Increasing Precision in Survey Experiments Without Introducing Bias[1]

Scott Clifford
University of Houston

Geoffrey Sheagley
University of Georgia

Spencer Piston
Boston University

5/30/2020

Abstract. The use of survey experiments has surged in political science as a method for estimating causal effects. By far, the most common design is the between-subjects design in which the outcome is only measured posttreatment. This design relies heavily on recruiting a large number of subjects to achieve adequate statistical power. Alternative designs that involve repeated measurement of the dependent variable promise greater precision, but are rarely used out of fears that these designs will bias treatment effects (e.g., due to consistency pressures). Across six studies, we assess this conventional wisdom by testing experimental designs against each other. Our results demonstrate that repeated measures designs substantially increase precision, while introducing little to no bias. These designs also offer new insights into the nature of treatment effects. We conclude by encouraging researchers to adopt repeated measures designs and providing guidelines for when and how to use them.

[1] We thank Alexa Bankert, Roberto Carlos, Mollie Cohen, Alex Coppock, Jamie Druckman, Christina Farhart, Jennifer Jerit, Yanna Krupnikov, Thomas Leeper, Brendan Nyhan, and Erik Peterson for helpful comments and suggestions. We thank Alexandria Putnam for valuable research assistance.

Introduction

Experiments have surged in popularity among political scientists as a method for generating unbiased estimates of causal relationships (Druckman et al. 2011). Perhaps nowhere have the benefits been clearer than in survey research. Survey experiments have allowed scholars of public opinion to isolate the causal effects of the news media (Iyengar and Kinder 1987), campaign advertisements (Valentino, Hutchings, and White 2002), political information (Zaller 1992), emotions (Banks 2014), partisan identity (Huddy, Mason, and Aarøe 2015), candidate characteristics such as race (Krupnikov and Piston 2015) and gender (Bauer 2017), and many more factors (see Mutz 2011 for a review).

However, survey experiments are only valuable insofar as the estimates of treatments effects are sufficiently precise. Estimates with wide confidence intervals provide little information to researchers. Moreover, noisy estimates are often unreliable; as a result, estimated effects that narrowly reject the null hypothesis are relatively unlikely to replicate (Open Science Collaboration 2015).

Precision in experiments can be improved either by collecting more observations or collecting more information about each observation, such as by including covariates when estimating treatment effects or by using blocking designs (for a discussion, see Bowers 2011). The most common design in political science – the between-subjects experiment in which outcomes are measured at a single point in time posttreatment – identifies a causal effect by comparing groups of subjects randomly assigned to different conditions. Precision in this design relies heavily on the collection of a large number of observations. As a result, this common design requires more resources than alternative designs and is often underpowered. For these reasons, scholars have suggested that the reliance on between-subjects designs, especially with

small sample sizes, has contributed to the replication crisis in psychology (Open Science Collaborative 2015).

Alternative experimental designs can yield greater statistical power by collecting more information about each observation. For example, pretest-posttest control group designs measure the dependent variable both before and after exposure to a treatment (e.g., Campbell and Stanley 1963), allowing an examination of how individual subjects' attitudes change throughout a study. Similarly, within-subjects designs expose subjects to multiple iterations of the same experiment. These designs increase statistical power by providing more information about subjects, but risk introducing bias in estimates of treatment effects. As Mutz (2011) points out, these types of designs are "underutilized, perhaps because of the common fear that one might arouse suspicion by asking about the same dependent variable more than once in a relatively short period of time" (p. 94). Similarly, a recent and influential paper recommends measuring pretreatment covariates in a prior survey wave and cautions researchers to "carefully separate pretreatment questions from their experiment and outcome measures to avoid inadvertently affecting the treatment effects they seek to estimate" (Montgomery, Nyhan, and Torres 2018, 773). Conventional wisdom, then, holds that there is a tradeoff between precision and bias when considering the optimal experimental design.

While there are legitimate concerns about the use of these more powerful designs in survey experiments, there is little empirical evidence regarding these concerns. Yet, political scientists heavily rely on the posttest-only between-subjects designs, as we show below, perhaps out of fears that alternative designs will bias treatment effects. If these fears are overblown, researchers have been forgoing more efficient and reliable designs. To help experimental researchers navigate this potential tradeoff between bias and precision, *we test alternative*

experimental designs against each other, by randomly assigning respondents to different experimental designs targeting the same estimand (e.g., Jerit, Barabas, and Clifford 2013).

We begin by reviewing alternative experimental designs, how these designs increase precision, and how they might introduce bias. Then, we conduct a content analysis of the experimental literature, which shows that political scientists overwhelmingly rely on the posttest-only design. Next, we conduct six studies on varied topics and samples that compare these designs against each other. Overall, including through an internal meta-analysis, we find little evidence that estimated treatment effects differ between the typical posttest-only between-subjects experiments and more powerful alternative designs. We also provide some evidence that this may be, in part, because few respondents are aware of how their attitudes change throughout a survey.

Not only do repeated measures designs not introduce bias, but they consistently provide much more precise estimates of treatment effects. We also illustrate an added benefit of repeated measures designs – the ability to examine how attitudes change throughout a study, providing new insights into treatment effect size and heterogeneous treatment effects. In short, our findings suggest that in many cases researchers can adopt more powerful experimental designs that require fewer resources and generate more precise and more informative estimates of treatment effects.

Experimental Designs in Political Science

Randomized experiments are an essential tool for political scientists seeking to estimate causal effects, and they have grown in popularity in recent years (Druckman et al. 2006; Morton and Williams 2010). Our focus here is on survey experiments, which take a variety of forms. To clarify the differences between designs, we display them graphically in Table 1, following the

notation introduced by Campbell and Stanley (1963). The first three entries – posttest, prepost, and quasi-prepost – are variations of the between-subjects design. The last entry represents the within-subjects design.

The between-subjects posttest design is the simplest and most common design in political science. In these experiments, researchers assign subjects to different groups (e.g., a treatment and a control group) and, after exposure to the experimental stimulus, measure the outcome variable. This is illustrated in the top row of Table 1, where R represents random assignment, O represents measurement of the outcome, and X represents the implementation of the treatment. The average treatment effect (ATE) is estimated by comparing the average value on the outcome for the control group to that of the treatment group. For example, a researcher might ask respondents about their support for a policy, but first expose a random half of the respondents to learn about a group's endorsement of the policy (e.g., Nicholson 2011). In practice, researchers often utilize more complicated versions of these designs, which might contain more than one treatment (e.g., Druckman, Peterson, and Slothuus 2013), omit a control (e.g., Nelson et al. 1997), or employ factorial designs that combine multiple treatments at the same time (e.g., Chong and Druckman 2007). What the posttest-only designs have in common is that they draw inferences by comparing group levels of the dependent variable measured post-treatment.

Researchers can increase statistical power in the between-subjects posttest design by adding additional information about individuals in the study. At the most basic level, researchers can include controls for covariates that are associated with the dependent variable in order to increase precision. The gains from covariates are limited, however, by the strength of the relationship between the covariates and the dependent variable. If the covariates are weakly related or entirely unrelated to the outcome of interest then there will be little to no gain in

reducing error (Gerber and Green 2012) and researchers may in fact make it more difficult to precisely estimate a treatment effect (Mutz 2011). Additionally, covariates can introduce bias if they are not independent of treatment assignment (Montgomery, Nyhan, and Torres 2018). As a result, some argue that best practice is to rely only on covariates that were measured pre-treatment (though see Klar, Leeper, and Robison 2019).

Table 1 – Comparisons of Experimental Designs

			T_1		T_2
Posttest	R			X	O_1
	R				O_2
Prepost	R		O_1	X	O_2
	R		O_3		O_4
Quasi-Prepost	R		Q_1	X	O_1
	R		Q_2		O_2
Within	R	X	O_1		O_2
	R		O_3	X	O_4

Note: R = Randomization assignment to a group. O = observation of the dependent variable. Q = observation of a variable closely related to the dependent variable. X = exposure to a treatment.

Researchers can make larger gains in precision through *repeated measures designs* that involve measuring the outcome variable at more than one point in time during a study. One such design is the pretest-posttest (or "prepost") design. The prepost is a between-subjects design that is identical to the posttest experiment save for one key difference: the dependent variable is also measured at T_1, prior to the experimental manipulation. The design is displayed in the second row of Table 1. The prepost design increases precision and statistical power through repeated measurement of the dependent variable, which provides more information about each respondent

than an equivalent posttest only design. The experiment can then be analyzed either by using a difference in change scores ($O_2 - O_1$ compared to $O_4 - O_3$) or by controlling for the pretreatment measure of O when comparing O_2 to O_4. Critically, the prepost design remains a between-subjects design because some respondents are never exposed to the treatment and respondents' T_2-T_1 difference scores are compared *between* groups.

Prepost designs promise gains in statistical power, but measuring the outcome pretreatment could bias estimated treatment effects, as discussed in detail below. This has led some researchers to propose a middle ground between standard covariate control methods and the prepost design, which is sometimes referred to as the quasi-pretest-posttest design (from here on out, the "quasi" design; Mutz 2011). This design, displayed in the third row of Table 1, follows the same structure of the prepost design with one difference – rather than directly measuring the dependent variable at T_1 (before the treatment) the researcher instead measures a closely related variable, or set of variables (Q). The goal here is to avoid the potential problems of repeated measurement of the dependent variable (e.g., bias caused by consistency pressures), while attempting to retain the gains in statistical power. Thus, researchers select a variable in advance that is strongly related to O and can serve as a proxy variable. For example, in an experiment on attitudes toward stem cell research, the researcher could measure attitudes toward cloning, which should strongly predict attitudes toward stem cell research. Similar to the prepost design, the quasi design increases statistical power by collecting more information about each respondent, though the gains will depend on how strongly Q relates to the outcome used in the study. Of course, it is possible that the quasi design introduces bias as well. For example, the quasi measures might prime considerations that are relevant to the dependent variable or create some form of consistency pressure.

The quasi design is quite similar to the posttest design, in that it uses pretreatment variables as controls in a statistical model. However, the quasi design has advantages over the typical use of posttest designs with covariates. First, by intentionally selecting and measuring a variable that is closely related to the dependent variable, the quasi design should increase precision more than whatever set of pre-treatment covariates happens to be available. Second, identifying good quasi measures requires foresight on the part of researchers. This design-based choice potentially reduces the arbitrariness of covariates, making it more amenable to pre-registration.

The final alternative design is the within-subjects experiment. While also a repeated measures design, the within-subjects design differs from the between-subjects posttest and prepost designs in that each subject is exposed to *all* experimental conditions and the dependent variable is measured after each condition (Aronson et al. 1976). The design is displayed at the bottom of Table 1. In this illustration, the dependent variable is measured twice for both groups and both groups receive the treatment and the control condition. The only difference is the order in which the conditions are administered, which allows researchers to rule out confounds between time and the treatment.[2] For example, in a study of the effects of incivility on political trust, Mutz and Reeves (2005) exposed respondents to both a civil and an uncivil version of a debate and measured respondents' physiological reactions to each. The within-subjects design

[2] If the order of conditions were not randomly assigned, it would be equivalent to a single-group design. Any effect of time or repeated measurement of the dependent variable would be confounded with the treatment, potentially biasing the estimated treatment effect.

maximizes statistical power by comparing respondents to themselves under different conditions. Thus, all individual differences are held constant in the analysis.[3]

Within-subjects designs have two main limitations. First, as we discuss in more detail below, they have the potential to introduce bias into estimated treatment effects. Second, within-subjects designs may not always be applicable to a particular research question. A within-subjects design demands that the manipulation can be "undone." Take, for example, common information experiments in political science. In the standard between-subjects design, respondents are randomly assigned to either receive a fact or not, then all are asked the dependent variable. A within-subjects design would require that we can undo that treatment, making treated subjects once again unaware of that relevant fact, so that we could again measure the dependent variable. For obvious reasons, this is not possible for information experiments, nor for a number of other theoretically interesting treatments.

In sum, posttest-only between-subjects designs offer the virtue of simplicity and the promise of unbiased estimates – which is likely why they are so common. However, the major shortcoming of these designs is relatively low precision. This shortcoming can be magnified with

[3] Recent years have seen the growth in popularity of the conjoint experiment (for an overview, see Bansak et al. 2020). These designs also feature repeated measurement of the dependent variable. Similar to within-subjects designs, conjoints expose respondents to multiple experimental conditions. However, respondents in this design are only exposed to a random subset of treatments, precluding within-subjects comparisons of each experimental condition. Additionally, conjoint designs are somewhat limited in their application – in fact, it is not clear that any of the studies in this paper could be conducted with this design.

more complex designs that involve multiple treatment conditions, moderators, or small effect sizes. Additionally, treatment effects can be more difficult to precisely estimate in demographically diverse samples (Mutz 2011) or when using abbreviated measures of the dependent variable that introduce measurement error (Ansolabehere, Rodden, and Snyder 2008). Thus, posttest-only designs necessitate recruiting a large number of subjects and therefore require greater resources. Repeated measure designs – prepost and within-subjects experiments – provide a design-based solution to this shortcoming through the additional information collected about subjects (Bowers 2011). However, these designs have the potential to bias the estimate of treatment effects due to the measurement of key outcomes or other covariates pre-treatment.

How Alternative Designs Might Introduce Bias

Based on the discussion above, prepost and within-subjects designs should be attractive methods for increasing statistical precision. However, the posttest design remains the most popular choice due to a variety of concerns about potential biases introduced by more powerful designs. In this section, we discuss some of the common concerns and the available evidence.

At the general level, repeated measures designs make an assumption of *no persistence* (Gerber and Green 2012). That is, researchers must assume that "the potential outcomes in one period are unaffected by treatments administered in prior periods." For example, in the canonical welfare question-wording experiment, this implies that being asked about support for "aid to the poor" at T_1 has no effect on how a subject later responds to a question about support for "welfare." However, as we discuss below, it is not only the delivery of multiple treatment conditions (as in the within-subjects design), but also the repeated measurement of the dependent variable itself (as in the prepost design), that might induce bias.

Demand Effects

One commonly cited threat to unbiased inferences in survey experiments is demand effects, a concern that was raised over 50 years ago (Orne 1962). As originally laid out by Orne, subjects attempt to make sense of the study they are participating in and the expectations for their behavior. An agreeable subject might then try to behave in line with the researcher's expectations, biasing treatment effects in the direction of the researcher's hypotheses. In other words, the design of the study creates a demand for hypothesis-supporting behavior.

Both prepost and within-subjects designs may increase the possibility of demand effects. Outside of an experimental design, there is little reason to ask a respondent the same question twice within a single interview. As a result, repeated measures of the dependent variable may stand out to respondents, drawing their attention to the design and goals of the experiment. The researcher's hypotheses are likely to be even more transparent when respondents are shown multiple versions of the same stimuli, as in a within-subjects design. For example, if a respondent were asked to evaluate two hypothetical scenarios involving war with a foreign country and only one detail differed between the two scenarios, the respondent might infer the researcher's hypothesis. For these reasons, researchers have often cited concerns over demand effects when considering repeated measures designs (e.g., Charness, Gneezy, and Kuhn 2012; Zizzo 2010).

However, the most recent and comprehensive study provides little evidence for demand effects (Mummolo and Peterson 2019). These authors replicated a series of experiments while manipulating the amount of information provided to respondents about the researcher's hypotheses. They find that "revealing the purpose of experiments to survey respondents leads to highly similar treatment effects relative to those generated when the purpose of the experiment

was not provided." Further, Mummolo and Peterson show that even financial incentives to support a researcher's hypothesis are largely insufficient to generate demand effects. Overall, while demand effects seem to be a reasonable concern, evidence for their impact in political science experiments is minimal.

Consistency Pressures

Another potential concern with measuring outcomes pretreatment or providing subjects with multiple versions of an experimental stimulus is that subjects may be motivated to provide responses that are consistent over time. Research on survey design shows that respondents can anchor responses on prior questions, especially when responding to questions that deal with the same issue (Schuman and Presser 1981). Psychological theories hold that people are motivated to maintain consistent beliefs and attitudes, and for others (e.g., the researcher) to perceive them as such (Cialdini, Trost, and Newsom 1995). As a result, an influential review on attitude measurement warns that "answers may undergo an editing process in which the answer is checked for consistency with prior answers" (Tourangeau and Rasinski 1988, 300). Consistency effects may be limited by subjects' memory, however. For example, in one panel study, Schuman and Presser (1996) find that only 16% of respondents correctly remembered that they had been asked an identical question a few months earlier. Overall, concerns about consistency pressure are widespread, but it is less clear how often it occurs and how long it lasts.

Interaction Between Testing and the Treatment

A general concern with measuring covariates or outcomes pretreatment centers on the potential for these measures to interact with the treatment in some way. For example, there is a large body of research on how repeatedly measuring an attitude increases its extremity (Downing, Judd, and Brauer 1992) and subjective importance (Roese and Olson 1994). Thus, in

a prepost design, the measurement of the dependent variable at T_1 may strengthen the focal attitude, making it more resistant to change. Notably, however, the effects of repeated measurement on attitude strength tend to emerge when an attitude is measured many times, rather than just two or three. A related concern is that the pretest measurement may increase the accessibility of certain considerations, which may affect how a subject responds to the treatment.

These concerns are often discussed in canonical texts on experimental design. For example, Aronson et al. (1976, p. 141) note that, "research on pretest sensitization indicates that its usual effect is to reduce the power of the independent variable to create change, so that investigators may erroneously conclude that their variable has no effect," although it is not clear what research the authors are referring to when making this claim. Campbell and Stanley (1963) discuss these concerns at various points in their essays, emphasizing in particular that researchers should avoid pretests when studying questions related to attitude change as the pretest may affect their susceptibility to treatment.

The general concerns about interactions between pretests and treatments highlight potential problems with the external rather than internal validity of a study. So long as covariates and other variables are measured independently of treatment, often before treatment occurs, the study will be internally valid. However, measuring a covariate or set of covariates creates a particular context for a given experiment – one that may differ from the context to which researchers hope to generalize. In short, pretests may interact with treatment effects such that results from a pretested experiment do not generalize to a population that was not pretested.

That scholars avoid using within-subjects or prepost experiments makes sense given these concerns. Scholars often prioritize the unbiased estimation of a treatment effect above all other concerns when selecting an experimental design. However, the most recent evidence on

demand effects fails to find any evidence of their presence across a variety of experimental contexts (Mummolo and Peterson 2019). Furthermore, while there is certainly potential for consistency effects, interaction effects, and possibly others, there is also scant evidence that these biases affect the types of studies commonly conducted by political scientists. In short, it is not yet clear whether the potential bias from using within-subjects or prepost designs outweighs the benefits these designs offer for precision.

Experimental Design Practices in Political Science

Having identified the possible costs and benefits of each type of design, we now turn to how the field has used these designs in practice. To do so, a research assistant identified all articles using experimental methods published between 2015-2020 in five major journals: *American Journal of Political Science, American Political Science Review, Journal of Politics, Political Behavior,* and *Political Psychology*. This yielded a population of 457 articles. We then selected a random sample of 55 articles to code in detail, and retained only those using survey experiments, leaving 41 articles with 67 studies. The authors coded each study for the type of design used (posttest, quasi, prepost, or within-subjects) and the use of covariates (see Appendix for coding details). The results are shown in Table 2.

Overall, 82% of the studies in our sample were posttest-only designs, demonstrating the dominance of this design in political science. Of these studies, 60% used control variables in at least some analyses. This suggests that many modern posttest-only designs in political science do take some steps to leverage additional information about observations. Only five studies (7%) used prepost designs, however. Of these five, three included the pretest measure in a prior wave of the survey, presumably to protect against bias. One of the two studies using a prepost design

within a single wave included a heavy caveat, stating that "Due to the limited amount of time between pretest and posttest measurements, it is possible that some subjects anchored their posttest response on their pretest response resulting in no change" in the dependent variable (Andrews et al. 2017). In short, prepost designs are rare, typically used in panel designs, and researchers are clearly concerned about the potential for bias.

Table 2 – Frequency of Experimental Designs

Design	**% (n)**
Post-test only	82% (55)
No controls	40% (22)
Controls	60% (33)
Quasi	4% (3)
Pretest-posttest	7% (5)
Within-subjects	7% (5)

Note: Does not add up to 100% because one study was classified as both pre-post and within-subjects. Number of studies in parentheses. Total N=67.

While prepost studies are rare, only three studies (4%) utilized quasi-prepost designs, in spite of being promoted as a method to increase precision while reducing the chance of bias (Mutz 2011). Thus, while researchers seem concerned about the use of prepost designs, few seem to be taking up a close alternative.

Finally, only five studies (7%) used a within-subjects design. However, three of these studies were conjoint designs and a fourth was quite similar. Thus, there was only one true within-subjects design, suggesting that within-subjects designs are also rarely used in political science. Overall, the content analysis affirms our claim that the posttest design is dominant in the discipline and that conventional wisdom holds that alternative designs will introduce bias.

Overview of Studies

To test the effects of repeated measures designs on precision and bias, we conducted six experiments that involve randomly assigning respondents to alternative designs. Each experiment roughly replicates a past study on topics that cover a range of common experimental paradigms, including question wording effects, information effects, partisan cues, and framing. In each study respondents are randomly assigned to one of up to three experimental designs. Every study includes a standard, between-subjects, posttest-only design (or posttest design, in brief) as the baseline for comparison. We use the treatment effect from this design as the benchmark under the assumption that the lack of relevant pretreatment measures means it produces the most unbiased estimates, and also because it is the most common survey experimental design in extant scholarship.[4] In addition, respondents are randomized into one of three alternative designs: a prepost design, a quasi design, or a within-subjects design. In all studies, the experiments were placed near the end of a larger survey and all pretest and quasi-pretest measures were placed near the beginning of the survey. In short, we sought to maximize the distance between repeated measurements to the extent possible in a standard survey. The six studies are summarized in Table 1. We review each below and additional details are available in the Appendix.

[4] To put it in terms of potential outcomes, we are assuming that $E[Y_i(1, \text{prepost}=0) - Y_i(0, \text{prepost}=0)]$ is an unbiased estimate of the ATE and testing whether $E[Y_i(1, \text{prepost}=1) - Y_i(0, \text{prepost}=1)]$ is equivalent.

Table 3. Overview of Experimental Studies

Study	Topic	Manipulation	Sample Source	Dates	Sample Size	Posttest	Within-Subject	Prepost	Quasi
1	Welfare	Question wording	Student	Spring 2018	900	X	X		
2	Foreign aid	Information	MTurk	March 2018	1,209	X		X	
3	Education	Information	MTurk	May 2018	1,206	X		X	X
4	Estate tax	Information	Lucid	Spring 2018	2,462	X		X	X
5	Prescription drugs	Party cues	Forthright	July 2019	1,531	X		X	X
6	GMOs	Framing	Student	Spring 2020	965	X		X	X

Study 1 – Welfare Question-Wording Experiment

Our first study sought to replicate the canonical welfare question-wording experiment (Smith 1987). For the key manipulation, respondents were randomly assigned to either a question asking about spending levels on "welfare" or to a question about "assistance to the poor." Responses were recorded on a three-point scale ("too much," "about the right amount," "too little"). To test for design effects, we randomized respondents into either a posttest design or a within-subjects design. In the posttest design, respondents were randomly assigned to either the welfare or the poor condition near the end of the survey. In the within-subjects design, respondents were randomly assigned to either the welfare or the poor condition early in the survey, then received the other question near the end of the survey.

Study 2 – Foreign Aid Information Experiment

In our second study, we sought to replicate a landmark information experiment about foreign aid (Gilens 2001). For the key manipulation, a random half of respondents were informed that spending on foreign aid makes up less than one percent of the federal budget. All respondents were then asked whether spending on foreign aid should be increased or decreased (on a five-point scale). Respondents were randomly assigned into either the posttest or pretest design.

Study 3 – Education Spending Information Experiment

In our third study, we conducted another information experiment; this one included a test of both the prepost and the quasi design. For the substantive manipulation, a random half of respondents were informed of the average annual per pupil spending on public schools. All respondents were then asked if taxes to support public schools should be increased or decreased (for a similar experiment, see Schueler and West 2016). In this study, respondents were randomized into either the posttest, pretest, or quasi design. For the quasi measure, respondents were asked whether they supported increasing or decreasing teacher salaries.

Study 4 – Estate Tax Information Experiment

The substantive manipulation in our fourth study involved informing a random half of respondents that the estate tax applies to those with an estate over $11.18 million, which makes up the wealthiest 0.0006% of Americans (cf., Piston 2018). All respondents were then asked whether they favor or oppose the estate tax on a seven-point scale. Respondents were randomized into a posttest, prepost, or quasi design. For those assigned to the quasi design, there were two quasi measures: whether Congress should reduce the budget deficit through spending cuts or tax increases, and whether the country would be better off if we worried less about how equal people are.

Study 5 – Prescription Drugs Party Cue Experiment

Our fifth study focused on party cues. In this study, we expected a greater likelihood of consistency pressures or other sources of bias, as compared to information experiments. This is because most respondents likely see changing one's mind in response to new information as more normatively desirable than changing one's mind in response to a party cue. Indeed, many people want to avoid being seen as partisan (Klar and Krupnikov 2016). And when asked about

the sources of their attitudes, people report that policy content is much more important than partisan endorsements (Cohen 2003). Thus, if following a party cue is normatively undesirable, respondents should experience consistency pressure and be less likely to follow the cue, leading to muted treatment effects. In contrast, in the posttest design, respondents have not reported a prior attitude that would create any consistency pressure.

The party cue experiment focused on support for allowing the importation of prescription drugs from Canada. For the treatment, a random half of respondents were informed that "Democrats tend to favor and Republicans tend to oppose" the policy, while this information was omitted from the control condition. Respondents were randomized into one of three designs: posttest, prepost, or quasi. For the quasi measure, respondents were asked if they support or oppose making it easier for people to import prescription drugs from other countries.

This experiment was embedded in the second wave of a panel study, allowing additional tests. First, we measured the dependent variable in the first wave of the survey for respondents in all experimental conditions. The first wave was administered approximately one month prior to the second wave, making it unlikely this measurement had any effect on the experiment. Thus, respondents in the prepost condition answered the same question about prescription drugs three times: once in wave 1, once at the start of wave 2, and again at the end of wave 2. This allows us to control for wave 1 pretreatment attitudes in all conditions to increase statistical power, while presumably avoiding any potential design effects (for discussion, see Montgomery, Nyhan, and Torres 2018).[5] A second benefit of the panel design is that it enables us to test whether the

[5] Of course, this is an assumption. We cannot be sure that measuring an outcome roughly one month before the relevant experiment avoids the previously discussed concerns about bias.

presumed gains in precision due to using the prepost design differ based on when the dependent variable is measured (i.e., in Wave 1 or Wave 2).

Study 6 – GMOs Framing Experiment

Our sixth study focused on framing effects on the topic of GMOs (e.g., Druckman and Bolsen 2011). Respondents were randomized to receive either a pro-GMO frame focusing on how foods can be modified to be more nutritious or an anti-GMO frame focusing on potential harmful health effects.[6] The dependent variable measured support for the production and consumption of GMOs on a seven-point scale. Respondents were randomized into either the posttest, prepost, or quasi design. For the quasi variables, respondents were asked two questions about their support for banning the use of chemical pesticides and the use of antibiotics on livestock, both of which tend to be related to attitudes toward GMOs (Clifford and Wendell 2016). Additionally, to test for the possibility of consistency bias, we included an item at the very end of the study asking respondents in the prepost condition to report how they believed their attitude had changed throughout the study.

How Design Influences Estimates of Effect Size

In this section, we analyze each experiment with a focus on treatment effect size, taking on the topic of precision in the next section. To this end, within each study we analyze each design separately, making use of covariates to increase statistical power for comparisons of treatment effect sizes. To maintain similarity, we analyze prepost designs by controlling for T_1 measures, rather than modeling difference scores (for discussion, see Blair et al. 2019; Gerber

[6] There was no pure control condition as our goal was to maximize treatment effects.

and Green 2012). Specifically, we control for partisanship and ideology in all designs, as well as pretest and quasi measures, as available. All effects are plotted in Figure 1. We compare effect sizes across models using a Wald test.

Study 1 – Welfare Question-Wording Experiment

Starting with the standard between-subjects design, respondents receiving the "welfare" wording were significantly less supportive of spending (b = -.25, p < .001) than respondents receiving the "poor" wording. Turning to the within-subjects design, we analyze the results using a paired *t*-test. Respondents again expressed less support in the "welfare" condition than in the "poor" condition (b = -.28, p < .001). Thus, the two designs yielded effects of similar magnitude that are statistically indistinguishable (p = .611).[7]

Study 2 – Foreign Aid Information Experiment

In the posttest design, respondents receiving the treatment were significantly less supportive of cutting foreign aid (b = -.34, p < .001) than respondents not receiving the treatment. In the prepost design, the treatment again significantly reduces support for cutting foreign aid (b = -.13, p = .002). However, the treatment effect is significantly smaller in the prepost design than in the standard posttest only design (p = .021).

[7] To test the equality of coefficients, we stacked the data so that respondents in the within-subjects condition provided two observations. We then estimated a regression model with respondent random effects, a treatment dummy, a dummy for prepost design, and an interaction between the two.

Study 3 – Education Spending Information Experiment

The standard posttest design yielded a substantively small treatment effect that is not distinguishable from zero (b = .06, p = .517). Because we are typically concerned that the prepost and quasi designs might *reduce* the effect size, this study is less informative on the question of bias. Nonetheless, treatment effects were similarly null in both the prepost (b = -.03, p = .439) and quasi designs (b = .12, p = .170), and neither of these effects differed from the effect in the posttest design (p = .354, p = .642, respectively). In any case, as we discuss below, this study is still informative for how designs affect the precision of estimates.

Study 4 – Estate Tax Information Experiment

In the posttest design, the treatment increased support for the estate tax (b = 1.03, p < .001). The treatment had a similar effect in both the prepost design (b = 1.15, p < .001) and the quasi design (b = 1.10, p < .001). In contrast to Study 2, the effect in the posttest design did not differ from the effect in the prepost design (p = .500) or the quasi design (p = .737).

Study 5 – Prescription Drugs Party Cue Experiment

Because the effects of party cues should be moderated by partisan identity, we take a different modeling approach for Study 5. We regress policy support on a treatment indicator, a dichotomous indicator of partisan identity (with pure independents excluded from the analysis), and an interaction between the two. To simplify our discussion, we focus on the interaction term, which indicates how much the treatment increased partisan differences in policy support.[8] As expected, the treatment significantly increased partisan disagreement in the posttest design (b =

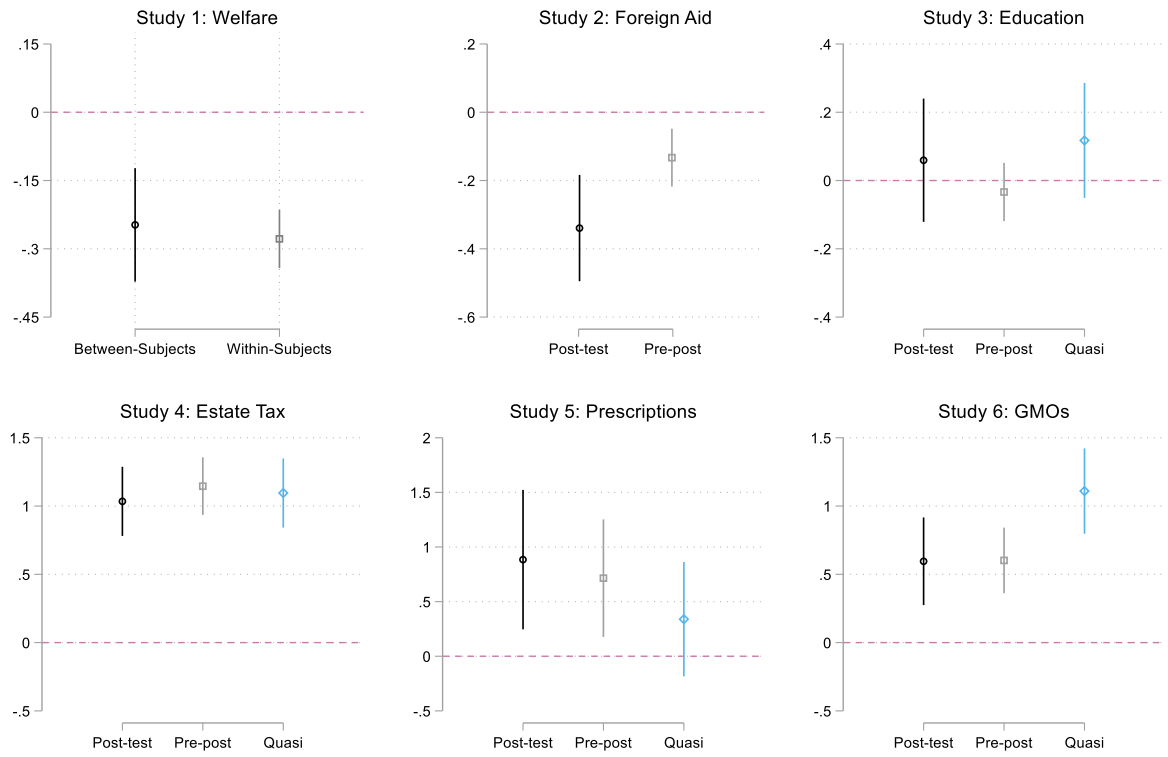
[8] This approach has the benefit of avoiding assumptions about partisan symmetry in responsiveness. Alternative modeling approaches yield substantively identical results.

.88, p = .007) and the pre-post design (b = .71, p = .009). In the quasi design, the effect on partisan disagreement was also in the expected direction, but not statistically significant (b = .34, p = .203). However, the effect in the posttest condition was not distinguishable from the effect in the other conditions (prepost: p = .682; quasi: p = .182).

Study 6 – Framing GMOs Experiment

In the posttest design, the pro-GMO frame increased support for GMOs, relative to the anti-GMO frame (b = .60, p < .001). The effect was quite similar in the prepost design (b = .60, p < .001), but somewhat larger in the quasi design (b = 1.11, p < .001). While the effects in the posttest and prepost designs did not significantly differ from each other (p = .975), the effect was significantly larger in the quasi design (posttest: p = .023, prepost: p = .010).

Figure 1. Treatment Effects by Experimental Design



Internal Meta-Analysis

Across six studies, we found little evidence that repeated measures designs distort treatment effects. But it is possible that there is a relatively small and systematic bias that could not be detected in any single study. We address this problem with an internal meta-analysis, which provides a precision-weighted estimate of the average effect of experimental design across all six studies (for discussion, see Goh, Hall, and Rosenthal 2016). First, we rescaled the dependent variable in each study to range from zero to one, and recoded the direction of the variable so that all treatment effects carry the same sign. Then, within each study, we estimated the difference in treatment effects between the posttest design and the repeated measures design. These six differences in treatment effects represent the observations in our internal meta-analysis.[9]

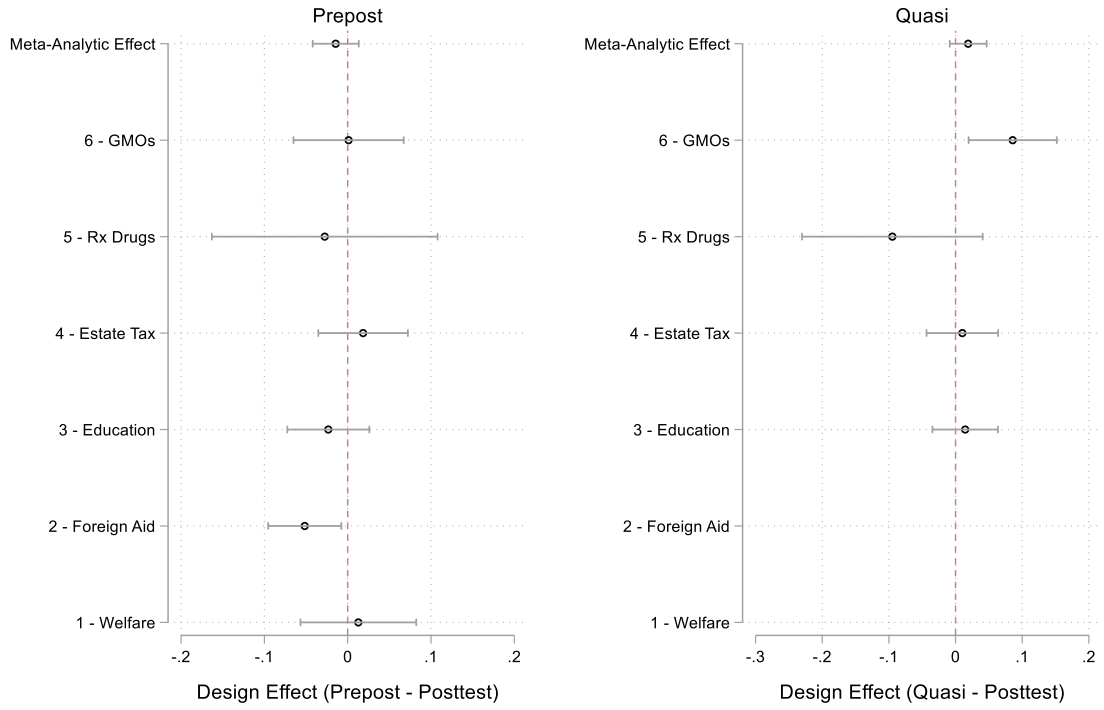
Figure 2 summarizes how repeated measures designs influence treatment effects, relative to the posttest design, for each study. Negative values indicate that the repeated measures design led to smaller treatment effects than the posttest design, which would be consistent with common concerns. The left-hand panel shows the effects of the prepost design, while the right-hand panel displays the effects of the quasi design. The top row of both panels displays the meta-analytic effect.[10] The meta-analytic effect of the prepost design is -0.014, which does not significantly

[9] We also took an alternative approach in which each of the 16 treatment effects constituted our dependent variable in the meta-analysis, which we modeled as a function of design and study. The effects are substantively similar. See Appendix for details.

[10] Meta-analyses were conducted using the metareg package in Stata (Harbord and Higgins 2008).

differ from zero (p = .355).[11] For reference, the meta-analytic treatment effect *within* the six posttest experiments is 0.104, implying that the average effect in the prepost design is roughly 86% of the size of the average effect in the posttest design.[12] Again, this small effect cannot be distinguished from zero.

Figure 2. Meta-Analysis of Design Effects



[11] The low I^2 for the model, 2.6%, suggests that variation in design effects between studies is overwhelmingly due to sampling error, rather than variance in the size of design effects.

[12] Additionally, if we exclude Study 3, on the grounds that this study did not yield a significant treatment effect in any condition, the results are stronger. With this exclusion, the design effect remains statistically insignificant (p = .581) and the effect within the prepost design is 91% of the size of the effect in the posttest design.

Turning to the quasi design, the effect is positive (0.021), implying that it leads to larger effects than the posttest design. But it cannot be distinguished from zero either (p = .333). Overall, these results suggest that repeated measures designs tend to yield the same substantive results.

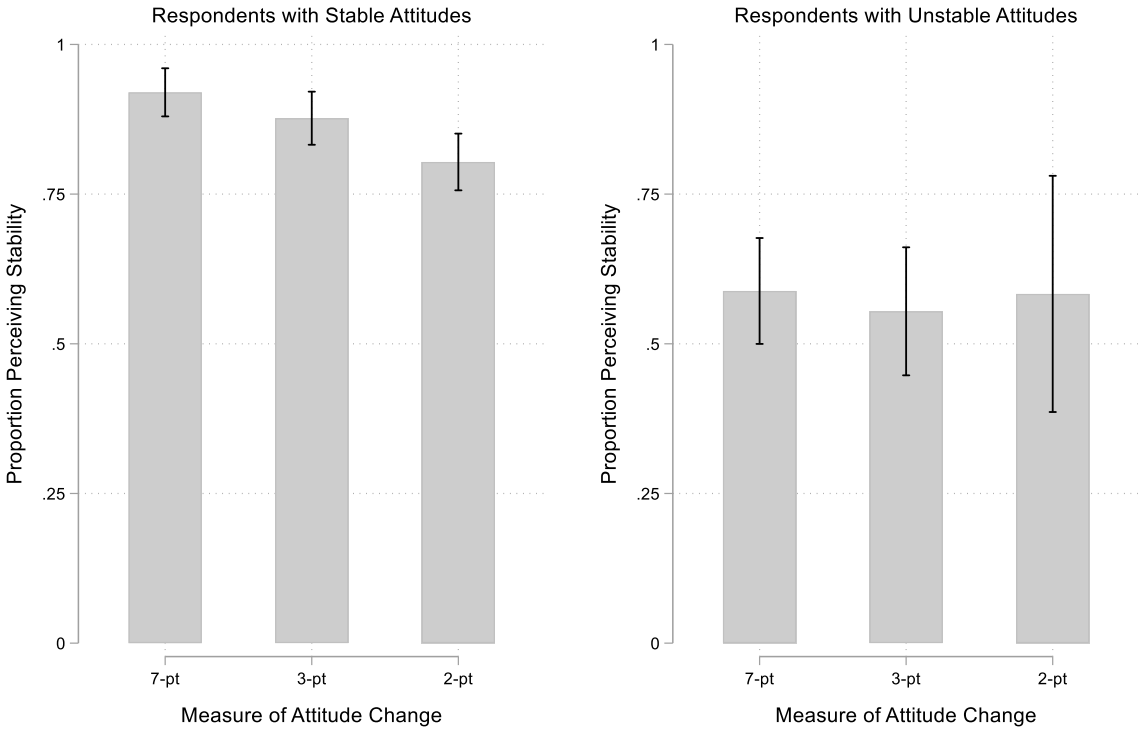
Are Respondents Aware of their Attitude Change?

Our results provide little evidence that repeated measurement of the dependent variable biases treatment effects. One possible explanation for this pattern is that respondents may not remember their initial stance, eliminating consistency pressures.[13] We tested this possibility in the prepost condition of Study 6, which included a measure of respondents' perceptions of their attitude change throughout the course of the study. Immediately following the second measurement of the dependent variable, all respondents in the prepost condition received the following question: "As you may remember, we also asked you about your support for genetically modified foods at the beginning of the survey. To the best of your memory, how has your support for genetically modified foods changed since the beginning of the survey?" Response options were increased (1), decreased (-1), or stayed the same (0) since the beginning of the survey. We compared self-reported change to actual change from the pretest to posttest measure. Because respondents might interpret change in different ways, we operationalized attitude change in three different ways: 1) any change on the seven-point scale, 2) any shift between favor, oppose and the scale midpoint, and 3) any change from favor to oppose or vice-versa.

[13] Of course, it is possible that consistency pressures do not require memory of the initial position, but this is a common assumption (Gerber and Green 2012, 256).

Figure 3 displays the proportion of respondents who perceived their attitudes as stable among those whose attitudes were actually stable (left panel) and among those whose attitudes actually changed (right panel). The three bars in each panel represent the three different operationalizations of attitude change described above. Among those with stable attitudes, between 80% and 92% of respondents correctly perceived their attitudes as stable, depending on the particular measure. However, among respondents whose attitudes *did* change, between 55% and 59% of respondents also perceived their attitudes as stable. Overall, most respondents believed their opinions did not change, even when they did. This weak relationship between perceptions and actual change creates an upper limit on any potential consistency effects.

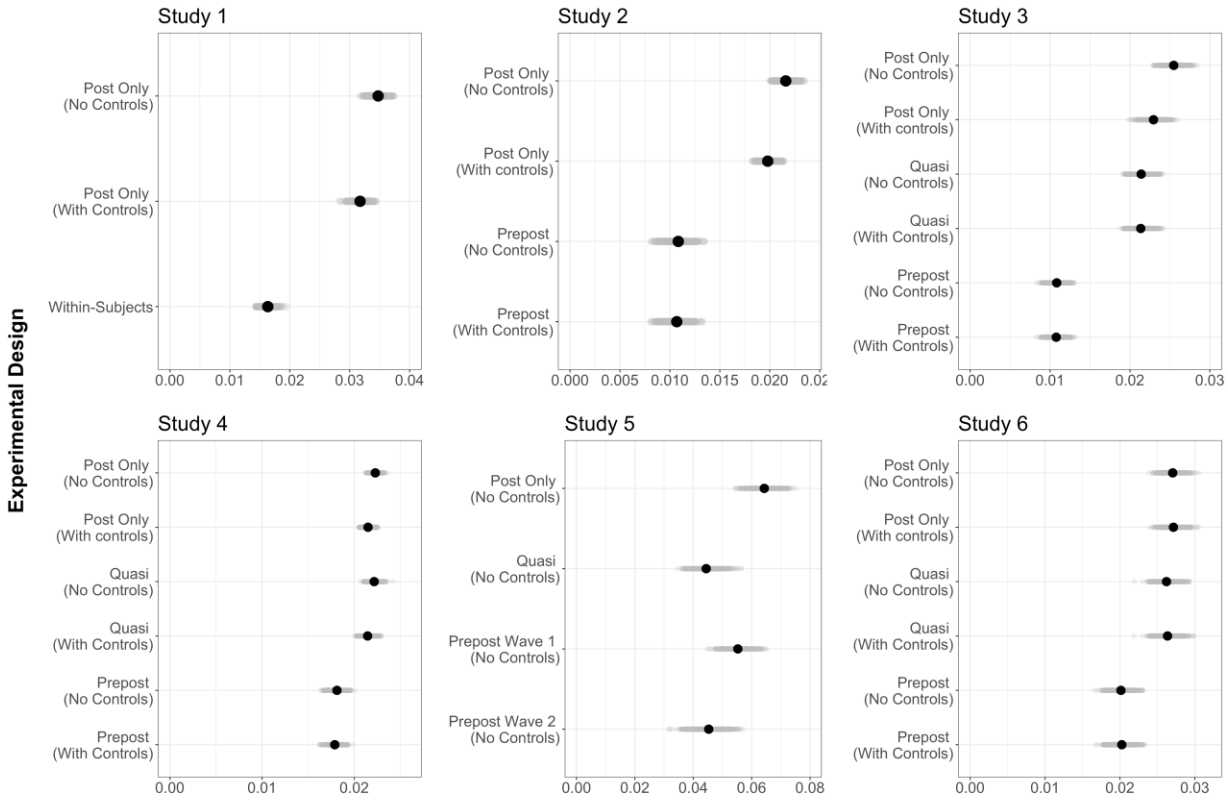
Figure 3. Perceptions of Attitude Change by Actual Change



How Design Influences Precision

In this section we assess how different experimental designs and analysis strategies affect the precision with which we estimate treatment effects. To facilitate comparison, we rescale all dependent variables to range from 0 to 1. In Figure 4, we plot the standard error of the estimated treatment effect for each design and for alternative analysis strategies for each study. To illustrate the uncertainty in estimates of the standard error, we plot bootstrapped estimates of the standard error (gray) and the mean estimate (black).

Figure 4. Standard Errors of the Estimate by Study, Design, and Analysis



As expected, in all six studies, the posttest design had the treatment effect estimate with the largest standard error. Including control variables (partisan identity and self-placement

ideology) typically results in slightly smaller standard errors, although the degree of change depends on the study. The reductions are effectively zero in Study 4 and Study 6, but more apparent in the other studies. The quasi designs tend to result in smaller standard errors than the posttest designs, even when including controls, although the reductions vary by study. In Studies 3 and 5, the quasi design yields sizable reductions in the standard error. However, the gains are more modest in Study 6, and in Study 4 the size of the standard error in the quasi design is larger than the posttest design with controls. That said, all six studies show that repeated measures designs yield substantially more precise estimates. In study 1, the within-subjects design clearly yields the greatest precision. In Studies 2 through 6, the prepost designs consistently have much smaller standard errors.

Study 5 also allows a comparison between measuring the pretest variable in a prior wave (Wave 1) or in the same wave as the experiment (Wave 2). This question is important, as the majority of prepost designs in our content analysis of the literature used a panel design, which is costly and raises concerns about attrition. Our results show that measurement within the same wave leads to clear gains in precision, suggesting that panel designs yield weaker benefits. Overall, clear patterns emerge. Standard political controls can lead to small gains in precision and quasi controls sometimes improve upon these gains. However, repeated measures designs consistently yield substantially more precise estimates.

How Design Choice Affects Statistical Power

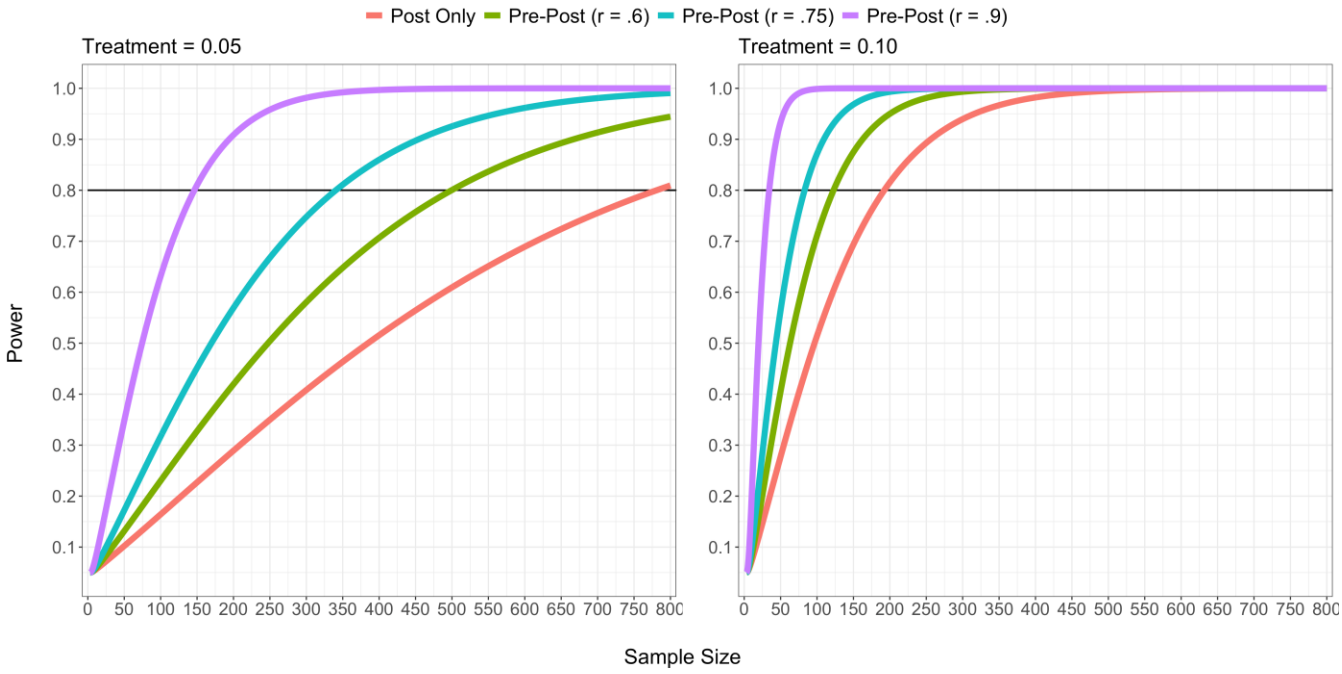
There are clear gains in precision from the prepost design, while the benefits are less consistent for the quasi design, and minimal for basic political controls. In this section, we further examine how the use of the prepost design affects statistical power and how this varies by

sample size and effect size. Additionally, we investigate how much bias would have to be introduced by the prepost design to offset the gains to statistical power from increased precision.

The two panels in Figure 5 plot statistical power on the y-axis with sample size on the x-axis.[14] The left panel corresponds to an effect size of .05, which is the smallest effect size observed in our experiments (Study 2). It is roughly equivalent to a quarter of a scale-point on a five-point scale, or a Cohen's d of 0.2. The red line represents estimates for the posttest design. With this effect size and design, a total sample size of 780 is required to reach a common standard of 80% power. The remaining three lines depict power for the prepost design under three different assumptions about the gain in precision from controlling for the pretreatment measure of the outcome. The green line represents the smallest gain in precision observed in our studies (a correlation between the pretest and posttest measures of $r = .60$), the purple line represents the largest gain in precision ($r = .90$), and the blue line falls in between ($r = .75$). Even with the weakest gains, the prepost design achieves 80% power at $N = 499$, a substantial reduction from the 780 required by the posttest design. In the most optimistic scenario, the prepost design requires only 147 respondents to achieve 80% power. Thus, with a small effect size, which is likely common in political science, prepost designs require a substantially smaller number of participants.

[14] Power was calculated using the retrodesign function (Gelman and Carlin 2014). For all of these analyses, we fix the standard deviation of our outcome to .25. This value is roughly the average of what we observed in our studies, which had standard deviations ranging from .23 (Study 4) to .28 (Study 2). See Appendix for additional details about the simulation procedure.

Figure 5. Simulated Statistical Power by Design-Type and Sample Size



Of course, if the prepost design also introduces bias in the estimated coefficient, this might offset the gains in power. If common concerns are correct, and the prepost design reduces the effect size, it will also reduce the statistical power of the test. To investigate potential tradeoffs between bias and precision in achieving adequate statistical power, we conducted a series of simulations in which we varied the amount of bias introduced by the prepost design. Specifically, we held all other variables constant, while shrinking the size of the coefficient in the prepost design, relative to the posttest design. Our simulations suggest that with the weakest gains in precision from the prepost design (r = .6), the prepost design will yield the same level of statistical power as the posttest design when the prepost design reduces the effect size to only 80% of the size of the effect in the posttest design. Under the assumption of moderate gains in precision (r = .75), the prepost and posttest designs achieve equal power when the prepost design reduces the effect size to 65%. And under the assumption of large gains (r = .9), the designs

reach equal power when the effect size is reduced to 43% of the original effect size. Notably, these three points at which designs reach the same power are unaffected by assumptions about the original effect size or the sample size.

Overall, these results make clear the potential gains from prepost designs. When no bias is induced, an assumption supported by our empirical results above, the prepost design yields substantially greater power, requiring much smaller sample sizes. This holds true for a variety of effect sizes and for a range of empirically-based assumptions about the precision gains from prepost designs. Finally, our results show that the prepost design would have to induce substantial bias, reducing the effect size to between 80% and 43% of the original effect size, to fall to the same statistical power as the posttest design.

Additional Benefits of the Prepost Design

In addition to greater statistical precision, prepost designs also allow the opportunity to gain further insight into the nature of treatment effects. Scholars have long been interested in treatment effect heterogeneity. For example, researchers often examine moderators of treatment effects under the expectation that the size or direction of effects depend on respondent characteristics (for discussion, see Kam and Trussler 2016). Treatment effect heterogeneity has also been central to debates over the generalizability of convenience samples (e.g., Druckman et al. 2011). If treatment effects are largely homogeneous, then they will generalize across different populations. Finally, the literature on motivated reasoning has proposed a backlash effect in which some respondents move in the opposite direction of the treatment (e.g., Nyhan and Reifler 2010), which is fundamentally an issue of treatment effect heterogeneity. Prepost designs offer a

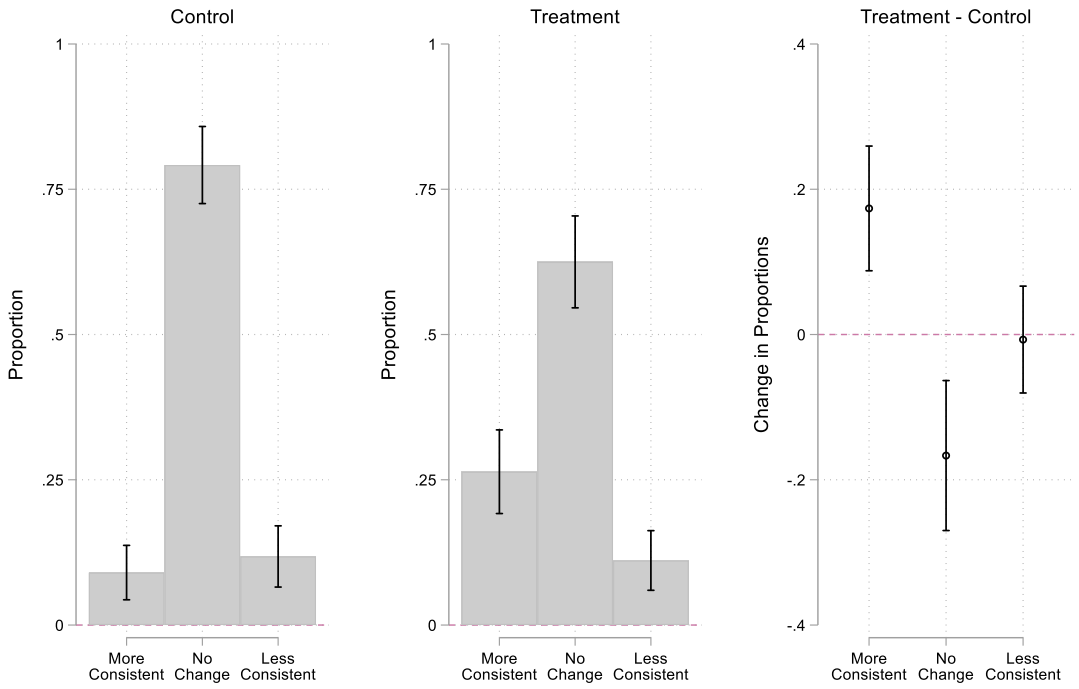
closer look at treatment effect heterogeneity by allowing an analysis of how respondents change their opinions throughout an experiment.

We illustrate these benefits with an analysis of the prepost arm of our party cues experiment (Study 5). In the section above, we reported evidence that partisans, on average, move toward their party's position. This average effect is consistent with a pattern of homogeneous treatment effects, in which most or all partisans undergo a small change in opinion. However, it is also consistent with a variety of patterns of heterogeneous effects, such as a large movement among a small number of partisans and no movement among most partisans. In other words, the standard posttest design cannot tell us how broadly party cues affect partisans and how large the effect is among those who are responsive.

To examine treatment effect heterogeneity, we turn to analyzing difference scores (posttest minus pretest). Using respondents' own pretreatment measure of party identification, we then classify respondents into three groups based on their difference score: those whose opinions became *more* consistent with the party's position, those whose opinions became *less* consistent with the party's position, and those whose opinions *did not change* throughout the study. The proportion of respondents falling into each group in the control condition are shown in the left-hand panel of Figure 3. Most respondents, 79%, did not change their opinions at all over the course of their study. But 9% become more party-consistent, and 12% became less party-consistent, likely due in part to measurement error or satisficing. The middle panel of Figure 3 shows the results for the treatment condition. Here, 63% did not change their opinion, but 26% moved toward the party's position and 11% moved away. Of course, these latter patterns are a combination of the treatment effect, measurement error, and any effects of time and other survey content. However, we can estimate the proportion of respondents that the

treatment caused to undergo each type of attitude change by differencing the proportions between the treatment and control condition, which is shown in the right-hand panel of Figure 6. The party cue treatment increased the percentage of respondents moving toward their party's position by 17 percentage points. In other words, only 17 percent of respondents followed the party cue! The treatment also decreased the percentage of respondents with stable attitudes by about 17 points, and had no discernible effect on the percentage of respondents moving away from their party's position. This latter finding suggests that there was no backlash to the treatment in this case.[15]

Figure 6. Most Respondents are Unaffected by Party Cues



[15] We would not theoretically expect to find backlash to party cues, but this serves to illustrate how the analysis could be useful in searching for backlash effects. Notably, we also find no evidence of backlash in the foreign aid or estate tax information experiments.

Of course, it could be that few partisans shifted their opinions in response to the treatment because they already supported their party's position. The pretest variable also allows us to examine this possibility. As it turns out, 42% of respondents strongly favored the policy, removing the possibility of becoming more supportive. Following the procedure described above, we can estimate that among respondents who initially opposed their party's position (nearly entirely Republicans), the treatment increased the percentage of respondents moving toward their party's position by 33 points ($p < .001$). However, most of this change is in attitude *intensity*, rather than in attitude *position*. Among those initially opposing their party's stance, the treatment caused only about 8% (CI: -5% to 20%) to move to the midpoint and 5% (CI: -4% to 13%) to switch to supporting their party's stance. Overall, the results suggest that, even on a relatively complex and low-salience issue, most partisans do not change their opinions when exposed to a party cue. And most of the opinion change that does occur is in intensity, while very few partisans actually change positions. Of course, this is hardly the last word on the subject of party cues. But our analysis demonstrates that prepost designs offer new insights into treatment effects that would be missed in a standard posttest design.

Conclusion

Scholars conducting survey experimental research should seek to minimize bias while maximizing precision. Researchers, however, have overwhelmingly opted for a posttest design that relies heavily on large sample sizes, ignoring concerns about precision in order to avoid bias. This design choice is supported by conventional wisdom that more statistically powerful designs are likely to introduce a variety of biases. Yet this conventional wisdom has not been thoroughly tested in political science. Across six experimental studies, each of which was based on a

common framework used in applied political behavior research, we find that prepost and within-subjects designs can offer dramatic improvements in statistical power, and little evidence that these designs introduce bias into estimated treatment effects. Additionally, in Study 6 we found evidence that most respondents were unaware that their attitudes had changed, potentially placing a limit on consistency effects. As a result, it seems that conventional wisdom has been too conservative, leading researchers to devote more resources to weaker designs.

While researchers have long relied on covariate controls to increase the precision of treatment effect estimates, this practice does not seem to consistently pay off. Across six studies, controls for partisanship and ideology led to gains in precision that were small in magnitude; these gains were dramatically outperformed by the gains from prepost and within-subjects designs. The evidence was less consistent for quasi control variables that were designed to closely relate to the dependent variable. The benefits ranged in size from similar to the gains from standard controls (Study 4) to being on par with the large gains from a prepost design (Study 5). Thus, the gains from controls seem to depend heavily on the quality of the measures selected and the specific design. We recommend that researchers employing a quasi design select multiple items for controls, and that they use pre-existing datasets to identify a set of controls that maximizes predictive power over the dependent variable.

Given the clear gains in precision and weak evidence for bias, we recommend that researchers use prepost and within-subjects designs whenever possible. Not only are these designs more powerful, but they also offer deeper insight into the topic of study by allowing a detailed examination of treatment effect heterogeneity, as illustrated with our study on party cues. Of course, our results are limited to only six studies, and we cannot be sure how they generalize to other topics. However, our studies covered several common experimental

paradigms and were applied to many different substantive topics. Additionally, our studies used a variety of subject pools, including respondents from Mechanical Turk, who may be particularly suspicious of researchers' intentions (Krupnikov and Levine 2014). Taken together, our evidence suggests repeated measures designs can offer dramatic gains in precision and require fewer resources.

Of course, there are some instances in which a standard posttest design may be the best option. Within-subjects designs are limited to cases in which the effect of the independent variable can be removed. As a result, research on information effects, for example, is not easily amenable to a within-subjects design. There may also be cases in which prepost designs do in fact lead to bias. In all of our studies, we sought to maximize the distance between the pretest measure and the experiment. As a result, we cannot be sure that our findings would hold if the pretest measure of the dependent variable had to be placed immediately before the experiment. We also suspect that bias is particularly likely when studies are being conducted on sensitive topics or when respondents may perceive treatment effects as normatively undesirable. For example, researchers are often wary of measuring racial attitudes prior to an experiment out of fear that it may prime racial considerations (Klar, Leeper, and Robison 2019), though some work finds no support for this concern (Valentino, Neuner, and Vandenbroek 2018). In our view, future research would do well to investigate the conditions under which consistency pressures are most likely to influence responding, biasing estimates of treatment effects.

When researchers are particularly concerned about bias from repeated measures designs, quasi designs appear to be the best option. In these cases, there is no need for the covariates to be causally related to the dependent variable. As such, pre-existing data can be mined in order to identify variables that have tight relationships with the outcome of interest. For example, in our

replication and extension of the estate tax experiment conducted by Piston (2018), we used his publicly available data to find two independent variables that jointly maximized predictive power over the dependent variable. While our quasi controls varied in their effectiveness across studies, overall the results suggest that carefully selected controls can substantially increase statistical power relative to standard controls for partisanship and ideology.

Ultimately, we believe that researchers would do well to acknowledge that both bias and precision are important. Unfortunately, current design practices heavily emphasize bias without acknowledging precision, requiring researchers to justify any deviation from the standard posttest design. In contrast, we believe researchers must justify their design choice both in terms of bias and precision, regardless of which design they choose. Fortunately, our results suggest that there is often little tradeoff between the two. In our view, therefore, the default should shift away from the posttest design and toward repeated measures designs.

References

Andrews, Amelia C., Rosalee A. Clawson, Benjamin M. Gramig, and Leigh Raymond. 2017. "Finding the Right Value: Framing Effects on Domain Experts." *Political Psychology* 38(2): 261–78.

Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102(02): 215–32.

Aronson, Elliot, Phoebe C. Ellsworth, J. Merrill Carlsmith, and Marti Hope Gonzales. 1976. *Methods of Research in Social Psychology*. McGraw-Hill.

Banks, Antoine J. 2014. *Anger and Racial Politics: The Emotional Foundations of Racial Attitudes in America*. New York: Cambridge University Press.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2020. "Conjoint Survey Experiments." In *Cambridge Handbook of Advances in Experimental Political Science*, eds. James N. Druckman and Donald P. Green. Cambridge University Press.

Bauer, Nichole M. 2017. "The Effects of Counterstereotypic Gender Strategies on Candidate Evaluations." *Political Psychology* 38(2): 279–95.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Use Change Scores or Control for Pre-Treatment Outcomes? Depends on the True Data Generating Process." *DeclareDesign*. https://declaredesign.org/blog/2019-01-15-change-scores.html.

Bowers, Jake. 2011. "Making Effects Manifest in Randomized Experiments." In *Cambridge Handbook of Experimental Political Science*, Cambridge University Press, 459–80.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental*

Designs for Research. Boston: Houghton Mifflin Company.

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

Chong, Dennis, and James N. Druckman. 2007. "A Theory of Framing and Opinion Formation in Competitive Elite Environments." *Journal of Communication* 57(1): 99–118.

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): 318–28.

Clifford, Scott. 2019. "How Emotional Frames Moralize and Polarize Political Attitudes." *Political Psychology* 40(1): 75–91.

Clifford, Scott, and Dane G. Wendell. 2016. "How Disgust Influences Health Purity Attitudes." *Political Behavior* 38(1): 155–78.

Cohen, Geoffrey L. 2003. "Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs." *Journal of Personality and Social Psychology* 85(5): 808–22.

Downing, James W., Charles M. Judd, and Markus Brauer. 1992. "Effects of Repeated Expressions on Attitude Extremity." *Journal of Personality and Social Psychology* 63(1): 17–29.

Druckman, James N., and Toby Bolsen. 2011. "Framing, Motivated Reasoning, and Opinions About Emergent Technologies." *Journal of Communication* 61(4): 659–88.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100(4): 627–35.

———. 2011. “Experimentation in Political Science.” In *Cambridge Handbook of Experimental Political Science,* New York: Cambridge University Press.

Druckman, James N., Erik Peterson, and Rune Slothuus. 2013. “How Elite Partisan Polarization Affects Public Opinion Formation.” *American Political Science Review* 107(01): 57–79.

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.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton.

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

Goh, Jin X., Judith A. Hall, and Robert Rosenthal. 2016. “Mini Meta-Analysis of Your Own Studies: Some Arguments on Why and a Primer on How.” *Social and Personality Psychology Compass* 10(10): 535–49.

Harbord, Roger M., and Julian P. T. Higgins. 2008. “Meta-Regression in Stata.” *The Stata Journal* 8(4): 493–519.

Huddy, Leonie, Lilliana Mason, and Lene Aarøe. 2015. “Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity.” *American Political Science Review* 109(01): 1–17.

Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters*. Chicago: University of Chicago Press.

Jerit, Jennifer, Jason Barabas, and Scott Clifford. 2013. “Comparing Contemporaneous Laboratory and Field Experiments on Media Effects.” *Public Opinion Quarterly* 77(1): 256–82.

Kam, Cindy D., and Marc J. Trussler. 2016. “At the Nexus of Observational and Experimental Research: Theory, Specification, and Analysis of Experiments with Heterogeneous Treatment Effects.” *Political Behavior*: 1–27.

Klar, Samara, and Yanna Krupnikov. 2016. “Independent Politics: How American Disdain for Parties Leads to Political Inaction.” : 212.

Klar, Samara, Thomas J. Leeper, and Joshua Robison. 2019. “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(01): 59–80.

Krupnikov, Yanna, and Spencer Piston. 2015. “Racial Prejudice, Partisanship, and White Turnout in Elections with Black Candidates.” *Political Behavior* 37(2): 397–418.

Montgomery, Jacob M., Brendan Nyhan, and Michelle Torres. 2018. “How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It.” *American Journal of Political Science* 62(3): 760–75.

Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge University Press.

Mummolo, Jonathan, and Erik Peterson. 2019. “Demand Effects in Survey Experiments: An Empirical Assessment.” *American Political Science Review* 113(2): 517–29.

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

Mutz, Diana C., and Byron Reeves. 2005. “The New Videomalaise: Effects of Televised Incivility on Political Trust.” *American Political Science Review* 99(1): 1–15.

Nelson, Thomas E. et al. 1997. “Media Framing of a Civil Liberties Conflict and Its Effect on

Tolerance." *American Political Science Review* 91(03): 567–83.

Nicholson, Stephen P. 2011. "Dominating Cues and the Limits of Elite Influence." *Journal of Politics* 73(4): 1165–77.

Nyhan, Brendan, and Jason Reifler. 2010. "When Corrections Fail: The Persistence of Political Misperceptions." *Political Behavior* 32(2): 303–30.

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

Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *American Psychologist* 17(11): 776–83.

Piston, Spencer. 2018. *Class Attitudes in America: Sympathy for the Poor, Resentment of the Rich, and Political Implications*. New York: Cambridge University Press.

Roese, Neal J., and James M. Olson. 1994. "Attitude Importance as a Function of Repeated Attitude Expression." *Journal of Experimental Social Psychology* 30(1): 39–51.

Schueler, Beth E., and Martin R. West. 2016. "Sticker Shock." *Public Opinion Quarterly* 80(1): 90–113.

Schuman, Howard H., and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. San Diego, CA: Academic Press.

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

Tourangeau, Roger, and Kenneth A. Rasinski. 1988. "Cognitive Processes Underlying Context Effects in Attitude Measurement." *Psychological Bulletin* 103(3): 299–314.

Valentino, Nicholas A., Vincent L. Hutchings, and Ismail K. White. 2002. “Cues That Matter: How Political Ads Prime Racial Attitudes During Campaigns.” *American Political Science Review* 96(01): 75–90.

Valentino, Nicholas A., Fabian G. Neuner, and L. Matthew Vandenbroek. 2018. “The Changing Norms of Racial Political Rhetoric and the End of Racial Priming.” *The Journal of Politics* 80(3): 757–71.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.

Zizzo, Daniel John. 2010. “Experimenter Demand Effects in Economic Experiments.” *Experimental Economics* 13(1): 75–98.