Repeated Measure Designs are Superior for (Most) Experimental Survey Research Applications

Diana Jordan
Duke University

Trent Ollerenshaw
University of Houston

Andrew Trexler*
Duke University

February 14, 2025

Abstract

An influential study in the *American Political Science Review* by Clifford, Sheagley, and Piston (2021) finds that including pre-treatment measures of outcome variables in survey experiments does not bias treatment effect estimates and greatly improves precision, prompting many researchers to adopt repeated measure designs. In a large-scale partial replication, we experimentally manipulate the design of six classic experiments in political science and field all six experiments in three separate samples of U.S. adults (total $N_i = 13,163$). We also provide three extensions that assess the broader suitability of repeated measure designs, specifically by fielding a larger set of within-subject experimental designs, by manipulating repeated measures' proximity, and by fielding our experiments on both probability-based and non-probability samples. In contrast to the original study, we find consistent evidence of a small attenuation of treatment effects in repeated measure designs. However, this average attenuation bias is sufficiently small that we largely affirm the original authors' recommendation to prefer repeated measure designs in most research applications, because the large gains to statistical precision will (in expectation) typically produce a more accurate estimate ATE. Further, we provide robust evidence that repeated measure designs are appropriate for within-subject and between-groups experiments, for extremely short surveys, and for both probability and non-probability samples.

Acknowledgements: 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 with the AmeriSpeak panel, particularly Dan Costanzo and Alyssa Kahle. The authors thank Scott Clifford, Jon Green, Sunshine Hillygus, Chris Johnston, Carlisle Rainey, Geoff Sheagley, and two anonymous TESS reviewers for helpful comments that improved the research. This research was previously presented at the Society for Political Methodology 2024 summer meeting.

* Corresponding author. Please direct correspondence to andrew.trexler@duke.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 different treatment conditions and measures the outcome variable(s) only post-treatment. The average treatment effect (ATE) is then estimated by comparing the outcome means or other sample statistics across the treatment groups. However, this predominant design suffers from low precision when estimating treatment effects, may miss small or varying treatment effects (Mutz 2011), and risks overestimating effects (Gelman and Carlin 2014; Loken and Gelman 2017). Given the increasing evidence that low precision contributes to low replicability rates in social scientific research (Arel-Bundock et al. 2022; Gelman and Carlin 2014), improving experimental design is essential for advancing research in political science and related fields.

A common alternative experimental design is the "repeated measure" design, in which outcomes are measured pre-treatment in addition to post-treatment, either within the same survey or on separate waves in a panel design. Researchers have often been reluctant to implement repeated measure designs, especially in the same survey, because of concerns that pre-treatment measurement of outcomes may inadvertently bias treatment effect estimates by priming respondents to the treatment, inducing consistency pressures, or creating demand incentives. However, a recent influential study by Clifford, Sheagley, and Piston (2021, referred to as CSP hereafter) in the *American Political Science Review* offers compelling evidence that the repeated measure design significantly improves the precision of treatment effect estimates compared to post-only designs but crucially does not bias the estimated ATE. CSP therefore conclude that traditional concerns about repeated measure designs distorting the ATE through priming, consistency, or demand effects can be largely dismissed, recommending "that researchers use pre-post and within-subject designs whenever possible" (Clifford, Sheagley, and Piston 2021, 1062). This recommendation has gained traction, with many researchers now using these designs to improve the precision and reliability of their findings. As of January 2025, CSP has been cited 198 times on Google Scholar, with 88 of

these citations solely justifying the use of repeated measure designs.

The rapid adoption of repeated measures designs is a promising development in survey research and speaks to the importance of CSP's findings. Yet, CSP's conclusion rests on just six experiments, all conducted using convenience samples with relatively professionalized survey respondents, who may respond to repeated measure designs differently than other sampled populations. While CSP's results are promising, the substantial shift in experimental practice since CSP's publication warrants further investigation with large-scale replications in new samples to confirm their findings. Additionally, beyond the fundamental concerns of bias and precision discussed by CSP, researchers lack information on key design considerations that could impact the utility of repeated measure designs in some experimental settings. For example, CSP and other researchers have typically placed pre- and post-treatment measures as far apart as possible within a single survey, or even on separate waves of a panel survey, out of concern that placing them too close together might make the repetition of measurement more apparent and thus introduce bias. We do not yet know whether repeated measure designs are suitable for very short surveys—or perhaps even most effective when closely spaced, by reducing the random noise that may accumulate between measurements. Best practices around the implementation of repeated measure designs thus remain understudied and underdeveloped.

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

We do so by experimentally manipulating the design of six previously published political science experiments, which include three within-subject experiments and three between-groups experiments to allow for a robust comparison across experiment types. We also varied the proximity of repeated measures in our experiments to evaluate how this design consideration affects bias and precision. We fielded all six experiments in omnibus surveys on three distinct samples of U.S. adults ($N_j = 18$ studies, $N_i = 13,163$ total respondents, $N_{ij} = 78,978$ total observations). These samples include a probability-based sample from the AmeriSpeak panel maintained by NORC ($n_i = 4,033$) and two non-probability samples (Lucid $n_i = 4,869$, Prolific $n_i = 4,261$) with varying levels of respondent professionalization and quality (Stagnaro et al. 2024). These large samples provide us with substantially greater statistical power to detect small design effects and assess potential moderators.

In contrast to CSP's original findings, we observe a small but consistent attenuation of treatment effects in repeated measures designs compared to post-only designs. Nevertheless, our findings largely affirm the original authors' recommendation to favor repeated measure designs over post-only designs in (most but not all) practical research applications, as the vast improvement in statistical precision is usually more important than the weak attenuation bias introduced by the design effect, providing a more accurate ATE in expectation in most circumstances. 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 even in surveys contexts in which pre- and post-treatment measures must necessarily appear in close proximity. While we do identify some rare circumstances where post-only designs may still be preferable, our findings broadly reinforce the field's nascent shift toward repeated measures designs and the enhanced precision they offer.

Repeated Measure Designs in Social Science

Survey experiments are among the most widespread tools for social inquiry, with the "post-only," between-groups design being the most common. In this design, participants are randomly assigned to either treatment or control conditions and exposed to different stimuli, after which outcomes are measured and compared across conditions. Here, outcomes are measured solely post-treatment, and differences between the groups' outcomes are interpreted as average treatment effects (ATE). Under a set of relatively weak assumptions—successful randomization, the stable unit treatment value assumption (SUTVA), and no differential attrition—the post-only design can provide unbiased estimates of the ATE.

While the post-only design effectively minimizes bias, it often suffers from poor statistical precision. These designs are prone to high levels of statistical noise, requiring large sample sizes to produce reliable estimates (Peters 2017). Although bias is a significant concern for researchers, imprecision can also negatively impact knowledge production. Imprecise studies risk failing to detect small treatment effects and variations in effects (Mutz 2011) and may lead to overestimation or misinterpretation of effect sizes (Gelman and Carlin 2014; Loken and Gelman 2017). Given the structural incentives to publish studies with "positive" findings that meet conventional significance thresholds, published experiments based on noisy data can accrue and misrepresent evidence in support of certain theories (Gerber, Green, and Nickerson 2001; Kühberger, Fritz, and Scherndl 2014). Statistical imprecision is increasingly recognized as a major contributor to the low replicability rates observed across the social sciences (Arel-Bundock et al. 2022; Gelman and Carlin 2014). Thus, addressing statistical precision in survey experiments is crucial. As CSP argues, researchers must balance considerations of both bias and precision to when designing experiments.

Repeated measure designs offer one way to improve on standard post-only designs in terms of precision. In this type of experimental design, outcomes are measured both before and after exposure to treatment or control stimuli. By accounting for respondents' pre-treatment outcome measurement, repeated measures designs significantly enhance the

precision of treatment effect estimates. CSP demonstrates that this can substantially reduce the sample size required to achieve conventional levels of statistical power. Repeated measures designs come in two main types: within-subject, where all respondents receive both the treatment and control stimuli, and between-groups, where respondents are randomized to either treatment or control stimuli. Both approaches can offer substantial gains in precision.

Despite these advantages, many researchers have historically avoided repeated measures designs due to concerns about introducing bias in the ATE estimate. The three primary concerns for social scientists are: *priming*, where pre-treatment measures may lead respondents to focus on specific considerations (e.g., Klar, Leeper, and Robison 2020); *consistency*, where respondents may attempt or feel pressure to provide post-treatment responses that align with their previous 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).

Conventional wisdom thus suggests a trade-off between bias and precision when considering post-only versus repeated measures designs. In practice, most survey experiments in political science have prioritized concerns about bias over imprecision by defaulting to post-only designs. To our knowledge, however, CSP is the only study to date that empirically tests the bias-precision trade-off for repeated measure designs. Their meta-analysis of six experiments revealed no significant differences in estimated ATEs between the two designs. However, they found that repeated measures designs substantially improve precision, allowing researchers to achieve more with fewer participants. For instance, to achieve 80 percent power for detecting a treatment effect of 0.20 standard deviations, a post-only design would require around 1,000 respondents, whereas a repeated measures design would need only about 200 to 600 respondents, depending on the strength of the correlation between pre- and post-treatment measures. Given the substantial improvement in precision and minimal evidence of bias, CSP argue that there is no meaningful bias-precision trade-off and strongly

recommend that researchers employ repeated measure designs as the default.

Contribution

Since its publication in 2021, CSP has already received 198 Google Scholar citations. Of these, at least 88 were original studies referencing CSP to justify using repeated measure designs (see Appendix Table A.4.1 for the full list of studies). And while most citations to CSP are from political scientists, scholars in fields such as communication, criminology, economics, education, and environmental policy have also referenced CSP to justify using repeated measure designs. The broad influence of this single article on experimental practice is already quite clear, and is likely to grow in the coming years as disciplines become more critical of low statistical power in experiments amid the ongoing replication crisis (Arel-Bundock et al. 2022; Ioannidis, Stanley, and Doucouliagos 2017; Open Science Collaboration 2015).

CSP provide a valuable and overdue examination of the bias-precision trade-off in repeated measure designs. However, the existing empirical literature on this design choice remains limited, and several key questions about best practices remain unanswered. First, as indicated in Appendix Table A.4.1, 31 percent of studies citing CSP have utilized within-subject designs rather than between-groups repeated measure designs, only one of which appears in CSP's original study (a $N = 900$ replication of Smith's classic 1987 question wording experiment on welfare). This ultimately constitutes a fairly limited basis for such a large shift in empirical practice. Our study expands the evidence base by replicating three within-subject experiments (the same Smith 1987 study plus two question wording studies from Wilson et al. 2008 and de Benedictis-Kessner and Hankinson 2019) on each of our three omnibus surveys, for nine total studies with a combined sample size over 43 times larger than the single study analyzed by CSP. Simultaneously, we replicate three of CSP's between-groups experiments (one of which is itself a replication of Gilens 2001) on the same omnibus surveys to further expand the evidence base for between-groups repeated measure

deisgns. 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 has assessed how the proximity between repeated measures affects bias and precision. An intuitive hypothesis suggests that increasing the distance (i.e., adding more survey content) between repeated measures can reduce bias by mitigating priming and enabling respondents to "forget" their earlier pre-treatment responses, potentially reducing pressure (or ability) to provide consistent responses. Indeed, given this intuition, many studies with repeated measure designs employ multi-wave panel surveys, introducing days or weeks of separation between measures. This consideration is also reflected in CSP's own design choices to place their pre- and post-treatment questions at opposite ends of their surveys. 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 quite close together. Further, close proximity may even be advantageous if it reduces random noise and strengthens the correlation between pre- and post-treatment measures, which increases precision. We investigate the impact of distance between pre- and post-treatment measures by manipulating these distances in our surveys. Specifically, we conduct a preregistered test of the conventional wisdom that:

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

Third, CSP conducted their six experiments using two student samples and four online non-probability samples. Their findings are important given the reliance on such convenience samples in experimental research (Jerit and Barabas 2023; Krupnikov and Levine 2014), but we do not yet know if the absence of bias they observe also applies to probability-based

[1] Anonymized preregistration materials are available here.

sampling designs. Student samples have long been known to 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 population, but also less professionalized, less prone to satisficing, and more attentive than members of opt-in panels used in common non-probability samples (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 ($N = 965$, student sample) revealed that many respondents whose outcome response changed between measures also self-reported (inaccurately) that their opinions did not change. These inaccurate perceptions may be due simply to the unobtrusiveness of repeated measures—affirming the utility of such designs—but respondent inattentiveness or satisficing may also play a role. More attentive respondents may be more likely to recognize being asked the same (or very similar) questions twice and alter their post-treatment response accordingly. Similarly, highly professionalized respondents that constitute large shares of many online non-probability panels may be accustomed to repeated measure designs and react differently than less professionalized respondents would under the same conditions. We explore these possibilities by fielding our experiments on three distinct samples recruited from sample providers that use both probability-based and non-probability sampling designs. This enables us to evaluate how both vendor choice and respondent-level characteristics like professionalization and attentiveness affect bias-precision trade-offs in experimental design.

Data and Methods

We replicate six previously published survey experiments, summarized in Table 1, and randomly manipulate the experimental design of each (post-only vs. repeated measure).[2]

[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.

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 spending (from Gilens 2001), in which treated respondents are informed that foreign aid spending represents about 1 percent of the federal budget. We expect this information treatment to increase support. In Study 2, we replicate an original party cues experiment from CSP on policy support for allowing prescription drugs to be imported from Canada, in which treated respondents are given information that Democrats tend to support this policy and Republicans tend to oppose it. In this experiment, we analyze the second difference in support between Democrats and Republicans among those who were treated versus not treated. We expect the party cues treatment to increase support among Democrats and decrease support among Republicans, widening the gap in support between the parties. In Study 3, we replicate an original 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 positive framing 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), in which respondents are asked about their support for anti-poverty spending described as "welfare" or "assistance to the poor." We expect support to be higher when the policy is described as assistance to the poor, relative to welfare. In Study 5, we replicate a classic question wording experiment on support for affirmative action (from Wilson et al. 2008), in which respondents are asked about their support for affirmative action for women or 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 clinic's location 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 when the proposed clinic would be located further away.

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.

Table 1: Summary of Replicated Survey Experiments

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

These six studies were selected because they are each brief and have previously found large treatment effects[3], and could be appropriately fielded with post-only and repeated measure designs. We consciously selected replication studies to cover a range of 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 them with two additional experiments from Wilson et al. (2008, denoted as study 5 here) and de Benedictis-Kessner and Hankinson (2019, study 2, denoted as study 6 here) to increase the number of within-subject studies we could analyze. In total, we thus have three between-groups and three within-subject repeated measure designs, all of which can be compared

[3]While repeated measure designs are appropriate and even advantageous for studying small treatment effects given their increased power, we replicate studies that previously found large effects to avoid floor effects. When estimating the bias introduced by repeated measure designs, replicating experiments with small effects could obscure whether an apparent lack of bias is due to floor effects estimating small effects or a true null.
[4]Studies 1, 2, 3, and 4 described here correspond to studies 2, 5, 6, and 1 in CSP.

against otherwise equivalent post-only designs.

Experimental Design

We fielded all six studies on each of the three omnibus surveys and manipulated the experimental designs in a multi-stage, structured randomization procedure. All respondents in each sample (combined $N_i = 13,163$ respondents) completed all six experiments (combined $N_{ij} = 78,978$ observations). In the first randomization stage, we randomly selected two experiments (at the respondent level) to use 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 question wording experiments, this assignment dictated which question 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 experiments assigned to a repeated measure design. These four questions (the "pre-treatment" block) were displayed 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 in the fifth stage. In the full-random procedure, all sub-blocks in the post-treatment block appeared 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 appeared subsequently. 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.

We hypothesized that design effects might be more pronounced when repeated measures are placed close together on the survey. When repeated measures are close together, respondents should be more likely to remember answering the same or similar question, potentially strengthening any priming effect, consistency pressure, or demand incentive. To test this hypothesis, we fielded relatively more repeated measure designs and employed a complex randomization procedure to create a distribution of "distance" between pre- and post-treatment measures (defined as the amount[5] of survey content separating the repeated measures) that oversamples proximate distances. We also include two three-question sub-blocks of unrelated NFL content to expand the right tail of this distribution and provide additional distractor items.[6] With this randomization procedure, we generated the distribution shown in Figure 1.[7] This purposeful distribution allows us to test whether design effects are more pronounced when repeated measures are closer together, while also providing a long right tail to explore whether design effects may change non-linearly as the distance between measures increases.

The structured randomization procedure provides us with unbiased estimates of design effects while maximizing statistical power where we expected (a priori) that it would mat-

[5]We follow the Time-Sharing Experiments for the Social Sciences (TESS) guidelines for defining "units" of survey content as our operational measure of distance. For most experiments, the pre-treatment item and treatment/post-treatment items each count as one unit, but some experiments also include short paragraphs or additional attitude change question or covariate question (see Appendix B for details). In our surveys, each repeated measure experiment was separated by between 0 and 20 TESS units of distance.

[6]For this purpose, NORC bundled our AmeriSpeak survey with six unrelated questions about the NFL fielded by an uninvolved researcher. We maintained these unrelated items in the Lucid and Prolific samples. See B.3 for more details.

[7]While Figure 1 shows the pooled observations from all experiments and samples, the distribution is similar within each sample and within each experiment.

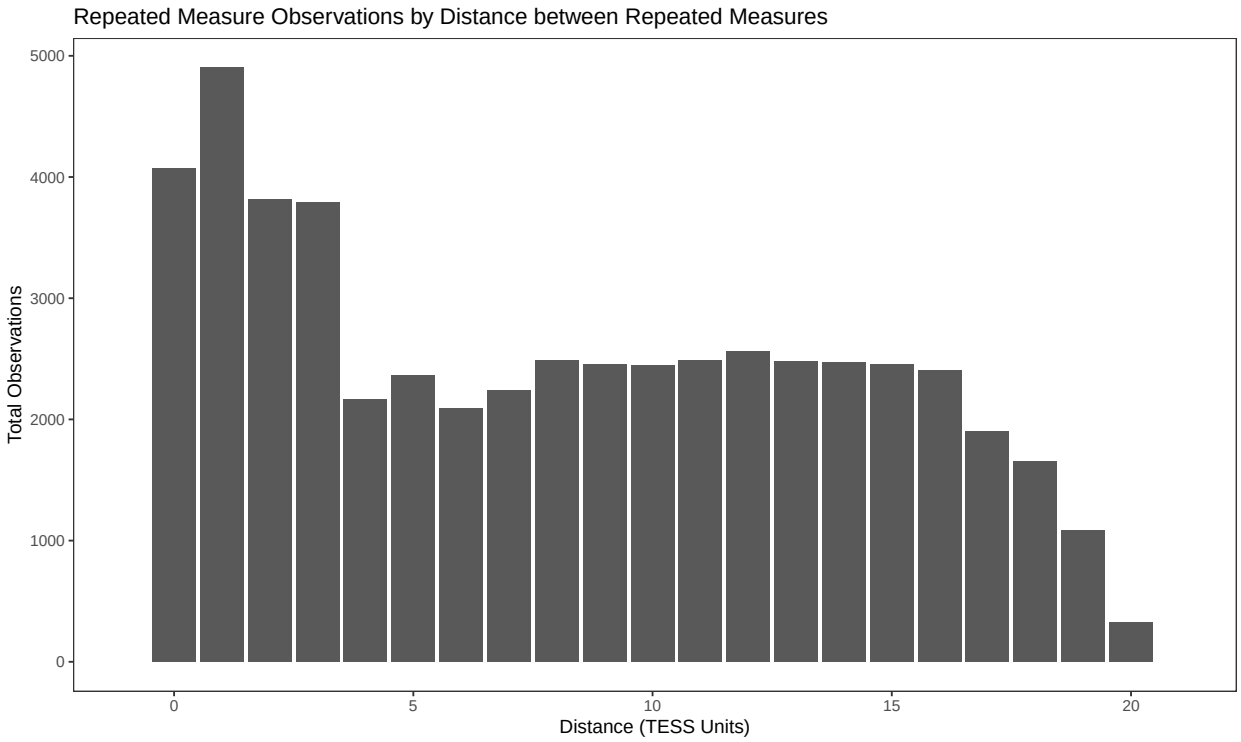

Figure 1: Histogram of distances between repeated measures. Figure shows the observed distances, in standard TESS units, separating the pre- and post-treatment measures for observations in the repeated measure design setting. Data includes pooled observations from all experiments in all samples.

ter most. By comparing point estimates for the ATE under the post-only versus repeated measure design, we can identify bias introduced from repeated measure designs (addressing H1). We can also identify precision gained from repeated measure designs by comparing standard errors for the post-only and repeated measure designs (using bootstrapped regressions with equivalent sample size 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 effectively test whether this proximity moderates the size of any design effects (addressing H2).

Sampling Approach

We fielded our experiments on three separate samples with concurrent omnibus surveys from June 27$^{\text{th}}$ through July 15$^{\text{th}}$, 2024. Building on CSP's original studies, which were

fielded exclusively with convenience samples drawn from undergraduate participant 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, attentiveness, and quality, as well as the credible representativeness of the sampling design (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 Methodology**	**Surveys per Month (Median R)**	**Panel Memberships (Median R)**	**Analysis Sample (N_i)**
1	AmeriSpeak	Probability-based	2	1	4,029
2	Prolific	Non-probability	30	2	4,261
3	Lucid	Non-probability	15	4	4,869

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, if any; however, a small minority of Americans take many surveys frequently for income or entertainment (Hillygus, Jackson, and Young 2014). Professionalized respondents constitute an out-sized share of non-probability panels like Prolific and Lucid because high-propensity respondents can voluntarily sign up to join such panels and take surveys on demand. In contrast members of probability-based panels like AmeriSpeak can only join if randomly sampled and recruited, and the organizations that manage such panels invite panelists to take only select surveys on an occasional basis. We find that our AmeriSpeak respondents are much less professionalized than our Prolific and Lucid respondents, as evidenced by the number of surveys they have taken in the last 30 days and the number of unique survey panels they recently participated in (Table 2).

Professionalization may induce survey respondents to react differently to repeated measure experiments through regular exposure, raising the possibility that such designs may bias estimated ATEs in less professionalized samples. More professionalized respondents could be less affected by treatments through previous exposure to similar content; alternatively, less professionalized may feel greater pressure to provide more consistent answers. Demand effects could be more pronounced in respondents who are more familiar with surveys, as they may be more attuned to the study's purpose and adjust their responses accordingly, particularly if they are motivated by financial incentives. Alternatively, professionalization may inure respondents to the use of repeated measures, dampening the risk of demand effects.

The second relevant dimension is response quality, which we define as respondent attention and sincere effort. A perennial issue in survey research is that respondents do not always pay close attention or put much effort into their responses, introducing random noise at best but possibly also introducing bias (Berinsky et al. 2021). Issues of response quality 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 opt-in panels also typically provide monetary compensation, some participants engage in extreme satisficing or speeding to earn revenue as quickly as possible (Hillygus and LaChapelle 2022), and can use generative AI and other automated tools to do so (Veselovsky et al. 2023; Veselovsky, Ribeiro, and West 2023). Some participants also complete surveys mostly for entertainment, sometimes with the express intent to provide phony responses and troll researchers (Lopez and Hillygus 2018). In repeated measure designs, respondents who are less attentive may still be subject to issues like priming, consistency, and demand effects but their lack of engagement might reduce the likelihood or strength of these biases, while troll respondents may provide reactions that differ from higher-quality respondents.

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) to take

surveys and weeding low-quality panelists. Other vendors largely leave it to the researchers to manage quality control. 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 for details) 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 analysis sample after we exclude 38 respondents who failed at least two. Lucid respondents failed an average of 0.684 checks, which falls to 0.279 in the analysis sample after we exclude 681 who failed at least two. Thus, the Lucid respondents tend to be less attentive and effortful than Prolific respondents, variation which we can exploit to test whether these respondent characteristics affect the performance of repeated measure designs.

In summary, our study shares important similarities with, but also provides key advances on, CSP's original evaluation of repeated measures designs. Like CSP, we replicate six survey experiments to test if repeated measure designs introduce design effects (i.e., attenuation or exaggeration of the ATE). We replicate four studies from CSP and add two additional question wording experiments from the political science literature. All six experiments are fielded on three separate omnibus surveys, yielding a total of 18 studies with a combined $N_{ij} = 79,978$—nearly ten times larger than CSP's combined samples from their six studies. Our larger samples not only provide greater power to detect small design effects, but also enables us to test for potential heterogeneity in design effects on across several critical design considerations: experiment type (between-groups or within-subject), the relative proximity of repeated measures, and vendor sampling designs and consequent respondent characteristics. Our study thus provides both well-powered tests of CSP's influential claims and

[8]NORC discourages attention checks, out of concern that AmeriSpeak respondents are unaccustomed to the practice and may discontinue participation. Instead, we exclude AmeriSpeak respondents who skipped more than half of the questions (17 respondents) or finished in under one-third of the median time (135 respondents). We assume that the vast majority of the remaining respondents are high quality, given the quality metrics we have (e.g., completion times) and NORC's rigorous recruitment and management for AmeriSpeak panelists.

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 with a series of internal meta-analyses using our 18 individual experiments. We then examine potential heterogeneity in design effects on several key dimensions.

Summary of Experimental Results

For each of the six experiments, 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, 2, and 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 variable as a function of a binary treatment indicator, with the pre-treatment outcome included as a covariate in the repeated measures design[9] We then combine these regressions via seemingly unrelated regression estimation and conduct a linear combination test for equivalence between treatment coefficients across the two designs.

For the within-subject experiments (Studies 4, 5, and 6), we compare the difference in ATEs using random effects models. These models regress the dependent variable on an indicator for treated observations interacted with an indicator for whether the observation occurs in a repeated measure setting, clustering standard errors at the respondent level. The coefficient on the interaction term estimates the difference in ATEs between the designs.

As preregistered, we follow prior authors' inclusion of specific covariates (such as parti-

[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.

sanship, ideology, etc.) in the estimation 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.

Study 1: Foreign Aid

In this between-groups experiment, we regress support for foreign aid spending on a treatment indicator for whether the respondent received an informational treatment noting that foreign aid spending is currently about 1% of the federal budget. Following CSP, we include partisanship and ideology as covariates. In all three samples, we replicate CSP's finding (and that of Gilens 2001, etc.) that the information treatment 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 repeated measure design for all three samples, although this difference is significant only in the Prolific sample ($p = 0.002$); the estimated design effect is smaller and not significant in both 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 whether the respondent received a party cues treatment, interacted with an indicator for identification with the Democratic party (we exclude true independents from this analysis). The coefficient on the interaction term thus provides a measure of polarization in attitudes between the parties. In all three samples, we replicate CSP's finding that the party cues treatment increases attitude polarization between the parties in both the post-only and repeated measure designs.[10] We again find an attenuation of this treatment effect in the repeated measure design for all three samples, although this difference is signif-

[10]The repeated measure design allows us to directly test how party cues polarize attitudes for specific partisan subgroups. We measure polarization as the difference in standard deviations of pre-post change scores in treatment vs. control conditions, where larger differences indicate greater polarization. Party cues markedly polarize strong partisans' views ($\Delta_\sigma = 0.058$), modestly polarize weak partisans' views ($\Delta_\sigma = 0.047$), and only weakly polarize partisan-leaning independents' views ($\Delta_\sigma = 0.031$).

Table 3: Summary of Experimental Results

		Post-Only	*Repeated*	*Design Effect*		
Experiment	Sample	Est. ATE	Est. ATE	Estimate	SE	Δ in ATE
Foreign Aid	AmeriSpeak	0.089^{***}	0.065^{***}	-0.023	0.015	-26.1%
	Prolific	0.111^{***}	0.068^{***}	-0.043^{**}	0.014	-38.6%
	Lucid	0.063^{***}	0.056^{***}	-0.008	0.016	-12.0%
Drug Imports	AmeriSpeak	0.125^{***}	0.055^{***}	-0.070^{*}	0.029	-56.0%
	Prolific	0.096^{***}	0.072^{***}	-0.024	0.032	-25.1%
	Lucid	0.110^{***}	0.077^{***}	-0.033	0.032	-30.4%
GMOs	AmeriSpeak	0.162^{***}	0.129^{***}	-0.033^{*}	0.017	-20.2%
	Prolific	0.180^{***}	0.162^{***}	-0.017	0.016	-9.7%
	Lucid	0.144^{***}	0.124^{***}	-0.021	0.016	-14.2%
Anti-poverty	AmeriSpeak	0.202^{***}	0.159^{***}	-0.044^{*}	0.020	-21.3%
	Prolific	0.165^{***}	0.110^{***}	-0.055^{***}	0.017	-33.1%
	Lucid	0.169^{***}	0.135^{***}	$-0.033^{\dagger}$	0.019	-20.0%
Affirm. Action	AmeriSpeak	0.095^{***}	0.079^{***}	-0.017	0.022	-17.3%
	Prolific	0.094^{***}	0.053^{***}	$-0.040^{\dagger}$	0.022	-43.2%
	Lucid	0.094^{***}	0.079^{***}	-0.014	0.020	-15.0%
Opioid Clinic	AmeriSpeak	0.113^{***}	0.101^{***}	-0.012	0.018	-10.8%
	Prolific	0.071^{***}	0.126^{***}	0.055^{**}	0.019	$+76.5\%$
	Lucid	0.047^{**}	0.050^{***}	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.

icant only in the AmeriSpeak sample ($p = 0.017$); the estimated design effect is smaller and 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 treatment message (as compared to an anti-GMO message). Following CSP, we include partisanship and ideology as covariates. In all three samples, we replicate CSP's finding that the positive framing treatment increases support for GMOs in both the post-only and repeated measure designs. We again find an attenuation of this treatment effect

in the repeated measure design for all three samples, although this difference is significant only in the AmeriSpeak sample ($p = 0.049$); the estimated design effect is smaller and 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 public spending to address poverty on an indicator for whether these efforts are described as "assistance to the poor" (1) versus "welfare" (0), interacted with an indicator for whether measurement was taken in the two-question repeated measures setting (1) versus the single-question post-only setting (0). Following CSP, we include partisanship and ideology as covariates. In all three samples, we replicate CSP's finding (and that of Smith 1987, etc.) that support for spending is greater when the policy is described as "assistance to the poor," in both the post-only and repeated measure designs. We again find an attenuation of this treatment effect in the repeated measure design for all three samples. This negative design effect is strongest in the Prolific sample ($p = 0.001$), but still large in the AmeriSpeak sample ($p = 0.025$) and narrowly insignificant in the Lucid sample ($p = 0.071$).

Study 5: Affirmative Action

In this within-subject experiment, we regress support for affirmative action policies on an indicator for whether the policies are aimed at women (1) versus racial minorities (0), interacted with an indicator for whether measurement was taken in the two-question repeated measures setting (1) versus the single-question post-only setting (0). In all three samples, we 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 again find an attenuation of this treatment effect in the repeated measure design for all three samples, although this design effect is significant (at the 0.10 level) only in the Prolific sample ($p = 0.071$); the estimated design effect is smaller and 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 to address opioid addiction on an indicator for whether the clinic would be located a quarter mile away (1) versus two miles away (0), interacted with an indicator for whether measurement was taken in the two-question repeated measures setting (1) versus the single-question post-only setting (0). In all three samples, we 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 large positive design effect (exaggerating the treatment effect) from the repeated measure design in the Prolific sample ($p = 0.004$), but the estimated design effects are much smaller and not significant in the AmeriSpeak (negative estimate, $p = 0.481$) and Lucid (positive estimate, $p = 0.821$) samples.

Repeated Measure Designs Cause (Slight) Attenuation Bias

Across six experiments replicated in three large samples, we estimate nearly every design effect to be negative. These differences in the estimated ATE are often large, as shown in the final column of Table 3, and we observe a median 20.1 percent reduction in the ATE (from repeated designs relative to the post-only design) across the 18 estimated design effects. The consistent pattern in the estimated design effects suggests that there may be a true consistent attenuation of treatment effects from repeated measure designs that individual experiments are not well-powered to detect. We conduct a series of preregistered internal meta-analyses to test this possibility. We first rescale the design effect and standard error in each experiment as the proportional change from the post-only design's ATE (that is, a 20.1 percent attenuation of the ATE is a design effect of -0.201).[11] We then meta-analyze: all six experiments, the three between-groups experiments, and the three within-subject experiments, both within each sample and across all three samples. The results are shown

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

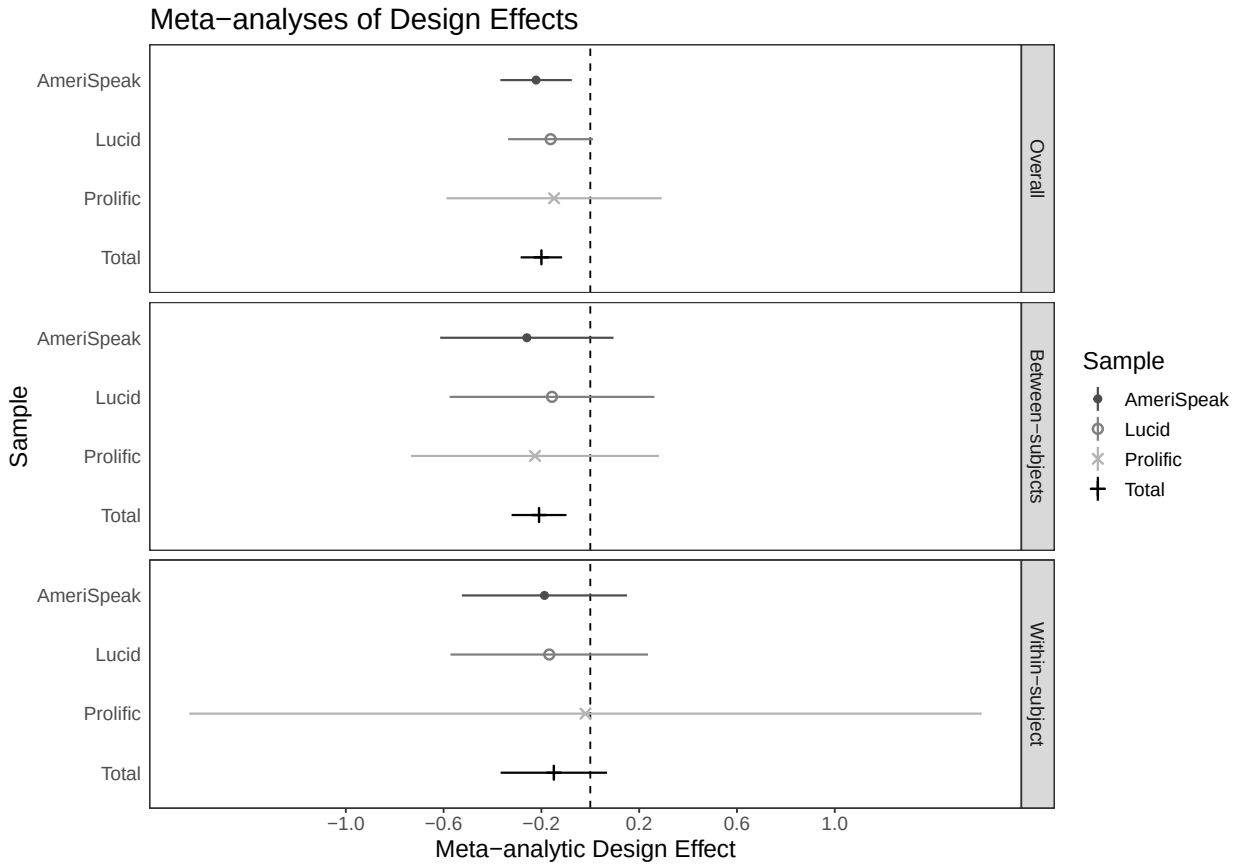

Figure 2: Internal Meta-Analyses. Figure displays estimated design effects from internal meta-analyses of experiments within each sample and across all three samples.

in Figure 2 and provided in tabular form in Appendix A.1.

When analyzing the 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 ($p = 0.011$), with slightly smaller meta-analytic design effects in the Prolific (estimate -0.148, $p = 0.426$) and Lucid (estimate -0.162, $p = 0.061$) samples. Across all three samples, we find a precisely estimated meta-analytic design effect of -0.200 ($p < 0.001$, 95 percent CI $= [-0.285, -0.115]$) when meta-analyzing all 18 experiments—that is, we would expect a 20.0 percent attenuation of the ATE with a repeated measure design 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 single sample,[12] the between-groups estimate across samples is very similar and 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 prescription drug experiment only. A meta-analysis of the within-subject experiments excluding this single outlier provides very similar results to the between-groups experiments (design effect estimate -0.227, $p = 0.003$).

Repeated Measure Designs Increase Statistical Power

Although we find some evidence of attenuation of treatment effects in repeated measure designs, CSP note that repeated measure designs also offer significant gains to precision, and may still be preferable for that reason. Because our experimental design assigns each respondent to complete twice as many repeated measure experiments as post-only experiments (and thus produces roughly twice as many repeated measure observations for each individual experiment), a direct comparison of the standard errors under each design for each experiment would artificially privilege the precision of the repeated measure design. We therefore re-estimate the results of each experiment under each design via a bootstrapping procedure that uses samples of identical size across 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 of observations (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 provides an estimate of the relative precision across experimental designs for samples of the same size. These estimates are provided in Table 4, which shows the pooled ATE and standard errors under each design, as well as the percentage change in standard error and root mean squared error (RMSE)

[12] 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, $p = 0.088$).

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%

†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, estimated with 1,000 runs of equivalent sample size under each design.

that the repeated measure design provides.

As Table 4 shows, we find that repeated measure designs provide large gains to precision with the same sample size, affirming CSP's findings of the same. We observe a median 49.4 percent reduction in the standard error across all 18 experiments, and in every experiment we observe a reduction of at least 31.1 percent or more. Similarly, we observe a median 41.0 percent reduction in the RMSE, with a minimum observed reduction of 28.6 percent. These consistently large reductions in standard error and RMSE confirm that repeated measure designs offer significant improvement in statistical precision relative to post-only designs. As we detail in the Discussion, this major advantage of repeated measure designs generally outweighs the disadvantage of slight attenuation of the estimated treatment effect in most experimental 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 longitudinal surveys—to minimize the probability that respondents will recall being previously asked the same question and change their post-treatment response through priming, consistency, or demand effects. Given our finding that repeated measure designs slightly attenuate treatment effects on average, one might reasonably be concerned about using such a design on a short survey or module, in which the proximity between measures might heighten respondents' recall and induce greater bias in the estimated ATE. Because our experimental design randomly varies the distance between pre- and post-treatment measures in the repeated measure design experiments, we can test for moderation by the design feature of distance explicitly. We estimate a series of ATEs at each discrete distance between pre- and post-treatment measures (in discrete TESS units, ranging from 0 to 19)[13] for each experiment in each sample. We standardize these estimated

[13] Although we observe distances as high as 20 TESS units between repeated measures, we typically have very few observations at the largest possible distance for each experiment (between 4 and 52 observations). We therefore exclude the estimated ATEs for the most extreme value of distance for each experiment.

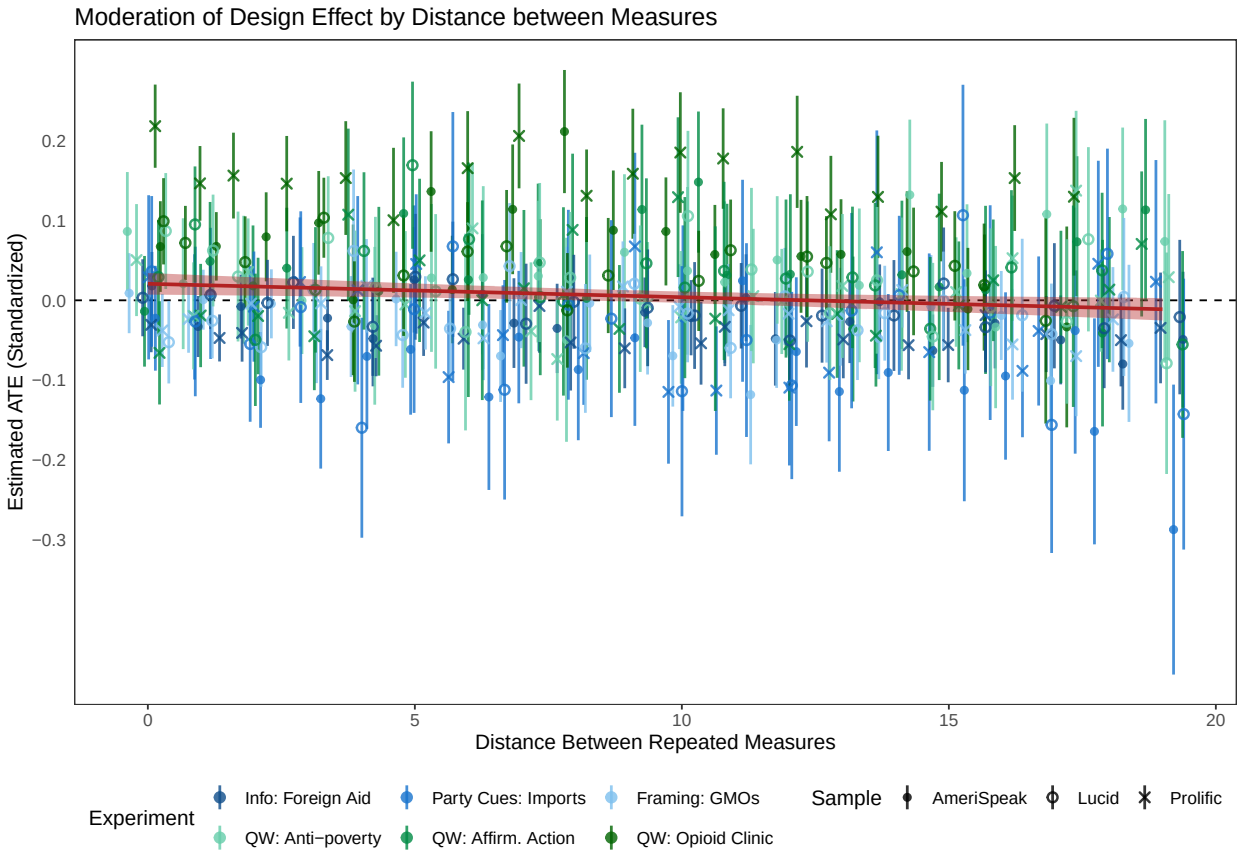

Figure 3: Estimated design effect by distance between repeated measures. Figure displays the estimated ATE at each distance between pre- and post-treatment measures (in discrete TESS units, x-jittered for visual clarity) in each experiment in each sample, standardized to the respective observed post-only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

ATEs relative to the post-only ATE observed for each study (see the first column of Table 3) and regress the standardized ATEs on the distance measure. Figure 3 shows the standardized ATEs and associated 95 percent confidence intervals for each experiment at each degree of separation; the linear regression is indicated by the red line, and the shaded areas indicate the 95 percent confidence interval.[14]

[14]Note that these analyses deviate from our preregistered intent to estimate spline regressions to assess potentially non-linear effects of distance on the design effect. We estimated a spline regression (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.

We find that the effect of distance between repeated measures is detectable but substantively small, as the red regression line in Figure 3 suggests. Each additional TESS item separating the pre- and post-treatment measures is estimated to attenuate the repeated measure ATE by -0.002 ($p = 0.010$) on average, or by about 1.4 percent of the mean ATE in our data. When we include fixed effects for the sample and experiment, we find very similar but more precise results: 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.[15] The slight influence of distance on the overall design effect suggests that repeated measure designs are about as well suited to shorter surveys and close placement as to separating the measures by several minutes within a single survey.

No Moderation by Respondent Professionalization

Many scholars today use non-probability sample providers for experimental research because of their convenience and relatively low costs (Jerit and Barabas 2023). Survey participants provided by online non-probability panels can typically complete many surveys on a regular basis, as a source of income or simply for personal enjoyment; respondents from these providers are thus both fairly professionalized and prone to satisficing to complete surveys quickly (Hillygus, Jackson, and Young 2014; Hillygus and LaChapelle 2022). In contrast, NORC's probability-based AmeriSpeak panel restricts participation frequency to keep respondent quality high and avoid excessive professionalization, meaning that respondent attention may be higher on average in this sample.

These differences in probability versus non-probability respondent pools could affect the design effect of repeated measure designs in several ways. Increased respondent attention

[15] This estimated effect is sufficiently small that it may best be considered negligible (Rainey 2014). Additionally, we note one potential threat to the inference that attenuation bias increases with separation between measures. Because we bundled several experiments together with a randomized order, observations with high distance between measures necessarily means that those same respondents encountered other repeated measure experiments in between—and previous exposure to this experimental design (even on other substantive topics) may have altered their response patterns on later observations, in a way that additional non-experimental distractor content would not. We therefore encourage readers to consider even our small but statistically significant moderation effect with the appropriate caution.

could increase recall of a pre-treatment measure or response, and therefore elevate consistency, priming, or demand pressures on post-treatment responses. Satisficing and speeding could reduce recall of a pre-treatment measure and suppress these same pressures, but could also reduce exposure to treatment (Hillygus, Jackson, and Young 2014). Increased professionalization may inure respondents to the use of repeated measures, improving their effectiveness. Conversely, higher rates of respondent trolling (Lopez and Hillygus 2018) by respondents recruited from less well-managed panels could exacerbate or alter demand effects in a repeated measure setting.

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 measure experiments. At the end of each survey, we asked respondents to indicate 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, we find that our Prolific and Lucid respondents are much more professionalized than the AmeriSpeak panelists: the median respondent in the AmeriSpeak sample reported completing just 2 surveys for 1 survey panel in the past 30 days, whereas the median Prolific respondent reported completing 40 for 2 panels and the median Lucid respondent reported completing 17 surveys for 4 panels.[16]. 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, we find similar design effect sizes above and below the median within each sample on both professionalization measures. That is, our data suggests that respondent professionalization does not exacerbate or mitigate the design effect of repeated measure experiments. This result using individual-level measures of profession-

[16]For these analyses, we pre-registered excluding respondents who reported completing more than 1,000 surveys or working with more than 100 companies in the past 30 days, as these responses are likely not genuine.

alization helps explain why the design effects are similar across three samples with large differences in respondent professionalization.

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) responses after being exposed to the treatment. In their original pre-post study on GMOs, CSP asked whether respondents' support for GMOs had changed since earlier in the survey—that is, between the pre- and post-treatment measures. CSP found that 40.5 percent of respondents provided different answers on the two measures, while 58.8 percent reported that their attitudes had remained stable. CSP concluded that respondents may struggle to provide consistent responses in repeated measures studies, even if they feel pressure to do so, because many cannot recall their earlier responses. This, they argued, reduces the risk of attenuation bias in repeated measures designs.[17]

For our three between-subject experiments, we followed the post-treatment measure with 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 the previous single study).[18] Specifically, we asked whether the respondent's preferences about the relevant issue had changed since being asked about the same issue earlier in the survey; respondents could indicate whether their support had decreased, increased, or stayed about the same.[19] 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). Like CSP, we find that most respondents (69.9 percent) provide the same response both pre- and post-

[17] 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 potential fuzziness about pre-treatment question itself.

[18] Because the question wording experiments ask about plausibly different quantities (or quantities that are perceived to be different), attitude change is not conceptually well-grounded in that setting.

[19] See Appendix B.3 for the exact question wording and response options.

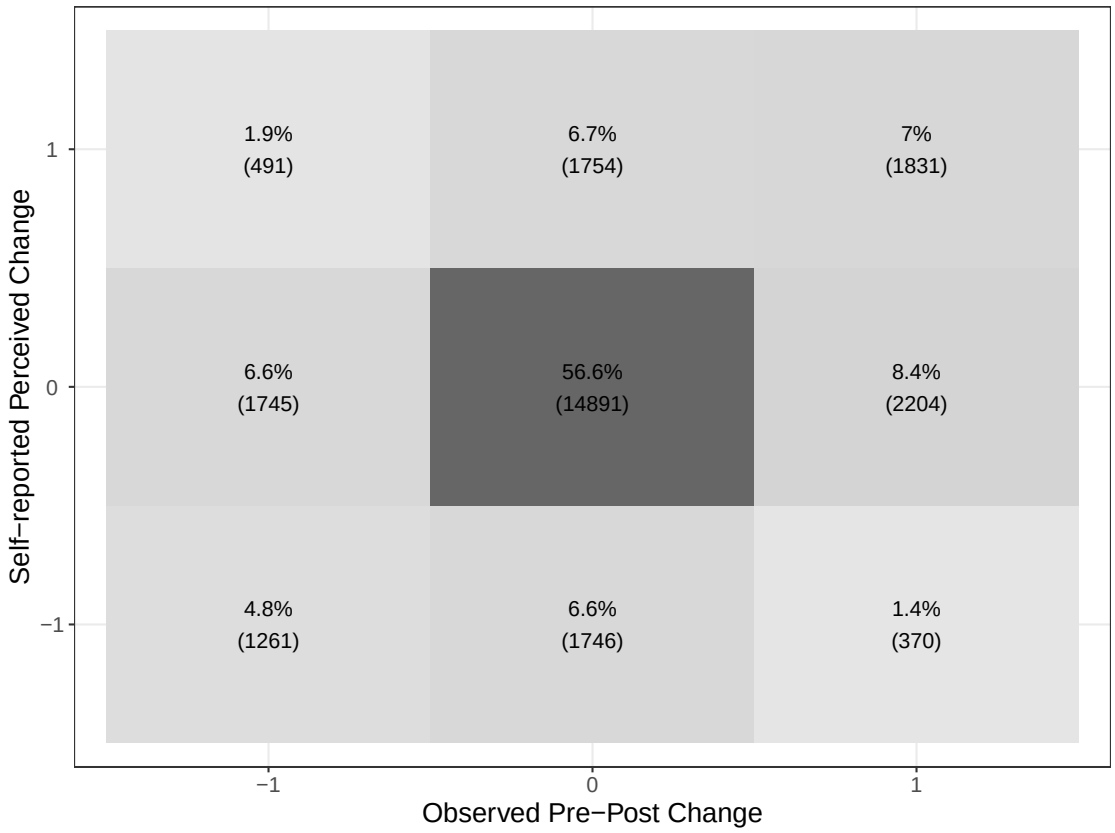

Figure 4: Respondent perception of attitude change. Figure shows a contingency table of pooled respondents (all between-groups repeated measure observations in all samples) by actual observed pre-post change in responses (columns) and self-reported perceived pre-post change (rows). Frequency counts for each cell are shown in parentheses.

treatment, as shown in the center column of Figure 4, 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.

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 subsets: those who accurately perceived their level of attitude change (that is, the three

cross-diagonal cells from the bottom left to top right in Figure 4) versus those who did not (all other cells).[20] We then re-estimate the design effect (as proportional change in the post-only design ATE) with each subset for each between-groups experiment in each sample, and finally meta-analyze these estimated design effects for accurate versus inaccurate respondents.

As shown in Table A.3.1, in all but one case we find that the attenuation bias is stronger among respondents who accurately perceived their level of pre-post change than those who did not, and often substantially so. Indeed, the design effect is often minimal and insignificant among those who fail to accurately assess their own level of change.[21] Across all nine between-groups experiments, the meta-analytic design effect among the accurate respondents is -0.538 ($p < 0.001$), but only an insignificant -0.020 ($p = 0.863$) among inaccurate respondents. These findings suggest that respondent attention—and the relative presence of consistency pressures—is a potential mechanism for producing the slight average attenuation bias we find in repeated measure designs.

That said, we caution readers against taking these findings as definitive evidence of the role of attention or consistency pressures in shaping design affects. 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 may partially de-randomize the assignment to treatment. This selection effect could

[20] 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.

[21] Though it should be noted that only 31.6 percent of respondents are classified as inaccurate, reducing statistical power for the estimation of design effects among these respondents.

artificially increase the design effect among recall-accurate respondents (that is, more attenuation relative to the post-only ATE) and decrease the design effect among recall-inaccurate respondents (closer to the post-only ATE), producing the same pattern of results we actually observe. Because we cannot distinguish between any mechanical effect from potential psychological effects at work, we offer this analysis merely as suggestive but not conclusive evidence that respondent attention and recall of their pre-treatment response contributes to the attenuation of treatment effects under repeated measure designs.

Discussion

Our study provides critical new evidence on the merits of repeated measure designs for experimental research. As in 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 standard errors across our 18 studies. Contra CSP, however, we also 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 traditional post-only design. Figure 5 summarizes the balance of our evidence on this fundamental trade-off. Given these findings, how then should experimental researchers proceed? And how should researchers weigh different types of experimental manipulation, sample provider, measure separation, and respondent attention when considering repeated measure designs? In this section, we assess our evidence and provide several practical recommendations to experimental researchers.

Repeated Measure Designs are Superior for (Most) Applications

Our experiments provide robust evidence that repeated measure designs reliably attenuate treatment effects.While post-only designs have the advantage of offering unbiased treatment estimates—as the absence of a pre-treatment measure allows randomization to guarantee unbiasedness in the outcome measure—a casual reader might infer that post-only designs

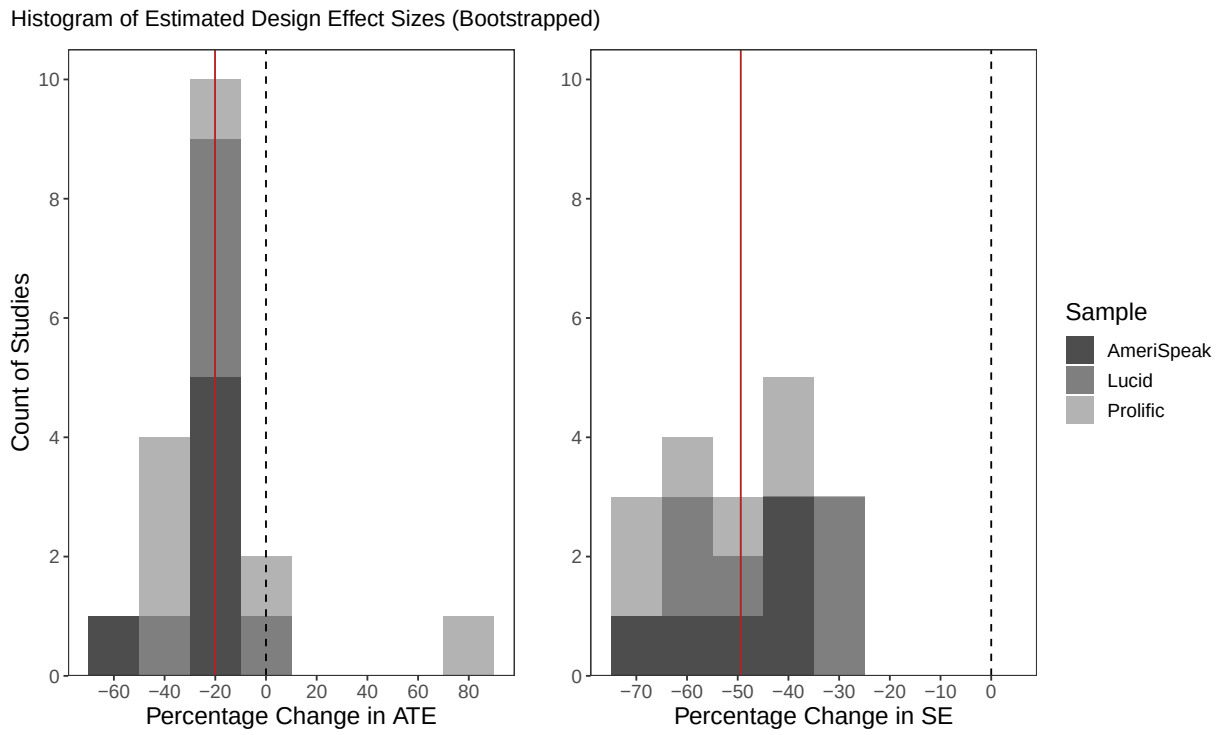

Figure 5: Histogram of design effects. Figure displays a histogram of observed design effects in terms of percentage change in estimated ATE (left panel) and standard error (right panel) in bootstrapped models with equal sample size across designs. The solid red line in each panel indicates the median percentage change in each statistic across all 18 experiments.

should be preferred over repeated measure designs. Yet survey experimental researchers are rarely concerned with identifying the exact value of the ATE; much more frequently, they aim to identify whether a given treatment moves the outcome variable (versus a null effect) and, often, whether the effect is in the hypothesized direction. For these purposes, statistical power is especially relevant, and in this regard, repeated measure designs clearly dominate. Even though repeated measure designs typically attenuate treatment effects to a slight degree, the substantially improved precision that these designs offer means that experiments with repeated measures are more likely to detect a true (albeit attenuated) effect in most scenarios.

Further, even when researchers are concerned with the actual magnitude of the ATE, repeated measure designs are still usually superior. For a given sample size, the gains to

precision mean that the estimated ATE of a repeated measure experiment is likely to be closer to the true value of the treatment effect despite some expected attenuation. That is, while post-only designs are unbiased in expectation, their relative imprecision means that actual estimated ATEs are likely to vary further from the true value of the ATE.

An exception, in which a post-only design may still be preferable, is for experiments with large samples and strong treatment effects expected a priori, such that the expected attenuation shift in the ATE from a repeated measure design is large enough—and the expected standard error in a post-only setting is small enough—that the post-only estimate will be closer in expectation to the true value of the treatment effect. To illustrate these considerations, we simulated 100,000 estimated ATEs under each design for a moderately-sized "true" treatment effect of $d = 0.200$, taking our observed 20 percent attenuation of the ATE and 50 percent reduction of the SE in the repeated measure design as ground truth. If the post-only treatment coefficient's standard error is 0.100 (sample size of ~ 400 observations), the post-only design results in a unbiased mean estimated ATE of 0.201 but the mean error from the true value is 0.079 and the estimated ATE is not significant 48.3 percent of simulations (power of just 0.517). The equivalent repeated measure design provides an attenuated mean estimated ATE of 0.160, but the mean error from the true value is nevertheless lower at 0.052 and the estimated ATE is significant in all but 10.7 percent of simulations (power of 0.893). In contrast, under the same conditions but with a sample roughly four times larger (~ 1600 observations) such that the post-only treatment coefficient standard error is 0.050, the post-only design's estimated ATE provides a lower mean error from the true value (0.040 versus 0.041) with statistical power that is nearly as strong as the repeated measure design (0.979 versus ~ 1.000).

In our view, because treatment effects 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, even researchers with access to large samples are likely better off with a repeated measure design. We therefore

concur with CSP in recommending that researchers employ repeated measure designs as the default in most practical applications.[22]

Repeated Measure Designs are Suitable for Both Between-groups and Within-subject Experiments

One of our aims was to increase the available evidence on repeated measure design effects for within-subject experiments. CSP's useful initial evidence comes from just one study on anti-poverty spending conducted on a student sample ($N = 900$). We analyze three within-subject question wording experiments (on anti-poverty, affirmative action, and opioid treatment policy) on three samples for a total $n_{ij} = 39,489$ across nine studies, providing robust evidence on the potential risks and benefits of repeated measure designs for these types of experiments. While we find that within-subject experiments—like between-groups experiments—are susceptible to some slight attenuation bias with repeated measure designs, we find that this bias is (if anything) smaller for within-subject 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 among 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 that researchers use repeated measure designs for both within-subject and between-groups experiments.

Repeated Measure Designs are Suitable for Probability and Non-probability Samples

Fielding all six of our experiments on three samples simultaneously, which vary in sampling design and respondent characteristics, allows us to assess the suitability of repeated measure designs for probability and non-probability samples. We observe some expected dif-

[22] An important second exception to this guidance concerns experiments with particularly sensitive topics or treatments, as we discuss below. Our experiments offer some variance in terms of their sensitivity, but we currently lack robust evidence on how repeated measure designs fare for experiments on especially sensitive topics.

ferences in respondent characteristics between the three samples, such as higher respondent professionalization among the non-probability samples and variation in respondents' ability to recall their pre-treatment responses (see Appendix A.3). Nevertheless, we find no consistent differences between the three samples in terms of overall design effects from repeated measure experiments, as shown in Figure 2. We further find that respondent professionalization does not have a major impact of design effect estimates within each sample (see Appendix A.2). We recommend that researchers use repeated measure designs with both probability and non-probability samples, with a caveat that attenuation bias may increase slightly as respondent attentiveness increases.

Repeated Measure Designs are Suitable for Brief Survey Modules

Common practice in experimental survey research is to place repeated measures as far apart as possible to enable respondents to "forget" their pre-treatment measurement or response, thus minimizing the risk of bias to the ATE. In our repeated measure designs, we randomly varied how early the pre-treatment measure appeared in the survey and how late the (treatment and) post-treatment question appeared, enabling us to assess how the separation between repeated measures moderates any design effects. We 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. Rather than exacerbating a design effect bias, placing pre- and post-treatment measurements very close together appears to have substantively similar results as placing them several minutes apart.[23] While our view is that researchers should still consider at least some distractor content between repeated measures (as we discuss below), we recommend that researchers use repeated measure designs even when constrained to very limited survey space that precludes providing much separation, and even when pre- and post-treatment measures

[23]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.

must be placed back-to-back.

Consider Distractor Content for Repeated Measure Experiments

We find tentative evidence that respondents' ability to recall their pre-treatment attitude may exacerbate the slight attenuation bias of repeated measure designs (see Table 4). For this reason, when possible we encourage experimenters to consider placing some unrelated content between the pre- and post-treatment measures to distract respondents' attention away from the previously measured concept, thus reducing their ability to accurately recall their previous response when completing the post-treatment measure. Given the small impact that we observe from separating repeated measures and the potential advantages of temporarily redirecting respondent attention prior to treatment, we still consider it best practice to place repeated measures further apart when possible—though we do not view this as strictly necessary, as discussed above.

When to Prefer Post-only Designs

Our evidence suggests that there are at least two circumstances in which a post-only design may be preferred. The first case is when researchers are especially concerned with identifying the precise magnitude of a treatment effect, not simply its presence or direction. Even here, however, experimenters should prefer post-only designs only if the expected magnitude of the ATE and the sample size are both sufficiently large such that the typical attenuation bias of a repeated measure design would outweigh the gains to precision and push the estimated ATE further away from the true value in expectation. Still, if the approximate magnitude of the treatment effect is not well known a priori or is not large, or if a large sample is not feasible to obtain, researchers are likely better served by a repeated measure design.

The second possible case for preferring a post-only design is when the experimental addresses an especially sensitive topic, such that that social desirability or similar pressures

substantially elevate the risk of consistency pressures or demand effects that could jeopardize the ATE estimate in a repeated measure design.[24] While our six experiments vary to some degree in terms of their sensitivity—the affirmative action and opioid clinic experiments might be considered more sensitive than the others, for example—our experiments are generally not as sensitive as (say) the topics often examined in list experiments (e.g., Aronow et al. 2015; García-Sánchez and Queirolo 2021; Redlawsk, Tolbert, and Franko 2010; Walsh and Braithwaite 2008). On especially sensitive topics, pre-treatment measurement of outcome variables may substantially heighten social desirability biases that could induce respondents to falsify their post-treatment responses (either to be more consistent or more responsive to the treatment, depending on the experiment). 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. Given that we find evidence of at least some attenuation bias in repeated measure experiments, researchers should still be cautious about employing repeated measure designs for experiments on very sensitive topics.

Concluding Remarks

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

Second, our 18 studies cover a lot of ground but necessarily leave much unexplored. In particular, our omnibus surveys remain relatively short (which is both a feature and a bug) and exclusively use the self-administered web survey mode. Repeated measure designs

[24] We note that CSP also acknowledge this potential exception (2021, 1062).

in other survey experimental contexts, such as face-to-face interviewing, may face other challenges that we cannot examine here. Nevertheless, because self-administered web surveys are quite common in experimental research (Jerit and Barabas 2023), we hope 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 (nearly) full-throated endorsement of this new standard for experimental design. There remain some circumstances in which the research aims can reasonably justify a traditional post-only design as preferable, but these are relatively rare in the discipline today. Our accumulated evidence suggests that the burden of justifying an experimental design choice should weigh more heavily on the use of post-only over repeated measure designs, rather than the historical reverse.

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.

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.

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.

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.

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.

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.

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

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

Lopez, Jesse and D. Sunshine Hillygus. 2018. Why So Serious?: Survey Trolls and Misinformation. Technical report SSRN.
URL: *https://dx.doi.org/10.2139/ssrn.3131087*

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.

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.

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.

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

Online Appendix to

Repeated Measure Designs are Superior for (Most) Experimental Survey Research Applications

Contents

A Supplemental Results

A.1 Internal Meta-analyses

In Table A.1.1, we report the tabular results of the 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.

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†	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.

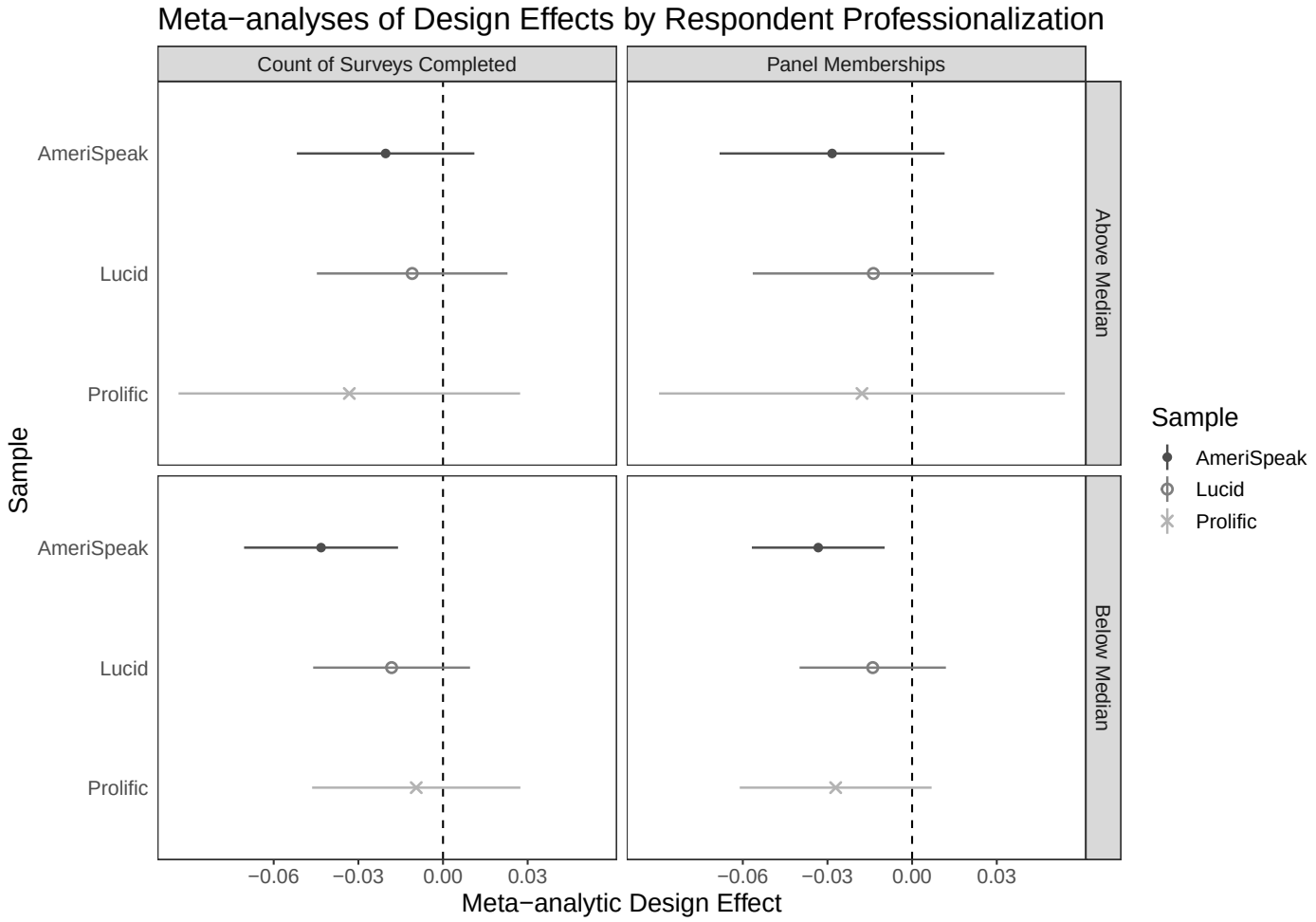

Figure A.2.1: Meta-Analytic design effect by respondent professionalization. The top (bottom) row of panels indicates the design effect among above (below) median respondents within each sample on each professionalization measure (columns).

A.3 Perceived Attitude Change

Table A.3.1 reports the results of each between-groups experiment, conditional on whether the respondent accurately or inaccurately self-reported (post-treatment) whether their attitude changed (and in which direction) since pre-treatment measurement. The final two columns report the respective conditional design effect relative to the post-only design. Among respondents who accurately self-report their level and direction of change (increase in support, decrease in support, or about the same), we observe an attenuation of the ATE relative to the post-only design in all nine between-groups studies, and this attenuation is significant in six. In contrast, among inaccurate respondents, we do not see this same pattern: the differences from the post-only design are smaller in magnitude in most studies, and the mean estimated design effect is quite close to zero.

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	Std. Err.	Est. ATE	Std. Err.	Accurate	Inaccurate
Foreign Aid	AmeriSpeak	0.036^{***}	0.005	0.121^{***}	0.014	-0.053^{***}	$0.032^{\dagger}$
Foreign Aid	Prolific	0.051^{***}	0.004	0.116^{***}	0.018	-0.060^{***}	0.005
Foreign Aid	Lucid	0.029^{***}	0.007	0.094^{***}	0.014	-0.034^{*}	0.031
Drug Imports	AmeriSpeak	0.029^{**}	0.009	0.086^{**}	0.030	-0.096^{***}	-0.039
Drug Imports	Prolific	0.020^{**}	0.007	0.189^{***}	0.035	-0.076^{*}	0.093^{*}
Drug Imports	Lucid	0.067^{***}	0.015	0.073^{**}	0.025	-0.043	-0.037
GMOs	AmeriSpeak	0.124^{***}	0.008	0.137^{***}	0.013	-0.038^{*}	-0.025
GMOs	Prolific	0.160^{***}	0.007	0.165^{***}	0.014	-0.019	-0.014
GMOs	Lucid	0.137^{***}	0.010	0.104^{***}	0.012	-0.007	-0.040^{*}

†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). The final two columns report the respective design effects (vs. post-only).

Figure A.3.1 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

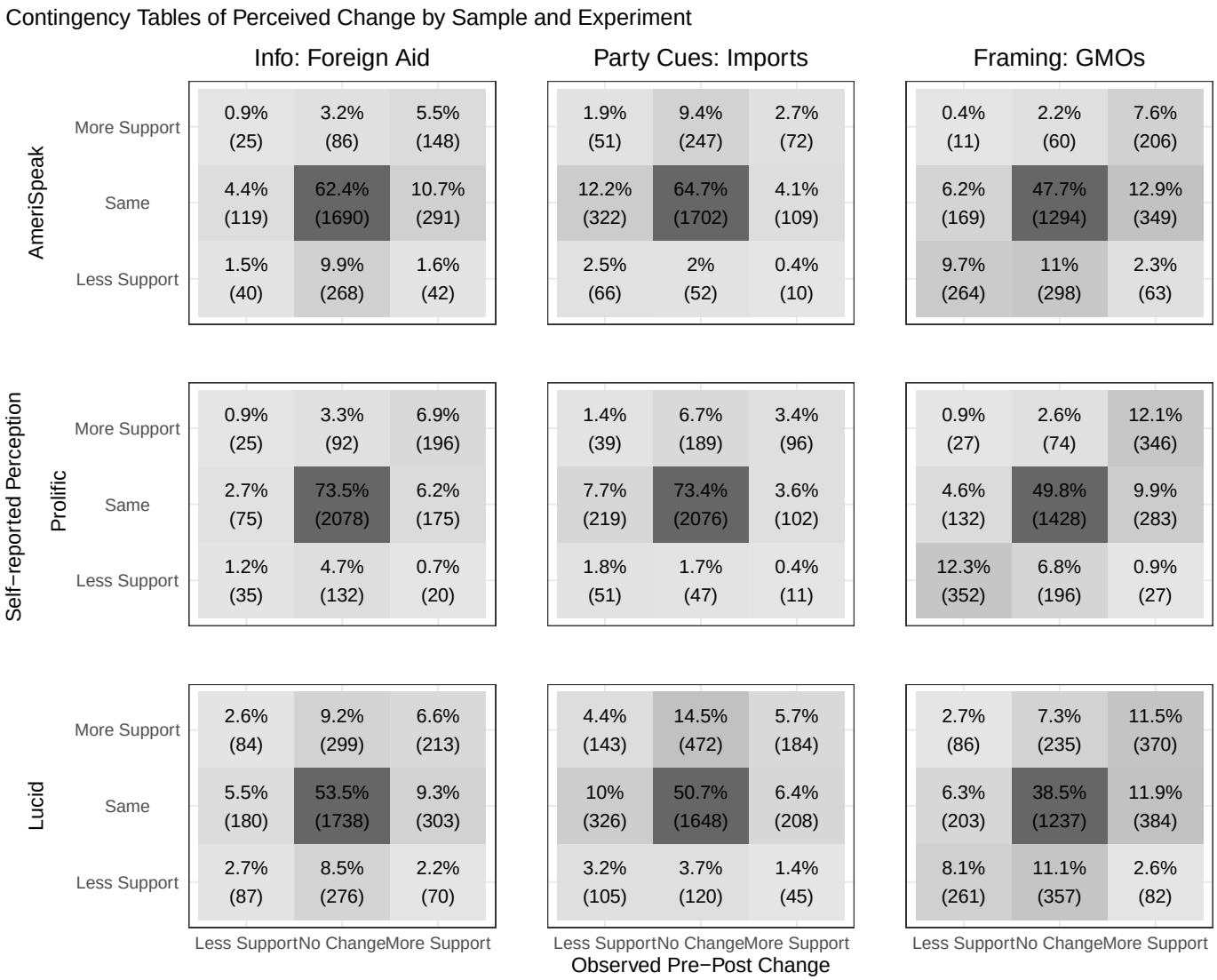

Figure A.3.1: Respondent perception of attitude change by sample and experiment. Figure shows contingency tables of respondents in each between-groups repeated measure experiment (panel columns) in each sample (panel rows) by actual observed pre-post change in responses (within-panel columns) and self-reported perceived pre-post change (within-panel rows). Frequency counts for each cell are shown in parentheses.

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.

A.4 Evidence of the Influence of Clifford, Sheagley, and Piston (2021)

As of January 2025, Clifford, Sheagley, and Piston (2021) have received 198 citations on Google Scholar. Of these, 88 are original studies referencing CSP to justify using repeated measure designs. In Table A.4.1 below, the pre-post designs have bolded titles and within-subject designs do not.

Table A.4.1: Citations Referencing CSP to Justify Repeated Measure Designs

No.	Title	Journal
1	**Depression and suicidality as evolved credible signals of need in social conflicts**	Evolution and Human Behavior
2	**Banklash: How Media Coverage of Bank Scandals Moves Mass Preferences on Financial Regulation**	American Journal of Political Science
3	**Latino-Targeted Misinformation and the Power of Factual Corrections**	Journal of Politics
4	**Testing Negative: The Non-Consequences of COVID-19 on Mass Ideology**	Unpublished
5	**Does Evidence Matter? The Impact of Evidence Regarding Aid Effectiveness on Attitudes Towards Aid**	The European Journal of Development Research
6	**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
7	**Does moral rhetoric fuel or reduce divides between parties and non-copartisan voters?**	Electoral Studies
8	**The Personal Vote in a Polarized Era**	American Journal of Political Science
9	**Descriptive, injunctive, or the synergy of both? Experimenting normative information on behavioral changes under the COVID-19 pandemic**	Frontiers in Psychology
10	**Media stereotypes, prejudice, and preference-based reinforcement: toward the dynamic of self-reinforcing effects by integrating audience selectivity**	Journal of Communication
11	**The Most Important Election of Our Lifetime: Focalism and Political Participation**	Political Psychology
12	**Career adaptability of interpreting students: A case study of its development and interactions with interpreter competences in three Chinese universities**	Frontiers in Psychology
13	**Paying for growth or goods: Tax morale among property owners in Lagos**	Journal of Experimental Political Science
14	**Imagined Otherness: Perceived Schematic Difference Can Fuel Dehumanization**	Unpublished
15	**Antiracism and its Discontents: The Prevalence and Political Influence of Opposition to Antiracism among White Americans**	Unpublished
16	**Making Issues Matter: Local Media and Policy-Based Evaluations of Politicians**	Unpublished
17	**Reliable Sources? Correcting Misinformation in Polarized Media Environments**	American Politics Research
18	**When Journalists Run for Office: The Effects of Journalist-Candidates on Citizens' Populist Attitudes and Voting Intentions**	International Journal of Communication
19	**Rules of Engagement: Elite Cues and Public Support for International Organizations**	Unpublished
20	**Confronting Core Issues: A Critical Test of Attitude Polarization**	Unpublished
21	**Beyond Changing Minds: Raising the Issue Importance of Expanding Legal Immigration**	Unpublished
22	**Can <3's Change Minds? Social Media Endorsements and Policy Preferences**	Social Media + Society
23	**Building intergroup trust through personal transfers: a field experiment in post-war Liberia**	Unpublished
24	**The Long Shadow of the Civil War: The Recurrent Historical Centrality of Anti-Black Political Threat in Eroding Public Support for American Democracy**	Unpublished
25	**Divestment as a Costly Signal: How Divestment Movements Affect Public Opinion**	Unpublished
26	**Winning Votes and Changing Minds: Do Populist Arguments Affect Candidate Evaluations and Issue Preferences?**	Unpublished
27	**Critical Race Theory and Asymmetric Mobilization**	Political Behavior
28	**The Holocaust, the Socialization of Victimhood and Outgroup Political Attitudes in Israel**	Comparative Political Studies

No.	Title	Journal
29	A randomized experiment evaluating survey mode effects for video interviewing	Political Science Research and Methods
30	Mass support for proposals to reshape policing depends on the implications for crime and safety	Criminology & Public Policy
31	Correcting the Misinformed: The Effectiveness of Fact-checking Messages in Changing False Beliefs	Political Communication
32	Politicized Battles: How Vacancies and Partisanship Influence Support for the Supreme Court	American Politics Research
33	Equality, Reciprocity, or Need? Bolstering Welfare Policy Support for Marginalized Groups with Distributive Fairness	American Political Science Review
34	Moral Rhetoric, Extreme Positions, and Perceptions of Candidate Sincerity	Political Behavior
35	Changes in Perceptions of Border Security Influence Desired Levels of Immigration	Journal of Conflict Resolution
36	Public support for phasing out carbon-intensive technologies: the end of the road for conventional cars in Germany?	Climate Policy
37	Partisan news versus party cues: The effect of cross-cutting party and partisan network cues on polarization and persuasion	Research & Politics
38	Women Experts and Gender Bias in Political Media	Public Opinion Quarterly
39	Active Student Responding and Student Perceptions: A Replication and Extension	Teaching of Psychology
40	From passerby to ally: Testing an intervention to challenge attributions for poverty and generate support for poverty-reducing policies and allyship	Analyses of Social Issues and Public Policy
41	Public Support for Professional Legislatures	State Politics & Policy Quarterly
42	Unilateral Inaction: Congressional Gridlock, Interbranch Conflict, and Public Evaluations of Executive Power	Legislative Studies Quarterly
43	Equating silence with violence: When White Americans feel threatened by anti-racist messages	Journal of Experimental Social Psychology
44	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
45	Greenwashing the Talents: attracting human capital through environmental pledges	Unpublished
46	Citizens as a Democratic Safeguard? The Sequence of Sanctioning Elite Attacks on Democracy	Unpublished
47	Can a constitutional monarch influence democratic preferences? Japanese emperor and the regulation of public expression	Social Science Quarterly
48	**Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments**	American Political Science Review
49	**Frontline employees' responses to citizens' communication of administrative burdens**	Public Administrative Review
50	**Correcting Myopia: Effect of Information Provision on Support for Preparedness Policy**	Political Research Quarterly
51	**Explaining the educational gradient in trust in politicians: a video-vignette survey experiment**	West European Politics
52	**No Evidence that Measuring Moderators Alters Treatment Effects**	American Journal of Political Science
53	**The Effect of International Actors on Public Support for Government Spending Decisions**	International Studies Quarterly
54	The policy acknowledgement gap: Explaining (mis)perceptions of government social program use	Policy Studies Journal
55	**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
56	**Varieties of Values: Moral Values Are Uniquely Divisive**	American Political Science Review
57	**Partisan Poll Watchers and Americans' Perceptions of Electoral Fairness**	Public Opinion Quarterly
58	**Going negative when spoiled for choice? Destabilizing and boomerang effects of negative political messaging in multiparty systems with multimember districts**	Political Research Exchange
59	**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
60	**Scientific supremacy: How do genetic narratives relate to racism?**	Politics and the Life Sciences
61	**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
62	**Anger expressions and coercive credibility in international crises**	American Journal of Political Science

No.	Title	Journal
63	**When partisanship and technocratic credibility collide: mass attitudes and central bank endorsements of fiscal policy in Canada and the USA**	Socio-Economic Review
64	The persuasive impact of athlete racial advocacy on individuals' cognitive responses: evidence from survey experiments in Japan	European Sport Management Quarterly
65	**Tweet no harm: Offer solutions when alerting the public to voter suppression efforts**	Communication and the Public
66	**Does the military lose public confidence without compliance with civilian control? Experimental evidence from Japan**	Journal of Peace Research
67	**Filling the EU information deficit mitigates negative EU attitudes among the least knowledgeable. Evidence from a population-based survey experiment**	Journal of European Integration
68	**Is Support for Authoritarian Rule Contagious? Evidence from Field and Survey Experiments**	Unpublished
69	Bureaucracy and Cyber Coercion	International Studies Quarterly
70	Supplemental online resources improve data literacy education: Evidence from a social science methods course	PLOS One
71	On motives and means: how approach and justification for court-curbing impact public trust	Democratization
72	**Imagined otherness fuels blatant dehumanization of outgroups**	Communications Psychology
73	**The policy basis of group sentiments**	Political Science Research Methods
74	**Beyond partisan filters: Can underreported news reduce issue polarization?**	PLOS One
75	**Public Reactions to Communication of Uncertainty: How Long-Term Benefits Can Outweigh Short-Term Costs**	Public Opinion Quarterly
76	Active Student Responding and Student Perceptions: A Replication and Extension	Teaching of Psychology
77	**The power of empirical evidence: assessing changes in public opinion on constitutional emergency provisions**	Public Choice
78	**(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
79	**Does the prospect of further sovereignty loss fuel Euroscepticism? A population-based survey experiment**	European Societies
80	**Anti-Black Political Violence and the Historical Legacy of the Great Replacement Conspiracy**	Perspectives on Politics
81	Moral Rhetoric, Extreme Positions, and Perceptions of Candidate Sincerity	Political Behavior
82	**Economic Inequality and Willingness to Pay for Collective Goods: Theory and Experimental Evidence**	Unpublished
83	**Race- and Class-Based Messaging and Anti-Carceral Policy Support**	Unpublished
84	**What Money Can't (or Can) Buy: Inward Foreign Direct Investment and Backlash against Globalization in the United States**	Unpublished
85	**Identity and the Social Construction of Reputation in World Politics**	Unpublished
86	**To Mitigate or Adapt? The Role of Climate Vulnerability on Policy Preferences**	Unpublished
87	**Digital Cloning of the Dead: Exploring the Optimal Default Rule**	Asian Journal of Law and Economics
88	Beyond Meating Climate Goals: The Unpopularity of Masculine-Threatening Climate Policies	Unpublished

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, funded and operated by NORC at the University of Chicago, 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 27$^{\text{th}}$ to July 15$^{\text{th}}$, 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 9$^{\text{th}}$. 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 Survey Questionnaire

B.3.1 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.3.2 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. Note that the order of items was randomized as discussed in the main text.

Foreign Aid (Study 1)

Foreign Aid Pretreatment/Control (1 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 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 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 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 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 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 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 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 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 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 unit): "Do you generally favor or oppose affirmative action programs for women?"

- Favor
- Oppose
- No opinion

Affirmative Action Race (1 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 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 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 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 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 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.3.3 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]*