
# Repeated Measure Designs in Survey Experiments: A Replication Study

| | |
|---|---|
| Journal: | *Time-Sharing Experiments for the Social Sciences* |
| Manuscript ID | TESS-1727 |
| Manuscript Type: | Original Article |
| Specialty Area: | Political Science, Methods/Measurement |
| | |

# Repeated Measure Designs in Survey Experiments: A Replication Study

*Submitted as part of the TESS Special Competition on Replications*

**Abstract:** A recent 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 increases precision when estimating treatment effect without introducing bias. In light of these promising findings, many researchers have since adopted repeated measure designs. We propose replicating and extending Clifford et al.'s work with a large, probability-based sample by having respondents complete six experiments with random assignment to one of three design conditions for each experiment: a post-only design, a repeated measure design with well-separated pre-post measures, and a repeated measure design with immediately sequential pre-post measures. In addition to providing further empirical evidence about the bias introduced and precision gained with repeated measure designs, our study will extend Clifford et al. (2021) by identifying how design considerations such as the distance between repeated measures and the use of within-subject designs affect the precision gained and bias introduced from repeated measures. Our study will offer critical insights into the promise and potential pitfalls of repeated measure designs amid heightened scholarly interest in improving survey experimental practices.

**Summary of Clifford et al. (2021):** Social scientists have long relied on the between-group experimental design with outcomes measured only post-treatment. In its simplest form, the post-only design randomly assigns respondents to either a control/placebo or treatment condition. The outcome of interest is only measured post-treatment, and differences in the groups' outcomes are interpreted as treatment effects. Many researchers have preferred post-only designs because the common wisdom holds that repeated measure designs—i.e., those measuring outcomes pre- *and* post-treatment, whether treating all respondents (within-subject) or some respondents (pre-post, between-subject)—risks introducing bias from demand effects, consistency effects, and priming.

However, post-only experiments tend to suffer low levels of precision. Low precision (or high uncertainty) becomes an issue when trying to identify small effects or effect heterogeneity, especially in demographically-diverse samples (Mutz 2011). Underpowered studies are also at higher risk of missing treatment effects or mis-estimating their sizes (Gelman and Carlin 2014; Loken and Gelman 2017). Post-only designs thus often need large samples to achieve adequate power—a potentially expensive requirement given the decreasing returns to precision for each one-unit increase in sample size. Repeated measure designs, by contrast, increase precision by accounting for respondents' baseline pre-treatment outcome values.

Conventional wisdom thus proposes a bias-precision tradeoff between post-only and repeated measure designs. Clifford et al. (2021) test this conventional wisdom by comparing bias and precision for repeated measure designs against otherwise identical post-only designs. Meta-analyzing six experiments (outlined in Table 1), Clifford et al. find no significant differences in the treatment effects from repeated measure and post-only designs, but find repeated measures substantially bolster precision. For example, to achieve 80 percent power to detect a 0.20 standardized effect requires a 1,000-respondent post-only study, but a pre-post study requires only about 200 to 600 respondents (depending on the level of correlation between the pre- and post-treatment measures). Clifford et al. recommend that "researchers use pre-post and within-subjects designs whenever possible" (p. 1062) because the increase in precision does not appear to be offset by an increase in bias—i.e., that there is no precision-bias tradeoff with repeated measure experiments.

In the two years since its publication, Clifford et al. (2021) has already received 112 Google Scholar citations. Of these, at least 47 were original studies referencing Clifford et al. to justify using repeated measure designs (see Appendix C for the full list of studies). And while

most citations to Clifford et al. (2021) are from political scientists, scholars in communications, criminology, economics, education, and environmental policy have also referenced Clifford et al. to justify using repeated measure designs. This single paper's influence on experimental practice is already quite clear, and is likely to grow in coming years as awareness increases for how low statistical power in experiments has contributed to the ongoing replication crisis (Arel-Bundock et al. 2022; Ioannidis, Stanley, and Doucouliagos 2017; Open Science Collaboration 2015).

**Reason for Replication Study:** While Clifford et al. (2021) provides a critical test of the bias-precision tradeoff, the empirical literature on this question remains scant, and even this advance leaves several important questions unanswered—which we propose testing. First, Clifford et al. fielded their experiments on two student samples and four online non-probability samples. Their findings are certainly important considering how much experimental research relies on these samples (Krupnikov and Levine 2014), but it is unclear whether the lack of bias (due to demand, consistency, and/or priming) they find also holds among probability samples, which recruit more representative, infrequent, and attentive respondents than professionalized opt-in online panels (Kennedy et al. 2016). In other words, we do not know whether design effects are heterogeneous across degrees of respondent professionalization. To give one example of why this might be an issue, in one pre-post experiment Clifford et al. find most respondents whose opinions changed post-treatment did not *think* that their opinions had changed; however, it is unclear whether this finding should be attributed to the unobtrusiveness of repeated measures, inattentiveness among non-probability and student samples, or the large number of survey questions these respondents tend to encounter. More attentive or less professionalized respondents could be more cognizant when they encounter the same question twice, raising the risk of consistency and demand effects.

Second, Clifford et al. do not test whether proximity between repeated measures matters.

An intuitive theory that is partly reflected in Clifford et al.'s design decisions posits that repeated measures with greater distance (i.e., survey content) between them are less likely to introduce bias. Clifford et al.'s decision to have respondents to their welfare question-wording experiment complete the first question "early in the survey" and the second "near the end of the survey" is surely intentional—and may increase the probability that (for example) inattentive respondents "forgot" their earlier responses and felt no pressure to answer consistently. It remains unclear if repeated measures are appropriate for short studies that do not include much content to separate pre- and post-treatment measures. Our study will randomize the distance between pre- and post-treatment measures to offer insight into whether bias increases with closer measure proximity.

Third, as shown in Appendix C, 40 percent of studies citing Clifford et al. used within-subject designs, not between-group pre-post designs. Clifford et al. analyze six experiments, of which only one (n = 900) uses a within-subject design: the canonical welfare question-wording experiment (Smith 1987). While Clifford et al. combine all six studies in their internal meta-analysis, they offer little evidence that within-subject designs, specifically, do not introduce bias. Given the apparent interest in within-subject designs, but limited ability of Clifford et al. to test whether these designs introduce bias, our study will expand the number of within-subject experiments from one to three, with a total stacked sample 13 times larger than the sole welfare question-wording experiment in their original analysis. Usefully, our inclusion of two additional question-wording experiments, drawn from Wilson et al. (2008) and De Benedictis-Kessner and Hankinson (2019), provides an opportunity to replicate two additional published studies.

**Replication Study Design:** Our proposed replication study follows and extends Clifford et al.'s (2021) research design towards identifying the degrees of bias introduced and precision gained with repeated measure designs vis-à-vis traditional post-only designs. We select six experiments

to replicate, four of which are drawn from Clifford et al.: Study 1, a canonical question wording experiment on welfare; Study 2, an information experiment on foreign aid; Study 5, a party cues experiment on prescription drug importation; and Study 6, a framing experiment on genetically modified organisms (GMOs). To these we add a question-wording experiment on affirmative action from Wilson et al. (2008) and a question-wording experiment on opioid policy from De Benedictis-Kessner and Hankinson (2019, Study 2).[1] Appendix A outlines each study in detail.

In our study, each respondent will participate in all six experiments in a random order, but with a randomly assigned design for each individual experiment. Specifically, each respondent will complete two experiments in the post-only condition, two experiments in the "proximate" repeated measure condition in which the first and second measurements immediately follow one another (before and after a treatment), and two experiments in the "distant" repeated measure condition in which the first and second measurements are separated by unrelated survey content. This structured randomization ensures respondents complete surveys of similar lengths; for our six proposed experiments, this is 13 or 14 TESS units per respondent. Again following Clifford et al., respondents will also be asked, for 2-3 repeated measure experiments, whether they thought that their attitudes changed since first measurement, bringing all respondents to 16 TESS units. To illustrate this structured randomization, two possible assignments are shown in Table 2.

Following Clifford et al. (2021), our analysis will first estimate treatment effects for all six experiments separately by the three design conditions. Then, we will compare the treatment effects across designs to estimate design effects on treatment effect size. The comparison of the post-only and "distant" repeated measure conditions directly follows from Clifford et al.;

[1] Notably, Wilson et al. (2008) finds a consistency effect on support for affirmative action towards women vs. racial minorities using a within-subject design. Our replication will test whether increasing the distance between measures attenuates the consistency effect, or whether this study needs a post-only design to eliminate consistency pressures.

however, we will also extend their work by comparing design effects between the post-only and "proximate" repeated measure conditions to test whether repeated measures immediately before and after treatment introduce bias. Again following Clifford et al., we will conduct internal meta-analyses aggregating the design effects for all six experiments, comparing the post-only, distant repeated, and proximate repeated designs. We will then separately analyze the three pre-post and three within-subject designs to determine if bias differs across these possible versions of repeated measure designs. Our analyses will directly reexamine Clifford et al.'s main findings and include extensions to identify possible biases arising in short repeat measure studies and within-subject designs, specifically. With a sample of 3,900, we are well-powered to conduct these tests, with power approaching 1 for identifying each experiment's main treatment effects and meta-analytic power above 0.9 to identify design effects as small as 0.06 standard deviations (Appendix B).

**Conclusion:** Our study will offer a rigorous replication of an influential paper that is already changing social scientific survey research. Our replication effort will retest Clifford et al.'s original claims, assessing their generalizability to probability-based samples, clarifying whether repeated measures in shorter surveys where large separations between measurements are not possible introduce bias, and providing a stronger test of design effects for within-subject experiments. Regardless of the outcome, our study will have important implications for survey experimental design. Should we confirm Clifford et al.'s claims, we will show that repeated measure designs are suitable for probability-based samples, where increasing statistical power is especially valuable as a cost-saving device, and that repeated measures can be employed even in short surveys. Should our results instead uncover bias in repeated measure designs, our study will offer a timely notice for caution when considering the original authors' recommendation that researchers use repeated measure experimental designs "whenever possible."

***Table 1.*** Overview of Experimental Methods from Clifford et al. (2021, Table 3)

| Study | Topic | Manipulation | Sample Source | Dates | Sample size | Post-only | Within-subject | Pre-post | Quasi |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Welfare | Question Wording | Student | Spring 2018 | 900 | X | X | | |
| 2 | Foreign Aid | Information | MTurk | March 2018 | 1209 | X | | X | |
| 3 | Education | Information | MTurk | May 2018 | 1206 | X | | X | X |
| 4 | Estate Tax | Information | Lucid | Spring 2018 | 2462 | X | | X | X |
| 5 | Prescription Drugs | Party Cues | Forthright | July 2019 | 1531 | X | | X | X |
| 6 | GMOs | Framing | Student | Spring 2020 | 965 | X | | X | X |

***Table 2.*** Example of proposed randomization for two hypothetical respondents.

| Respondent | Experiment | Manipulation | Condition | Measure | Units |
|---|---|---|---|---|---|
| Alice | Affirmative Action | Question Wording | Distant | Wording 1 | 1 |
| Alice | Opioid Policy | Question Wording | Distant | Wording 2 | 2 |
| Alice | Prescription Drugs | Party Cues | Proximate | Pre-Treatment | 1 |
| Alice | Prescription Drugs | Party Cues | Proximate | Post-Treatment | 1 |
| Alice | Prescription Drugs | Party Cues | Proximate | Attitude Change? | 1 |
| Alice | GMOs | Framing | Proximate | Pre-Treatment | 1 |
| Alice | GMOs | Framing | Proximate | Post-Treatment | 2 |
| Alice | Welfare | Question Wording | Post-Only | Wording 1 or 2 | 1 |
| Alice | Foreign Aid | Information | Post-Only | Post-Treatment | 1 |
| Alice | Affirmative Action | Question Wording | Distant | Wording 2 | 1 |
| Alice | Opioid Policy | Question Wording | Distant | Wording 1 | 2 |
| Alice | Opioid Policy | Question Wording | Distant | Attitude Change? | 1 |
| Alice | Opioid Policy | Question Wording | Distant | Covariate | 1 |
| | | | | | Total: 16 |
| Bob | Foreign Aid | Information | Distant | Pre-Treatment | 1 |
| Bob | Prescription Drugs | Party Cues | Distant | Pre-Treatment | 1 |
| Bob | Welfare | Question Wording | Proximate | Wording 1 | 1 |
| Bob | Welfare | Question Wording | Proximate | Wording 2 | 1 |
| Bob | Affirmative Action | Question Wording | Proximate | Wording 2 | 1 |
| Bob | Affirmative Action | Question Wording | Proximate | Wording 1 | 1 |
| Bob | Affirmative Action | Question Wording | Proximate | Attitude Change? | 1 |
| Bob | GMOs | Framing | Post-Only | Post-Treatment | 2 |
| Bob | Opioid Policy | Question Wording | Post-Only | Wording 1 or 2 | 2 |
| Bob | Opioid Policy | Question Wording | Post-Only | Covariate | 1 |
| Bob | Foreign Aid | Information | Distant | Post-Treatment | 1 |
| Bob | Foreign Aid | Information | Distant | Attitude Change? | 1 |
| Bob | Prescription Drugs | Party Cues | Distant | Post-Treatment | 1 |
| Bob | Prescription Drugs | Party Cues | Distant | Attitude Change? | 1 |
| | | | | | Total: 16 |

## References

Arel-Bundock, Vincent, Ryan Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and T. D. Stanley. 2022. "Quantitative Political Science Research Is Greatly Underpowered." OSF Preprints. https://doi.org/10.31219/osf.io/7vy2f.

Clifford, Scott, Geoffrey Sheagley, and Spencer Piston. 2021. "Increasing Precision without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments." *American Political Science Review* 115 (3): 1048–65.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6): 641–51.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127 (605): F236–65.

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

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

Loken, Eric, and Andrew Gelman. 2017. "Measurement Error and the Replication Crisis." *Science* 355 (6325): 584–85.

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

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

*Appendix to*
"Repeated Measure Designs in Survey Experiments:
A Replication Study"

## A. Study Descriptions

We propose replicating six studies. Four studies are drawn from Clifford et al. (2021); we add one question-wording experiment from each of de Benedictis-Kessner and Hankinson (2019) and Wilson et al. (2008). Below are short descriptions of each study that we will replicate.

### A.1 Clifford et al. (2021) Replications

In Clifford et al. (2021), the authors conducted six studies, of which we will replicate four. Each of these four studies are described in detail below.

**Study 1—Welfare Question-Wording Experiment.** Clifford et al. (2021, Study 1) replicated a welfare question-wording experiment originally conducted by Smith (1987). Using a student sample (*N=900*), Clifford et al. (2021) randomized respondents into post-only or within-subject designs. Respondents are randomly assigned to a question about spending levels on "welfare" or "assistance to the poor" on a three-point scale. In the post-only design, respondents at the end of the survey were randomly assigned one of the two questions. In the within-subjects design, respondents were randomly assigned one question early in the survey, and then they completed the other question toward the end of the survey. The two question wordings were:

1. [Welfare (1 unit)] "Generally speaking, do you think we're spending too much, too little or about the right amount on welfare?" *((1) Too much, (2) About the right amount, (3) Too little)*
2. [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?" *(1) Too much, (2) About the right amount, (3) Too little)*

**Study 2—Foreign Aid Information Experiment.** Clifford et al. (2021, Study 2) replicated an information experiment originally conducted by Gilens (2001). Using a convenience sample from Amazon's Mechanical Turk (*N=1,209*), Clifford et al. randomized respondents into post-only or pre-post designs. Respondents are randomly assigned to either receive or not receive information that spending on foreign aid makes up about 1% of the federal budget. Respondents are then asked whether foreign aid spending should increase, remain the same, or decrease. In the pre-post design, respondents answer the control question and then later answer either the control or treatment question version. Question wording was as follows:

1. [Pretreatment/Control (1 unit)] "Do you think spending on foreign aid should be increased or decreased?" *((1) Greatly increased, (2) Slightly increased, (3) Kept about the same, (4) Slightly decreased, (5) Greatly decreased)*
2. [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?" *((1) Greatly increased, (2) Slightly increased, (3) Kept about the same, (4) Slightly decreased, (5) Greatly decreased)*

**Study 3—Prescription Drugs Party Cue Experiment.** This Clifford et al. (2021, Study 5) experiment focused on support for allowing the importation of prescription drugs from Canada, a replication of Clifford et al. (2019). Using a Forthright panel (*N=1,531*), Clifford et al. compare post-only, pre-post, and quasi pre-post designs. All respondents complete the control question in wave 1 of the panel, then respondents are randomly assigned to design conditions in wave 2. The relevant question wordings for the post-only and pre-post designs are:

1. [Pretreatment/Control (1 unit)] "Do you support or oppose allowing individuals to import prescription drugs from Canada?" *((1) Strong support, (2) Somewhat support, (3) Slightly support, (4) Neither support nor oppose, (5) Slightly oppose, (6) Somewhat oppose, (7) Strongly oppose)*
2. [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?" *((1) Strong support, (2) Somewhat support, (3) Slightly support, (4) Neither support nor oppose, (5) Slightly oppose, (6) Somewhat oppose, (7) Strongly oppose)*

**Study 4—GMOs Framing Experiment.** Using a student sample (*N=965*), Clifford et al. (2021, Study 6) compare post-only, pre-post, and quasi pre-post designs. All respondents are assigned to receive either a pro-GMO frame or an anti-GMO frame (no control). The dependent variable is a seven-point measure of support for GMOs. Respondents in the pre-post condition also complete a pre-treatment version of the dependent variable without any pro/con framing. The relevant question wordings for the post-only and pre-post designs are:

1. [Pretreatment (1 unit)] "How strongly do you favor or oppose the production and consumption of genetically modified foods?" *((1) Strongly favor, (2) Favor, (3) Slightly favor, (4) Neither favor nor oppose, (5) Slightly oppose, (6) Oppose, (7) Strongly oppose)*
2. [Anti-GMO Frame (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 about you? How strongly do you favor or oppose the production and consumption of genetically modified foods?" *((1) Strongly favor, (2) Favor, (3) Slightly favor, (4) Neither favor nor oppose, (5) Slightly oppose, (6) Oppose, (7) Strongly oppose)*
3. [Pro-GMO Frame (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 about you? How strongly do you favor or oppose the production and consumption of genetically modified foods?" *((1) Strongly favor, (2) Favor, (3) Slightly favor, (4) Neither favor nor oppose, (5) Slightly oppose, (6) Oppose, (7) Strongly oppose)*

## A.2 Additional Replication Studies

In addition to the four replications from Clifford et al. (2021), we will replicate two additional question-wording experiments suitable for comparing post-only and within-subject designs.

**Study 5—Opioid Question-Wording Experiment.** In "Concentrated burdens: How self-interest and partisanship shape opinion on opioid treatment policy" (de Benedictis-Kessner and Hankinson 2019), the authors used the AmeriSpeak panel (*N=2,000*) and had respondents read

two policy proposals and answer a question about each. Here, we propose replicating their study on the second policy proposal about a medication-assisted treatment clinic. In the post-only condition, respondents will complete the near or far treatment clinic condition. In the within-subject condition, respondents will complete both the near and far treatment clinic conditions in a random order. All respondents will complete a question about their personal exposure to opioid addiction. The relevant question wordings are:

1. [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?" *((1) Strongly support, (2) Somewhat support, (3) Neither support nor oppose, (4) Somewhat oppose, (5) Strongly oppose)*
2. [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?" *((1) Strongly support, (2) Somewhat support, (3) Neither support nor oppose, (4) Somewhat oppose, (5) Strongly oppose)*
3. [Personal Exposure (1 unit)] "Do you personally know anyone who has ever been addicted to opioids, including prescription painkillers or heroin?" *((1) Yes, me (2) Yes, a family member (3) Yes, a close friend (4) Yes, an acquaintance (5) No, I do not know anyone who has ever been addicted to opioids)*

**Study 6—Affirmative Action and Gender Question Order Experiment.** In "Affirmative Action Programs for Women and Minorities: Expressed Support Affected by Question Order," Wilson et al. (2008) collected data from the 2003 Gallup survey which used a split-ballot experimental design (*N=1,385*) to study how asking respondents about support for two policies, gender-based and race-based, is impacted by question order (Moore, 2003). Respondents were randomly assigned to one of two conditions, either seeing a gender-based policy question then a race-based policy question or the reverse. We will randomize respondents to post-only design (in which a respondent only answers 1 question) or within-subject designs. The relevant question wordings are:

1. [Gender (1 unit)] "Do you generally favor or oppose affirmative action programs for women?" *((1) Favor, (2) Oppose, (3) No opinion)*
2. [Race (1 unit)] "Do you generally favor or oppose affirmative action programs for racial minorities" *((1) Favor, (2) Oppose, (3) No opinion)*

## A.3 Perceived Attitude Change

Clifford et al. (2021) asked a recall question in the GMO framing experiment (Study 6) that measured whether participants thought their attitude had changed. Depending on which designs each respondent sees for our six experiments (and thus completing either 13 or 14 TESS units of length), we will ask a similar question two or three times for each respondent, bringing all

respondents to 16 units of length. We will vary this question such that participants are asked to recall their attitude for two or three of the four studies in which they provided two responses (pre- and post-treatment). The recall question wording is as follows:

1. [Recall Previous Attitude (1 unit)] As you may remember, we also asked you about your support or opposition to [welfare / assistance to the poor / foreign aid / importing subscription drugs from Canada / genetically modified foods (GMOs) / medication-assisted treatment clinics / affirmative action][1] at the beginning of the survey. To the best of your memory, how has your support for [welfare / assistance to the poor / foreign aid / importing subscription drugs from Canada / genetically modified foods (GMOs) / medication-assisted treatment clinics / affirmative action] changed since the beginning of the survey? *((1) Support <u>increased</u> since the beginning of the survey, (2) Support <u>decreased</u> since the beginning of the survey, (3) Support stayed the same)*



[1] For Study 1 only, we will pipe in the correct wording depending on the wording that the participant was shown first.

## B. Power Analyses

Our proposed replication involves several components. First, we will attempt to directly replicate six experiments; our power to replicate each of these experiments approaches 1. Second, we will attempt to replicate Clifford et al.'s (2021) finding of a null design effect (pre-post or within-subject versus post-only) for four experiments, as well as a partial replication of their internal meta-analysis confirming the null design effect. Given our intent is to replicate a null finding, it is impractical to discuss our power to identify a true effect equivalent to the effect size they find. Instead, we focus on our power to identify a minimum detectable effect. We have power of 0.90 to uncover design effects as small as 0.13 (Cohen's *d*) for each of these experiments, or about two-thirds of a conventionally "small" effect. In the meta-analysis, we have power above 0.90 to uncover design effects as small as 0.06, or less than one-third of a conventionally "small" effect. This represents a substantial increase in power relative to that of the original meta-analysis to uncover a similarly small effect (approximately 0.70). Third, we offer an extension by examining design effects among question wording experiments (within-subject versus post-only) with a second internal meta-analysis. For this extension, we have power above 0.95 to uncover an effect as small as 0.10. Finally, we offer an extension that examines design effect from pre-post instruments that do not include substantial separation between the repeated measurements (the "proximate" condition). For this extension, we have power of 0.70 to uncover design effects in each individual experiment as small as 0.10, and in a meta-analysis we have power above 0.90 to uncover design effects as small as 0.06.

We discuss each of the power calculations for these components in detail below. In all calculations, we assume a two-tailed test of difference in means and use the standard significance threshold of $\alpha = 0.05$. As per our design, we also assume equal assignment to treatment and control conditions for the main effect replications, and equal assignment between the post-only, proximate pre-post, and distant pre-post conditions for the analysis of design effects.

### B.1 Main Effect Replications

**Study 1 - Welfare Question-Wording Experiment.** Comparing the welfare wording versus assistance for the poor wording in their within-subject design, Clifford et al. (2021, Study 1) find an effect size of 0.41 standard deviations (Cohen's *d*). We are extremely well powered to replicate this finding. With a sample size of 3,900 respondents—or even just the ~1,300 respondents assigned to the "distant" pre-post condition that most closely mimics their design—we have power approaching 1 to detect an effect of the same size. To detect an effect of half the size found in Clifford et al., we are powered at 0.96 with just the "distant" condition respondents, and our power approaches 1 with the full sample.

**Study 2 - Foreign Aid Information Experiment.** In their pre-post design condition, Clifford et al. (2021, Study 2) find an effect size of 0.11 standard deviations (Cohen's *d*) from the information treatment, and find larger effect sizes in their post-only condition (0.30 standard deviations) and full sample (0.24 standard deviations). We are well powered to replicate this finding. Our power approaches 1 to identify effect sizes equivalent to Clifford et al.'s post-only condition and full sample, and we have power of 0.93 to detect an effect of the size found in their pre-post condition (the smallest).

**Study 3 - Prescription Drugs Party Cue Experiment.** Clifford et al. (2021, Study 5) find an effect size of 0.45 standard deviations (Cohen's *d*) from the party cues treatment in their pre-post condition. We are extremely well powered to replicate this finding. With a sample size of 3,900 respondents—or even just the ~1,300 respondents assigned to the pre-post condition—we have power approaching 1 to detect an effect of the same size. To detect an effect of half the size found in Clifford et al., we are powered at 0.98 with just the "distant" condition respondents, and our power approaches 1 with the full sample.

**Study 4 - GMOs Framing Experiment.** Clifford et al. (2021, Study 6) find an effect size of 0.40 standard deviations (Cohen's *d*) from the framing treatment in their pre-post condition. We are extremely well powered to replicate this finding. With a sample size of 3,900 respondents—or even just the ~1,300 respondents assigned to the "distant" condition—we have power approaching 1 to detect an effect of the same size. To detect an effect of half the size found in Clifford et al., we are powered at 0.95 with just the "distant" condition respondents, and our power approaches 1 with the full sample.

**Study 5 - Opioid Question-Wording Experiment.** In their analysis of the distance treatment for (binarized) support or opposition to opening a new clinic, de Benedictis-Kessner & Hankinson (2019) find an effect size of 0.29 standard deviations (Cohen's *d*) in a post-only design. We are extremely well powered to replicate this finding. With a sample size of 3,900 respondents—or even just the ~1,300 respondents assigned to the post-only condition—we have power approaching 1 to detect an effect of the same size. To detect an effect of half the size found in the original study, we have power of 0.74 with just the post-only respondents, and power of 0.99 with the full sample. Further, our proposed design has power to detect design effects as small as 0.13 with power above 0.90, and power of 0.72 to uncover design effects as small as 0.10.

**Study 6 – Affirmative Action Question-Wording Experiment.** In within-subject design design, Wilson et al. (2008) find an (approximate[2]) effect size of 0.15 standard deviations (Cohen's *d*) for a question wording experiment on affirmative action. With a sample size of 3,900 respondents, we have power approaching 1 to detect an effect of the same size; with just the ~1300 respondents assigned to the "proximate" pre-post condition that most closely mimics the design of Wilson et al., we have power of 0.77 to detect an effect of a similar size.

## B.1 Design Effect Replications

**Study 1 - Welfare Question-Wording Experiment.** Clifford et al. (2021, Study 1) find a null design effect of 0.04 standard deviations (Cohen's *d*). We have power of 0.90 to uncover a true design effect as small as 0.13, and power of 0.72 to uncover a design effect of 0.10. This represents a significant increase in power over Clifford et al., who had power of 0.50 or 0.32 (respectively) to identify true effects of the same small size with a sample of 900 students.

**Study 2 - Foreign Aid Information Experiment.** Clifford et al. (2021, Study 2) find statistically significant design effect of 0.17 standard deviations (Cohen's *d*). We have power of 0.99 to identify a design effect of the same size, and power 0.72 to uncover a design effect as small as 0.10.

[2] Wilson et al. do not provide replication data. Using the percentages in Table 1 and Table 2 of their manuscript, we assembled a facsimile of their dataset to generate this effect size estimate and conduct power analyses.

**Study 3 - Prescription Drugs Party Cue Experiment.** Clifford et al. (2021, Study 5) find a null design effect of 0.11 standard deviations (Cohen's *d*). We have power of 0.90 to uncover a true design effect as small as 0.13, and power of 0.72 to uncover a design effect of 0.10. This represents a significant increase in power over Clifford et al., who had power of 0.62 or 0.41 (respectively) to identify true effects of the same small size with a sample of 1,206 respondents.

**Study 4 - GMOs Framing Experiment.** Clifford et al. (2021, Study 6) find a null design effect of 0.04 standard deviations (Cohen's *d*). We have power of 0.90 to uncover a true design effect as small as 0.13, and power of 0.72 to uncover a design effect of 0.10. This represents a significant increase in power over Clifford et al., who had power of 0.52 or 0.34 (respectively) to identify true effects of the same small size with a sample of 965 students.

**Internal Meta-analysis.** In an internal random-effects meta-analysis of six studies, Clifford et al. (2021) find a null design effect -0.014 (outcomes rescaled to vary between 0 and 1), with between-study variance ($\tau^2$) estimated to be 0.00026 and residual variance from heterogeneity ($I^2$) to be 0.026. Following Griffon (2021), we use the $I^2$ value of Clifford et al. to inform our power calculation, and assume an average of ~2,600 respondents for each of six studies to directly compare two design conditions (that is, for each experiment, one third of our sample will have been assigned to the third design condition, leaving two thirds of the sample in the direct test of two design conditions). In this internal meta-analysis, we have power of 0.93 to identify a true design effect as small as 0.06 (Cohen's *d*), allowing for heterogeneity (random effects), and power of 0.62 to identify a true design effect as small as 0.04 (Cohen's *d*), or one-fifth of a conventionally "small" effect. In comparison, Clifford et al. had power of only 0.70 or 0.37 (respectively) to identify true design effects of the same size. Even under much more conservative assumptions (increasing the $I^2$ value by an order of magnitude to 0.25), our meta-analysis retains power of 0.87 to detect a design effect as small as 0.06, and power of 0.56 to detect an effect as small as 0.04.

## B.3 Design Effects for Question Wording Experiments

A further strength of our proposal is our ability to focus on the design effects of within-subject versus post-only designs for question wording experiments. An internal meta-analysis of just the three studies we propose has power of 0.96 to identify a true design effect of 0.10 and power of 0.78 to identify a true design effect as small as 0.07. With more conservative assumptions about residual heterogeneity ($I^2 = 0.25$), our power to identify a true design effect of 0.10 only falls to 0.93, and to identify an effect as small as 0.07 our power falls only 0.71.

## B.4 Design Effects within Pre-Post Designs

Finally, we propose to estimate the design effects of pre-post designs that separate their repeated measures by other survey content versus those that do not. As with the replication of Clifford et al.'s analyses of design effects between pre-post and post-only designs, we have power of 0.72 to identify true design effects as small as 0.10 for any individual study (again assuming ~2,600 respondents for each experiment assigned equally across design conditions). Further, an internal meta-analysis of the six studies has power of 0.93 to identify a true design effect as small as 0.06 (Cohen's *d*), allowing for heterogeneity (random effects), and power of 0.62 to identify a true design effect as small as 0.04.

## C. Citation List

As of September 5, 2023, Clifford et al. (2021) had received 94 citations (Google Scholar). Of these, 47 were original studies referencing Clifford et al. (2021) to justify using repeated measure designs. The pre-post designs have bolded titles and within-subject designs do not.

| | **Title (Bolded=Pre-Post,** Unbolded=Within-Subject**)** | **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 ♥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 |
| 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 |

# References

Clifford, S., Leeper, T.J., & Rainey, C. 2019. "Increasing the Generalizability of Survey Experiments Using Randomized Topics: An Application to Party Cues." Paper presented at the Annual Meeting of the American Political Science Association, Washington, DC.

Clifford, S., Sheagley, G., & Piston, S. 2021. "Increasing precision without altering treatment effects: Repeated measures designs in survey experiments." *American Political Science Review*, 115(3), 1048-1065.

Clifford, S., & Wendell, D. G. 2016. "How disgust influences health purity attitudes." *Political Behavior*, *38*, 155-178.

de Benedictis-Kessner, J., & Hankinson, M. 2019. "Concentrated burdens: How self-interest and partisanship shape opinion on opioid treatment policy." *American Political Science Review*, 113(4), 1078-1084.

Gilens, M. 2001. "Political Ignorance and Collective Policy Preferences." *American Political Science Review* 95 (2): 379–96.

Griffon, J.W. 2021. "Calculating Statistical Power for Meta-Analysis Using metapower." *The Quantitative Methods for Psychology* 17 (1): 24-39.

Moore, D.W., 2003. "How Question Order Affects Attitudes on Affirmative Action." *PolTalk (on Gallup website)* (July 1).

Smith, T.W. 1987. "That Which We Call Welfare by Any Other Name Would Smell Sweeter: an Analysis of the Impact of Question." *Public Opinion Quarterly*, 51(1): 75-83.

Wilson, D. C., Moore, D. W., McKay, P. F., & Avery, D. R. 2008. Affirmative action programs for women and minorities: Expressed support affected by question order. *Public Opinion Quarterly*, *72*(3), 514-522.