences
56	Supplemental online resources improve data literacy education: Evidence from a social science methods course	PLOS One
57	**The Impacts of US Foreign Policy on Taiwanese Public Support for Independence**	Journal of Chinese Political Science

No.	Title	Journal
58	**The Personal Vote in a Polarized Era**	American Journal of Political Science
59	The policy acknowledgement gap: Explaining (mis)perceptions of government social program use	Policy Studies Journal
60	**The power of empirical evidence: assessing changes in public opinion on constitutional emergency provisions**	Public Choice
61	When partisanship and technocratic credibility collide: mass attitudes and central bank endorsements of fiscal policy in Canada and the USA	Socio-Economic Review
62	Active Student Responding and Student Perceptions: A Replication and Extension	Teaching of Psychology
63	**Anti-Black Political Violence and the Historical Legacy of the Great Replacement Conspiracy**	Perspectives on Politics
64	Breaking the rules, but for whom? How client characteristics affect frontline professionals' prosocial rule-breaking behavior	Journal of Public Administration Research and Theory
65	**Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments**	American Political Science Review
66	**Do Justifications Affect Tolerance for Administrative Burdens? Evidence From a Survey Experiment Among Policymakers**	Social Policy and Administration
67	**Does the military lose public confidence without compliance with civilian control? Experimental evidence from Japan**	Journal of Peace Research
68	**External coercion and public support: The case of the US–China trade war**	Journal of Peace Research
69	**Information, Uncertainty, and Public Support for Brinkmanship During the 2023 Debt Limit Negotiations**	British Journal of Political Science
70	Language Cues and Perceptions of Nationalism	Political Behavior
71	**Making Issues Matter: Local Media and Policy-Based Evaluations of Politicians**	Political Behavior
72	**No Evidence that Measuring Moderators Alters Treatment Effects**	American Journal of Political Science
73	On motives and means: how approach and justification for court-curbing impact public trust	Democratization
74	**Participating in a Digital-History Project Mobilizes People for Symbolic Justice and Better Intergroup Relations Today**	Psychological Science
75	**Priority Projects: Constituent Spending Demand and the Benefits of Congressional Credit Claiming**	Legislative Studies Quarterly
76	**(Small D-democratic) vacation, all I ever wanted? The effect of democratic backsliding on leisure travel in the American states**	Journal of Experimental Political Science
77	**Survey measures of democratic attitudes and social desirability bias**	Political Science Research and Methods
78	Testing theories of political persuasion using AI	Proceedings of the National Academy of Sciences
79	**The Effect of International Actors on Public Support for Government Spending Decisions**	International Studies Quarterly
80	**The policy basis of group sentiments**	Political Science Research Methods
81	To What End? Policy Objectives and US Public Support for Political Warfare	Foreign Policy Analysis
82	Tweet no harm: Offer solutions when alerting the public to voter suppression efforts	Communication and the Public
83	**Winning Votes and Changing Minds: Do Populist Arguments Affect Candidate Evaluations and Issue Preferences?**	British Journal of Political Science

B Study Information

The study was approved by [REDACTED UNIVERSITY]'s Institutional Review Board under protocol [REDACTED]. Anonymized preregistration materials for this study are available here.

B.1 Sampling Procedure

The data for this study come from a three omnibus surveys of the U.S. general adult population (combined $N_i = 13,163$) recruited from three vendors. We describe the sampling procedure for each sample in turn.

The first sample ($n_i = 4,029$) was drawn from the probability-based AmeriSpeak panel. This component of the study was funded by the National Science Foundation via the Time-Sharing Experiments for the Social Sciences (TESS) maintained by the University of Rochester. The AmeriSpeak panel, operated by NORC, is a probability-based panel designed to be representative of the US household population. Randomly selected US households are sampled using area probability and address-based sampling, with a known, non-zero probability of selection from the NORC National Sample Frame. These sampled households are then contacted by US mail, telephone, and field interviewers (face to face). The panel provides sample coverage of approximately 97 percent of the U.S. household population. Those excluded from the sample include people with P.O. Box only addresses, some addresses not listed in the USPS Delivery Sequence File, and some newly constructed dwellings. While most AmeriSpeak allows non-internet households can participate in AmeriSpeak surveys by telephone, this option was not included for this study; this study was also available only in English. Households without conventional internet access but having web access via smartphones are allowed to participate in AmeriSpeak surveys by web. More information about the panel and sampling design is available at AmeriSpeak.norc.edu.

For this study, NORC invited consented AmeriSpeak panelists to participate in the omnibus survey hosted directly by the authors on the Qualtrics platform. This survey was fielded from June 27th to July 15th, 2024. NORC invited 19,024 total panelists to participate, sending email reminders 3 days after initial invitation and every 5 days thereafter, plus a final email reminder on July 9th. The survey completion rate among invited participants was 21.2 percent. The weighted cumulative response rate (which accounts for panel recruitment, panel retention, and survey completion) is 3.7 percent. AmeriSpeak panelists were offered the cash equivalent of $2.00 for completing the survey. The median completion time was 6.1 minutes. A total of 4,250 panelists entered the survey; as preregistered, we exclude 82 who failed to complete the survey, and a further 139 for either extreme speeding (less than 1/3 of the median completion time) or item non-response on at least half of the survey questions. This provides our analysis sample of $N_i = 4,029$. Although provided by NORC, we do not apply sample weights in our analyses to preserve statistical power (Miratrix et al. 2018).

The second and third samples are non-probability convenience samples recruited via quota sampling from the Prolific ($n_i = 4,261$) and Lucid (now Cint, $n_i = 4,869$) opt-in online panels. This component of the study was funded by the Rapoport Family Foundation and by Bass Connections at Duke University. The Prolific sample recruited with the following quotas: sex (50.9% female, 49.1% male), age (11.8% age 18-24, 17.5% age 25-34, 17.0% age 35-44, 15.8% 45-54, and 37.9% age 55 or above), and party affiliation (29.5% Democrat, 27.7% Republican, 42.8% Independent). The Lucid sample was recruited with joint quotas on sex, age, and race/ethnicity as shown in Table B.1.1 (note that the "Other" category was not an explicit quota, but includes anyone who opted not to report their sex, age, or race/ethnicity to Lucid in the prescreen phase).

Table B.1.1: Lucid Demographic Quotas

Sex	**Age**	**Race/Ethnicity**	**Quota**	**Sex**	**Age**	**Race/Ethnicity**	**Quota**
Male	18-24	White	2.9%	Male	35-44	Black	1.2%
Female	18-24	White	3.0%	Female	35-44	Black	1.2%
Male	18-24	Hispanic	1.2%	Male	35-44	Other Race	0.6%
Female	18-24	Hispanic	1.3%	Female	35-44	Other Race	0.6%
Male	18-24	Black	0.8%	Male	45-54	White	4.2%
Female	18-24	Black	0.8%	Female	45-54	White	4.4%
Male	18-24	Other Race	0.5%	Male	45-54	Hispanic	1.5%
Female	18-24	Other Race	0.5%	Female	45-54	Hispanic	1.5%
Male	25-34	White	4.4%	Male	45-54	Black	1.0%
Female	25-34	White	4.5%	Female	45-54	Black	1.1%
Male	25-34	Hispanic	1.8%	Male	45-54	Other Race	0.6%
Female	25-34	Hispanic	1.9%	Female	45-54	Other Race	0.6%
Male	25-34	Black	1.2%	Male	55+	White	12.8%
Female	25-34	Black	1.3%	Female	55+	White	13.3%
Male	25-34	Other Race	0.7%	Male	55+	Hispanic	2.0%
Female	25-34	Other Race	0.8%	Female	55+	Hispanic	2.0%
Male	35-44	White	4.2%	Male	55+	Black	2.0%
Female	35-44	White	4.6%	Female	55+	Black	2.0%
Male	35-44	Hispanic	1.7%	Male	55+	Other Race	1.1%
Female	35-44	Hispanic	1.7%	Female	55+	Other Race	1.1%
Other							5.3%

Recruited panelists entered (separate) omnibus surveys hosted directly by the authors on the Qualtrics platform. These surveys were fielded from July 3rd to July 15th, 2024. Prolific respondents received $1.00 for completing the study; Lucid provided participants with an incentive to participate in our study, but these incentives differ by respondent and are not disclosed to the researcher. The median completion time for Prolific participants was 7.2 minutes and 7.3 minutes for Lucid participants. After consenting to participate the study, participants were screened for eligibility to confirm that they were at least 18 years of age and resided in the United States. We recruited a total of 4,398 eligible Prolific participants and 6,094 eligible Lucid participants into the study. As preregistered, we exclude 94 participants in the Prolific sample and 354 in the Lucid sample who failed to complete the respective survey, as well as 5 Prolific participants and 190 Lucid participants who failed an explicit attention check during screening (failing to select either "B" or "D" when asked to identify the second and fourth letters of the English alphabet). Finally, we exclude 38 Prolific participants and 681 Lucid participants for extreme speeding (completing the survey in less than 1/3 of the within-sample median time) or failing at least two of the following preregistered quality checks: self-reported age and birth year do not correspond, within a tolerance of +/- 2 years; self-reported state of residence and zip code do not match; speeding (completing the survey in less than 1/2 of the median time); scoring less than 0.65 on Qualtrics' internal reCaptcha measure; partially failing the pretreatment attention check by selecting either "B" or "D" but not both; or failing a second explicit pre-treatment attention check question about activities in the past 30 days (by self-reporting unlikely activities like purchasing an airline company, climbing a mountain on Mars, or having a fatal heart attack, or failing to self-report likely activities like eating a meal and using electricity). All of the screening and exclusion criteria were preregistered. The exclusions reduce the final analysis samples to $n_i = 4,261$ Prolific respondents and $n_i = 4,869$ Lucid participants. Appendix B.2 provides descriptive statistics for all samples. The observations are not weighted.

As with all survey research, the design and collection of data has limitations for all three samples, and resulting estimates may involve unmeasured error that limits representativeness to the target population.

B.2 Sample Characteristics

Table B.2.1: Sample Characteristics by Vendor (Unweighted)

Category	**AmeriSpeak**	**Lucid**	**Prolific**
Male	48%	46%	48%
Mean Age	49.47	50.08	46.68
White	66.16%	63.13%	69.67%
Black	11.74%	13.70%	12.91%
Hispanic	13.88%	8.95%	4.06%
Multi-race	2.81%	7.01%	6.13%
Other Race or Ethnicity	5.41%	7.21%	7.23%
Less than high school degree	4.39%	4.87%	0.77%
High school diploma or equivalent	18.82%	29.83%	12.20%
Some college/Associate degree	38.21%	33.99%	33.11%
Bachelor's degree	22.37%	20.44%	35.79%
Postgraduate degree	16.21%	10.87%	18.12%
Less than $60,000	43.87%	64.46%	42.86%
$60,000–$99,999	24.95%	20.94%	27.06%
$100,000–$149,999	17.50%	8.97%	18.27%
$150,000–$199,999	7.42%	3.14%	6.83%
$200,000 or more	6.26%	2.49%	4.98%
Democrat	46.06%	44.48%	49.43%
Independent	18.10%	16.31%	12.09%
Republican	35.84%	39.21%	38.48%

Note: Table reports unweighted percentages of respondents included in the final analysis samples.

B.3 Randomization Procedure

We randomized design conditions, treatment conditions, and the order of the experimental content for each respondent with a complex, multi-stage randomization procedure. In the first randomization stage, two of the six experiments (at the respondent level) were randomly assigned post-only designs, with the remaining four experiments assigned to repeated measure designs. In the second stage, we randomized assignment to treatment or control stimuli for each experiment. For the within-subject experiments, this assignment dictated which wording appeared first and which appeared second (if in the repeated measure condition). In the third stage, we randomly ordered the "pre-treatment" questions (or first wordings) for the four repeated measure experiments. These four questions (the "pre-treatment" block) were displayed to respondents sequentially.

All remaining experimental content was randomized as part of the "post-treatment" block. This block included eight sub-blocks: a treatment/control stimulus and immediate post-treatment measurement for each of the six experiments, plus six unrelated questions about attitudes regarding the National Football League (NFL) that were split into two sub-blocks of three questions each. The sub-blocks for the three between-groups experiments additionally included a question about perceived change in attitude (only if assigned to the repeated measures design), and the sub-block for Study 6 additionally included a post-treatment covariate measure about personal exposure to opioid addiction.

In the fourth randomization stage, we randomly assigned each respondent to one of two order-randomization procedures for the overall post-treatment block: either a "full-random" or a "forced-short" procedure, which was then executed as the final fifth stage. In the full-random procedure, all sub-blocks in the post-treatment block were displayed to respondents in a random order. In the forced-short procedure, the sub-blocks for the two experiments whose pre-treatment content appeared last (that is, the third and fourth pre-treatment items) were forced to appear immediately following the pre-treatment block, either in the same order or the inverse, with equal probability. All other sub-blocks were randomly ordered and were displayed to respondents subsequently in that order. This alternate procedure ensured that more repeated measure experiments appeared close together than was likely if we fully randomized the order of the sub-blocks.

To illustrate this complex randomization design, Table B.3.1 shows the realized randomization outcomes and respondent experience for two hypothetical respondents.

Table B.3.1: Example Randomization Outcomes

Randomization Stage	**Respondent A**	**Respondent B**
Stage 1: Assign Post-only	Clinic, Poverty	Foreign, Clinic
Stage 2: Assign Treatment Vector	{0, 0, 0, 1, 1, 0}	{1, 0, 1, 0, 0, 1}
Stage 3: Order Pre-treat Block	Drugs, Affirm, GMOs, Foreign	Affirm, Poverty, GMOs, Drugs
Stage 4: Select Procedure	Full-random	Forced-short
Stage 5: Order Post-block	Poverty, NFL-B, Clinic, GMOs, Drugs, Affirm, Foreign, NFL-A	Drugs, GMOs, Foreign, NFL-A, Affirm, NFL-B, Poverty, Clinic
Respondent Experience	Drugs: Pre	Affirm: Pre
	Affirm: Pre	Poverty: Pre
	GMOs: Pre	GMOs: Pre
	Foreign: Pre	Drugs: Pre
	Poverty: Treatment & Post	Drugs: Control & Post
	**NFL-B	Drugs: Attitude Change
	*Clinic: Control & Post	*GMOs: Treatment & Post
	Clinic: Covariate	GMOs: Attitude Change
	*GMOs: Control & Post	Foreign: Treatment & Post
	GMOs: Attitude Change	Foreign: Attitude Change
	Drugs: Control & Post	**NFL-A
	Drugs: Attitude Change	Affirm: Control & Post
	Affirm: Treatment & Post	**NFL-B
	Foreign: Control & Post	Poverty: Control & Post
	Foreign: Attitude change	*Clinic: Treatment & Post
	**NFL-A	Clinic: Covariate

Note: The upper panel provides two hypothetical example randomizations executed at each of five stages to determine the display order of the survey for individual respondents. The bottom panel provides the resulting display order experienced by the respondent. One star (*) indicates that the item counts double for the purpose of determining distance between repeated measures; two stars (**) indicates that the item counts triple (the NFL blocks each include three questions). To identify the distance between repeated measures, sum the number of items between them plus the number of stars (*).

As shown in Table B.3.1, all respondents first answers pre-treatment questions for all four repeated measure experiments. Respondent A then answers post-treatment questions for all six experiments (including the post-only experiments) in a completely random order, with the unrelated NFL content also included in the mix. In contrast, Respondent B, assigned to the "forced-short" procedure, completes the post-treatment questions for the two repeated measure experiments that appeared last in the pre-treatment block (the GMO and prescription drug experiments, in this case), guaranteeing that some repeated measure content appears close together, before answering all remaining post-treatment questions in a random order (again with the NFL content mixed in). Table B.3.1 also illustrates how we measure distance between repeated measures: the count of items in between the pre- and post-treatment measures, plus the count of stars in between, indicating that an item that counts twice (*, because the question is preceded by a longer paragraph) or thrice (**, because it denotes a block of three questions). For example, the calculated distance between measures in the prescription drugs experiment is 13 for Respondent A, and 0 for Respondent B because the measures appear back-to-back.

B.4 Survey Questionnaire

B.4.1 Screening and Demographics

This content was included prior to the experimental content in the Prolific and Lucid surveys only. This content was not included on the AmeriSpeak survey.

Screening

Age: What is your age in years? Please enter a whole number. *[Open-ended]*
State: In which state do you currently reside? *[List of U.S. states, DC, and Puerto Rico]*
Attention Check 1: What are the second and fourth letters of the English alphabet? This is an attention check question and the correct answer is B and D (please select both).

- A
- B
- C
- D
- E

Demographics

Gender: Which of the following best describes your gender?

- Male
- Female
- Something else

Race/ethnicity: Which racial or ethnic group best describes you? Please check all that apply.

- Asian or Asian-American
- Black or African-American
- Hispanic or Latino
- Middle Eastern
- Native American or Alaskan Native
- Native Hawaiian or other Pacific Islander
- White
- Something else

Education: Which is the highest level of education that you have completed?

- Less than a high school degree or equivalent
- High school degree or equivalent (for example: GED)
- Some college, but no degree
- 2-year college degree (Associate's degree)
- 4-year college degree (Bachelor's degree)
- Postgraduate degree (MA, MBA, MD, JD, PhD, etc.)

Employment Status: What is your current employment status?

- Employed full-time
- Employed part-time

- Unemployed
- Retired
- Full-time homemaker
- Student
- Something else

Household Income: Which of the following describes your total annual household income from 2023—that is, the total income everyone living in your household made together, before taxes, in 2023?

- Less than $20,000
- $20,000 to $39,999
- $40,000 to $59,999
- $60,000 to $79,999
- $80,000 to $99,999
- $100,000 to $119,999
- $120,000 to $149,999
- $150,000 to $199,999
- $200,000 or more

Year Born: In what year were you born? Please enter a 4-digit number. *[Open-ended]*
Zip Code: In which ZIP code do you currently reside? Please enter a 5-digit number. *[Open-ended]*
Attention Check 2: Which of the following have you done in the past 30 days? Please check all that apply.

- Eaten a meal
- Purchased an airline company
- Read a book
- Climbed the Olympus Mons
- Had a fatal heart attack
- Used electricity

B.4.2 Experimental Content

In this section, we provide the question wording and response options for all experimental content for studies 1-6. We specify the standard TESS unit length of each item (1 unit for most items, 2 units for those whose question is preceded by a longer paragraph). Note that the order of items was randomized as discussed in the main text.

Foreign Aid (Study 1)

Foreign Aid Pretreatment/Control (1 TESS unit): "Do you think spending on foreign aid should be increased or decreased?"

- Greatly increased
- Slightly increased
- Kept about the same
- Slightly decreased
- Greatly decreased

Foreign Aid Treatment (1 TESS unit): "Spending on foreign aid currently makes up about 1% of the federal budget. Do you think federal spending on foreign aid should be increased or decreased?"

- Greatly increased
- Slightly increased
- Kept about the same
- Slightly decreased
- Greatly decreased

Drug Imports (Study 2)

Drug Imports Pretreatment/Control (1 TESS unit): "Do you support or oppose allowing individuals to import prescription drugs from Canada?"

- Strong support
- Somewhat support
- Slightly support
- Neither support nor oppose
- Slightly oppose
- Somewhat oppose
- Strongly oppose

Drug Imports Treatment (1 TESS unit): "Democrats tend to favor and Republicans tend to oppose allowing individuals to import prescription drugs from Canada. Do you support or oppose this policy?"

- Strong support
- Somewhat support
- Slightly support
- Neither support nor oppose
- Slightly oppose
- Somewhat oppose
- Strongly oppose

GMOs (Study 3)

GMO Pretreatment (1 TESS unit): "How strongly do you favor or oppose the production and consumption of genetically modified foods?"

- Strongly favor
- Favor
- Slightly favor
- Neither favor nor oppose
- Slightly oppose
- Oppose
- Strongly oppose

Anti-GMO Control (2 TESS units): "As you may know, opponents of genetically modified foods point out that there have not been studies on the long-term health effects of genetically modified foods on humans. And a recent study on animals found that genetically modified potatoes damaged the digestive tracts of rats. How strongly do you favor or oppose the production and consumption of genetically modified foods?"

- Strongly favor
- Favor
- Slightly favor
- Neither favor nor oppose
- Slightly oppose
- Oppose
- Strongly oppose

Pro-GMO Treatment (2 TESS units): "As you may know, supporters of genetically modified foods point out that a recent study on genetically modified foods found that a type of rice ("golden rice") can be produced with a high content of vitamin A, which is used to prevent blindness. How strongly do you favor or oppose the production and consumption of genetically modified foods?"

- Strongly favor
- Favor
- Slightly favor
- Neither favor nor oppose
- Slightly oppose
- Oppose
- Strongly oppose

Perceived Attitude Change (Studies 1-3 Only)

Recall Previous Attitude (1 TESS unit): "As you may remember, we also asked you about your support or opposition to [foreign aid / importing subscription drugs from Canada / genetically modified foods (GMOs)] earlier in the survey. To the best of your memory, how have your preferences about [foreign aid / importing subscription drugs from Canada / genetically modified foods (GMOs)] changed since then?"

- Much more supportive
- A little more supportive
- Stayed about the same
- A little more opposed
- Much more opposed

Anti-poverty (Study 4)

Welfare (1 TESS unit): "Generally speaking, do you think we're spending too much, too little or about the right amount on welfare?"

- Too much
- About the right amount
- Too little

Assistance to the Poor (1 TESS unit): "Generally speaking, do you think we're spending too much, too little or about the right amount on assistance to the poor?"

- Too much
- About the right amount
- Too little

Affirmative Action (Study 5)

Affirmative Action Gender (1 TESS unit): "Do you generally favor or oppose affirmative action programs for women?"

- Favor
- Oppose
- No opinion

Affirmative Action Race (1 TESS unit): "Do you generally favor or oppose affirmative action programs for racial minorities?"

- Favor
- Oppose
- No opinion

Opioid Clinic (Study 6)

Opioid Clinic Near Condition (2 TESS units): "Medication-assisted treatment clinics provide help for people with substance abuse problems. They do this by providing needed medication (such as methadone) and follow-up that can keep them off dangerous opioids and prevent deadly overdoses. Would you support the opening of a new medication-assisted treatment clinic for opioid addiction a 1/4 mile (5 minute walk) from your home?"

- Strongly support
- Somewhat support
- Neither support nor oppose
- Somewhat oppose
- Strongly oppose

Opioid Clinic Far Condition (2 TESS units): "Medication-assisted treatment clinics provide help for people with substance abuse problems. They do this by providing needed medication (such as methadone) and follow-up that can keep them off dangerous opioids and prevent deadly overdoses. Would you support the opening of a new medication-assisted treatment clinic for opioid addiction 2 miles (40 minute walk) from your home?"

- Strongly support
- Somewhat support
- Neither support nor oppose
- Somewhat oppose
- Strongly oppose

Opioid Clinic Personal Exposure (1 TESS unit): "Do you personally know anyone who has ever been addicted to opioids, including prescription painkillers or heroin?"

- Yes, I personally know someone who has been addicted to opioids (such as a family member, a friend, an acquaintance, or myself)
- No, I do not know anyone who has ever been addicted to opioids

Unrelated Items (Distractor Content)

NFL Block 1 (3 TESS units): "We're interested in what people do in their spare time. How much attention would you say you pay to football games in the National Football League (NFL)?"

- A lot
- Some
- None

“Without consulting any sources, do you happen to know if any of the following slogans are associated with the NFL? It’s OK if you don’t know or aren’t sure, just tell us that.”

- “Intercept Cancer”
- “End Racism”
- “Inspire Change”
- “Salute to Service”
- “End Concussions”
- “It Takes All of Us”
- “Play It Safe”

“Should the NFL encourage people to do any of the following things?”

- Register to vote in upcoming elections
- Follow players on social media
- Place bets on upcoming games
- Recycle to save the planet
- Increase exercise to improve health
- Treat people equally regardless of their personal characteristics

NFL Block 2 (3 TESS units): “Which of the following teams played in the NFL Super Bowl in February of 2024? (select two)”

- New England Patriots
- Dallas Cowboys
- Kansas City Chiefs
- Philadelphia Eagles
- San Francisco 49ers
- I don’t know or am not sure

“Do you happen to know which of the following products the football player Patrick Mahomes endorses?”

- Apple
- State Farm Insurance
- Lexus
- Pepsi
- All of the above
- I don’t know or am not sure

“Do you happen to know which of the following people the football player Travis Kelce has dated?”

- Ariana Grande
- Taylor Swift
- Alexandria Ocasio-Cortez
- Kylie Jenner
- All of the above
- I don’t know or am not sure

B.4.3 Post-Experimental Content

This content was included on our surveys following the experimental content.

Professionalization Measures

"To the best of your memory, how many other online surveys have you completed in the past 30 days, not including this one?" *[Open-ended]*

"To the best of your memory, in the past 30 days, how many different online survey companies have you completed one or more surveys for, not including this one?" *[Open-ended]*