No Evidence That Measuring Moderators Alters Treatment Effects

Geoffrey Sheagley
Associate Professor
University of Georgia
104 Baldwin St.
Athens, GA 30602
geoff.sheagley@uga.edu
0000-0002-1743-6201

Scott Clifford
Associate Professor
University of Houston
3551 Cullen Blvd. Rm 447
Houston, TX 77204
(713) 743-3890
sclifford@uh.edu
0000-0002-9401-7481

Abstract. Social scientists are frequently interested in who is most responsive to a treatment. By necessity, such moderation experiments often rely on observed moderators, such as partisan identity. These designs have led to an ongoing debate about where to measure moderators – immediately prior to the treatment, after the treatment, or in a prior wave of a panel survey. Measuring a moderator prior to the treatment is the most efficient and avoids post-treatment bias, but raises concerns about priming. We contribute to this debate by systematically studying whether measuring moderators prior to an experiment affects the results. Across six different experiments, each involving a commonly used moderator, we find little evidence of priming effects, even when a moderator is placed immediately before the experiment. Our findings thus help resolve the debate, suggesting that researchers should measure moderators pretreatment. We conclude with advice on designing well-powered moderation experiments.

The data and materials required to verify the computation reproducibility of the results, procedures, and analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/0YHJRG

Word Count: 9,290

As experiments have become more popular in public opinion research, scholars are increasingly interested in "second generation" studies (Kam and Trussler 2016) that reveal "the boundaries of a given theory—the kind of people for whom it is true" (Mutz 2011, 98). For example, a researcher may want to investigate who is most likely to follow a partisan cue (Bakker, Lelkes, and Malka 2020) or whether informational corrections can "backfire" among some respondents (Nyhan and Reifler 2010). Studies such as these frequently rely on a measured moderator, such as partisan identity, to examine for whom a treatment is more (or less) effective. Indeed, a systematic review of the literature between 1990 and 2014 finds that 63% of all articles using an experiment included an observed moderator (Kam and Trussler 2016).

While experimental studies involving observed moderators are quite common, there is an unresolved debate about how to design these studies. Specifically, researchers disagree about *where* to measure the moderator. Many researchers have chosen to measure a moderator after the treatment due to concerns that measuring a moderator prior to the treatment will prime the measured concept and influence the experimental results (for discussion, see Klar, Leeper, and Robison 2020; Mendelberg 2008a; Valenzuela and Reny 2021). However, recent research makes clear that measuring the moderator posttreatment can introduce bias of unknown size and direction into treatment effect estimates if the moderator is affected by the treatment (Acharya, Blackwell, and Sen 2016; Montgomery, Nyhan, and Torres 2018). As a solution, some have turned to the use of panel studies that allow moderators to be measured in a wave prior to the experiment (e.g., Banks 2014; Hainmueller and Hopkins 2015; Horiuchi, Markovich, and Yamamoto 2021; Klar and McCoy 2021; Newman and Malhotra 2019). But these designs are more costly, sometimes infeasible, and introduce concerns about attrition.

Despite the stakes of this debate, there is relatively little evidence as to whether measuring a moderator prior to an experiment actually affects the results. If this threat has been overstated, then researchers can adopt a simpler, more efficient design that avoids the possibility of posttreatment bias. We clarify this debate by first reviewing the existing theory and evidence behind such "measurement priming" effects, then by providing systematic evidence on whether measuring moderators pretreatment alters estimated treatment effects. Using a series of panel studies, we replicate well-established survey experimental paradigms in American political behavior that use moderators, while randomizing the presence of the moderator. Our studies involve the most common moderators, including partisan identity, policy attitudes, political values, racial resentment, social dominance orientation, and political knowledge. These studies are fielded on a variety of samples, including those drawn from an undergraduate subject pool, Mechanical Turk, and the Cooperative Election Study (CES).

Across six different experiments, the estimated treatment effects are remarkably similar regardless of whether the moderating variable is measured in the same wave as an experiment. We also provide evidence that even the distance between the moderator and the experiment *within* a survey wave does not alter the substantive results. Finally, we directly test the hypothesis that pretreatment alters conditional treatment effects through priming with a thought-listing task and find no evidence that measuring a moderator increases the salience of that concept. Consistent with a recent line of work on experimental design (Brutger et al. 2022; Clifford, Sheagley, and Piston 2021; Mummolo and Peterson 2019), our findings suggest that treatment effects in survey experiments are highly robust to alternative design choices. Thus, researchers generally do not need to invest in costly panel designs or risk posttreatment bias. We

conclude with practical guidance for researchers on how to design and conduct well-powered moderation experiments.

Theory and Practice in Moderation Experiments

We focus our attention on treatment effect heterogeneity, rather than causal moderation effects, both for simplicity and because most applied research on the topic is not designed to credibly estimate a causal moderation effect (for discussion, see Bansak 2021). In formal terms, our focus is on the Average Treatment Moderation Effect (ATME). Thus, for a moderator, *S*, our estimand is:

$$E[Y_i(1) - Y_i(0)|S_i = 1] - E[Y_i(1) - Y_i(0)|S_i = 0]$$

In other words, our interest is in whether the magnitude of the treatment effect varies across levels of the moderating variable, rather than whether the moderating variable *causes* variation in the treatment effect. For example, the effect of a party cue treatment may differ between those who have a strong attachment to a political party compared to those with a weak attachment. Our focal question is whether that moderating effect, the ATME, differs depending on when the moderator is measured.

Researchers designing experiments that involve a measured moderator face three possible choices for when to measure the moderator: posttreatment, pretreatment, or in a prior wave of a panel study. Each design choice poses potential risks and benefits. We first briefly discuss the debate over posttreatment bias and the use of panel designs, then turn to discussing priming effects in greater detail.

The Threat of Posttreatment Bias in Moderation Experiments

Posttreatment bias results from conditioning on a variable that has been affected by the treatment, such as a moderator that is measured posttreatment (Aronow, Baron, and Pinson 2019; Coppock 2019; Montgomery, Nyhan, and Torres 2018). As succinctly stated by Coppock (2019), "conditioning on post-treatment outcomes 'de-randomizes' an experiment in the sense that the resulting treatment and control groups no longer have potential outcomes that are in expectation equivalent." This has led some to conclude that "conditioning on post-treatment variables should be avoided in all cases" (Coppock 2019, 3).

However, many researchers continue to measure moderators posttreatment (for a review of practices, see Montgomery, Nyhan, and Torres 2018) under the assumption that stable traits, identities, and demographics are unlikely to be affected by experimental treatments (Albertson and Jessee 2022; Klar, Leeper, and Robison 2020). While this assumption often seems safe, standard hypothesis tests cannot rule out small average treatment effects or heterogeneous effects that average out to zero. Even small deviations from the assumption of a sharp null effect can lead to potentially large bias (Aronow, Baron, and Pinson 2019). Thus, the absence of posttreatment bias remains an assumption even when there is no statistically significant treatment effect on the moderator.

Panel Studies as a Solution

As a solution to the tension between the risks of posttreatment bias and priming effects, some scholars have advocated for the use of panel studies. By measuring the moderator in a prior wave, researchers can largely rule out the possibility of a priming effect. For example, Hainmueller and Hopkins (2015, 535) measured moderators in a survey three weeks prior to their experiment, arguing that the design "enables us to measure potential moderating variables

without priming respondents or introducing differential measurement bias." Many others have adopted a panel design due to these same concerns (e.g., Banks 2014; Hainmueller and Hopkins 2015; Horiuchi, Markovich, and Yamamoto 2021; Klar and McCoy 2021; Newman and Malhotra 2019; Valentino, Neuner, and Vandenbroek 2018).

While panel studies are a common solution to this problem, they are not without limitations. Most obviously, panel studies are costly. Klar et al. (2020) estimate the cost of a panel study at approximately three times the cost of a single wave. The three-fold cost of a panel study may be cost-prohibitive for a researcher, or may force a reduction in sample size, and thus statistical power. Panel attrition also raises concerns about sample representativeness. Finally, as Klar et al. (2020) point out, panel studies are sometimes infeasible, such as when conducting exit polls or studying political rallies. Thus, while panel studies can resolve the tension between pretreatment and posttreatment measurement of a moderator, these designs come with several downsides and limitations.

Priming in Theory and in Practice

The influence of prior questions on experimental results is widely believed to occur through priming (e.g., Klar, Leeper, and Robison 2020) and these concerns are spurred on by the broader literature on survey design and attitude formation. The conventional model of survey response holds that respondents provide answers based on a sampling of considerations, "…including an oversample of ideas made salient by the questionnaire…and use them to choose among the options offered" (Zaller and Feldman 1992, 580). This model clearly highlights priming, or "changes in the standards that people use to make political evaluations" (Iyengar and

Kinder 1987, 63), as a central component of survey design and interpretation.[1] This perspective informs much of the subsequent scholarship raising concerns about priming in survey experiments. For example, in discussing issues of spillover effects between experiments, Transue, Lee, and Aldrich (2009, 19) appeal to "the fragility of survey response to question wording." Related research suggests that "any survey item might, by chance, induce unequal effects in the control and treatment groups of a later item" (Gaines, Kuklinski, and Quirk 2007, 18). In an extended debate over whether measuring racial attitudes prior to an experiment biases the results, Mendelberg (2008b, 116) draws on a "well-established literature [that] documents the influence of question order in surveys." In short, concerns about pretreatment measures stem, at least in part, from research demonstrating the fickleness of attitudes reported in surveys.

At the same time, many priming studies in social psychology have failed to replicate, such as the priming of flags (Klein et al. 2014), money (Caruso, Shapira, and Landy 2017; Klein et al. 2014), and mortality salience (Klein et al. 2019). Similar debates have played out in political science, such as over whether irrelevant events (e.g., football games) affect voting behavior (e.g., Fowler and Gollust 2015; Healy, Malhotra, and Mo 2010). Thus, while there is a broad literature supporting subtle priming effects, it should be treated with some skepticism.

In the following section, we briefly review some of the evidence for priming effects across a variety of moderators commonly used in experimental political science. The review is

[1] An alternative model of attitude formation focuses on spreading activation as a mechanism to explain priming (Lodge and Taber 2005b). In this model, information is stored in nodes that are linked to other associated nodes. When one node is activated, this activation spreads to other associated nodes.

not intended to be exhaustive, but to review some of the most relevant evidence that priming effects could result from simply measuring a concept. These are also the areas we drew from when designing the experiments used in the studies reported in this paper.

Issue Attitudes

A large literature suggests that priming an issue attitude increases the weight given to it when evaluating a candidate, however, many common designs cannot distinguish between informational (or learning) effects, priming effects, and projection effects (cf., Lenz 2012). Even experimental designs face this problem. For example, many designs involve complex treatments that include information, such as candidate positions (Druckman and Holmes 2004; Hart and Middleton 2014). Even designs evaluating question order effects face inferential problems, such as when the treatment variable is also used as a moderator, which is measured post-treatment among respondents in the control group (e.g., Cassino and Erisen 2010). Thus, while there are strong theoretical grounds for issue priming, the evidence is less clear than one might expect.

Group Identities and Attitudes

Many scholars have examined the impact of priming identities, particularly partisan identity. For example, there is evidence, though mixed, that merely asking about partisan identity affects economic evaluations and financial decisions (Bailey 2022; Heath et al. 2015; Morris, Carranza, and Fox 2008) and policy attitudes (Klar 2013). There is even suggestive evidence that a combination of six questions about partisanship, ideology, issue attitudes, and presidential approval affected self-reported personality (Bakker, Lelkes, and Malka 2021). Thus, the evidence suggests that merely measuring partisan identity can affect downstream responses.

There has also been an extensive debate about the nature of racial priming. Some scholars have argued that merely measuring racial resentment “acts as a cue that primes predispositions”

(Mendelberg 2008a), while others have cast doubt on this concern (Huber and Lapinski 2008). More recently, some have argued that the "baseline salience of racial attitudes in American politics may have increased dramatically in recent years," potentially eliminating the possibility of further priming racial attitudes (Valentino, Neuner, and Vandenbroek 2018, 761). Thus, scholars hold conflicting views, but there has been little attempt to empirically resolve this debate.[2]

Values, Traits, and Skills

Researchers have also used a variety of psychological traits, values, skills and other dispositions as moderators of treatment effects, including the Big Five personality traits and epistemic and existential needs (e.g., Bakker, Lelkes, and Malka 2020; Federico and Schneider 2007; Johnston, Lavine, and Federico 2017). Many of these are conceptualized as basic needs or orientations that are unlikely to be primed. However, there is evidence that value orientations can be primed through writing tasks (Waytz, Dungan, and Young 2013) or even subtle word unscrambling tasks (Maio and Rees Maio 2009). Even information processing style can be influenced by simple task instructions (Druckman and Leeper 2012; Tormala and Petty 2001) or other survey characteristics (Hauser and Schwarz 2015). Thus, it is not implausible that merely measuring these constructs might influence downstream outcomes, though we are not aware of any direct evidence on this question.

[2] One exception is Valentino et al. (2018), who vary whether they measure symbolic racism one week in advance of an experiment, immediately before the experiment, or post-treatment. They find no evidence that this design choice influences their results.

One exception is the case of political knowledge, which is often conceptualized as a domain-specific cognitive skill or resource. There is consistent evidence that measuring difficult political knowledge questions causes respondents to report lower levels of interest in politics (Bishop 1987; Bishop, Oldendick, and Tuchfarber 1982; Schwarz and Schuman 1997). However, the knowledge measures used in these studies are unusually difficult and it's unclear whether this effect extends beyond anything but closely related constructs.

Implications for Experimental Design

The evidence for priming effects from merely measuring a concept is mixed and varies across topics, but is clearly sufficient to give researchers pause. We now turn to the question of how a measurement prime would likely affect results in the context of a moderation experiment. We begin with the assumption that experimental designs fall along a spectrum ranging from an informational design to a priming design. In an informational design, the treatment consists of providing a piece of information (e.g., partisan or racial cue, policy stance, policy design) that enables respondents to connect their dispositions (beliefs, attitudes, values, identities) to the outcome variable. In contrast, priming designs assume that respondents have already formed an association between their disposition and the outcome variable (e.g., Lodge and Taber 2005a). The treatment is instead intended to make the association more accessible (while holding information constant), causing the disposition to exert a stronger influence on the outcome variable.

Political scientists, however, often use complex treatments that may work through some combination of priming and information (e.g., exposure to debates or news). For example, framing likely works by changing beliefs about the topic, changing beliefs about the importance

of relevant considerations, and increasing the accessibility of relevant considerations (e.g., Nelson et al. 1997; Slothuus 2008). Even seemingly pure informational or priming designs likely work through both mechanisms. For a treatment to work purely through priming, no new information should be conveyed through the treatment, though this is often not the case in practice. For example, one study primes American identity with a lengthy news article (Levendusky 2018), which surely conveys some information in addition to making that identity more accessible. For a treatment to work purely through information, it must not affect the salience of the target concepts, which is unlikely to be the case. For example, a party cue might affect an issue attitude at least in part because it makes partisan identity more salient (i.e., it primes the concept). Thus, most experiments vary along a dimension ranging from mostly informational (i.e., creating a new connection between a consideration and an outcome) to mostly priming (i.e., increasing the accessibility of a pre-existing connection between a consideration and an outcome).

The nature of the experimental design has important implications for whether the measurement of a moderator affects the results. The potential for a measurement prime to alter a treatment effect is highest when considering a pure priming design. By measuring the moderating variable prior to the experiment, it may increase the accessibility of that concept. As a result, "the prior question itself acts as a cue that primes predispositions" (Mendelberg 2008b, 116), which "renders the entire sample one big treatment group, washing out any effect" (Mendelberg 2008a, 137). For example, consider a design that primes partisanship and this prime is expected to interact with a respondent's own partisan identity to affect their attitude toward some outcome. By measuring partisan identity immediately prior to the experiment, the researcher increases the accessibility of partisanship. To illustrate the plausibility of this

problem, recall that some experimental designs use questions about partisan identity as a treatment to prime partisanship. Under the assumptions that 1) the effect of the measurement prime on accessibility is as large as the effect of the treatment on accessibility, and 2) the treatment does not have an additional effect on accessibility above and beyond the effect of the measurement prime, measuring a moderator immediately prior to an experiment will attenuate or eliminate the effect.

However, the pretreatment measurement of a moderator only threatens to undermine the experiment if the moderating variable is closely related to the concept being primed (e.g., racial resentment and a race prime). In some cases, the effects of priming are expected to be moderated by political knowledge under the assumption that these respondents are more or less responsive to the prime (e.g., Druckman and Holmes 2004). In this case, there is little reason to expect that the measurement of the moderator will influence the treatment effect.

Political scientists, however, more commonly use designs with an informational component. We start with a scenario in which a treatment exerts an effect only through information. For example, consider a candidate evaluation study in which the treatment involves information about the candidate's sex scandal, the effect of which is expected to be moderated by traditional values. In the absence of the informational treatment, respondents are unable to make the connection between the primed concept (traditional values) and the outcome (candidate

evaluation). Thus, even if the measurement of the moderator primes that concept, it should exert no effect on the outcome in the control condition.[3]

The implications differ for the treatment condition, however. If we again assume that the treatment only works through information, then priming the concept of traditional values could increase the likelihood that respondents rely on these values to evaluate the candidate, increasing the size of the ATME. However, it is likely that informational treatments also prime related concepts. Continuing with the example above, it seems likely that treating respondents with information about a candidate's sex scandal will also make the concept of traditional values more accessible. If measuring the moderator has no *additional* effect on accessibility, a measurement prime will also have no effect on the outcome in the treatment condition. However, if measuring the moderator does have an additional effect on accessibility, then it may *increase* the size of the ATME. Thus, depending on assumptions about the nature of accessibility, measurement priming may have no effect, or it may strengthen the ATME (contrary to concerns commonly expressed in the literature).

In summary, there is good reason to believe that a measurement prime may affect the results of a pure priming study, though these designs are less common in political science. For more common designs that have an informational component, the expectations are less clear and depend on assumptions about additive or interactive effects of priming. We focus our attention below on experimental designs that include an informational component.

[3] One exception is the case of pretreatment. If some respondents are already aware of the information connecting the moderator and the outcome, then priming the moderator may strengthen this relationship in the control condition, weakening the ATME.

Experimental Studies

To test whether measuring moderators alters treatment effects, we fielded seven panel surveys that included six different experimental designs. To identify studies to replicate, we selected the most common observed moderators from Kam and Trussler's (2016) systematic review of the experimental literature: partisan identity, political attitudes and values, racial attitudes, and political awareness. Additionally, we included a study involving social dominance orientation to extend our findings to a politically relevant psychological predisposition. Our selected moderators are not only the most common, but also include several that are frequently mentioned in the context of concerns about measurement priming (e.g., partisan identity, racial attitudes).

We then selected specific experimental designs that we expected would produce a strong moderation effect and could be easily replicated in a survey experiment without special samples or measurement strategies (e.g., the Implicit Association Test). We excluded pure priming studies, as these are relatively uncommon in political science, and the implications for bias are more straightforward. Thus, our studies all share some informational component, but they vary across several dimensions that are relevant to the likelihood of measurement priming effects. Some are purely informational (e.g., support for a fictional trade agreement), while others likely have an important priming component (e.g., issue attitudes and candidate evaluation). Some involve hypothetical policies, while others involve actual politicians about whom participants may have pre-existing opinions. Finally, our moderators range from single-item measures (e.g.,

issue attitudes) to an eight-item scale (e.g., SDO).[4] Thus, while our tests surely do not generalize to all political science experiments, a point we revisit in the conclusion, they do represent a wide range of common designs.

Each experiment was embedded in a two-wave panel survey and follows roughly the same design. In the first wave, the moderator was measured for all respondents. The second wave contained the experiment. However, prior to the experiment, the moderator was measured again for a random half of respondents, which we refer to as a *measurement prime*. Because tests of moderation create unique challenges for statistical power, we took several steps to increase the precision of our estimates. In addition to using relatively large sample sizes, we replicated some of our studies across multiple samples, which we pool together. When possible, we also make use of pretreatment covariates, including measures of the dependent variable embedded in the prior wave (Clifford, Sheagley, Piston 2021).

These studies were fielded on a diverse array of samples, which are summarized in Table 1 (see Appendix pgs. 1-3 for details). The last two columns of Table 1 list the experiments that were included in each sample, along with the relevant moderator. In Surveys 2 and 3, the experiments were placed consecutively with unrelated survey content between the modules. In Surveys 5 through 7, we randomized *which* of the two moderators was measured in the second wave, which entails an assumption that measuring a moderator does not affect an unrelated experiment. Although there is some possibility of spillover between experiments (e.g.,

[4] This latter test is particularly important given the demonstrated importance of using longer scales to measure moderating variables (Bakker and Lelkes 2018).

Westwood and Peterson 2020), we find largely similar results across separate samples and combinations of studies.

Table 1. Overview of Surveys

Survey	**Sample Source**	**Dates**	**Wave 1 Sample Size**	**Wave 2 Sample Size**	**Included Studies**	**Moderator**
1	Forthright	Summer 2019	1,500	998	Candidate Position-Taking	Issue Stance
2	MTurk	Summer 2020	1,200	1,001	Candidate Position-Taking	Issue Stance
					Partisan Cues	Partisan ID
3	Undergraduate	Fall 2020	995	428	Race-Targeted Policy	Racial Resentment
					Value Framing	Humanitarianism
4	CES	Fall 2020	1,406	962	Race-Targeted Policy	Racial Resentment
5	MTurk	Summer 2021	1,303	1,050	Partisan Cues	Partisan ID
					Race-Targeted Policy	Racial Resentment
6	Forthright	Spring 2022	1,584	994	Motivated Reasoning	Political Knowledge
					Candidate Position-Taking	Issue Stance
7	MTurk	Summer 2022	2,009	1,639	Motivated Reasoning	Political Knowledge
					Trade Attitudes	Social Dominance Orientation

Note – This table provides a description survey sample size and which experiment(s) it contained.

Results

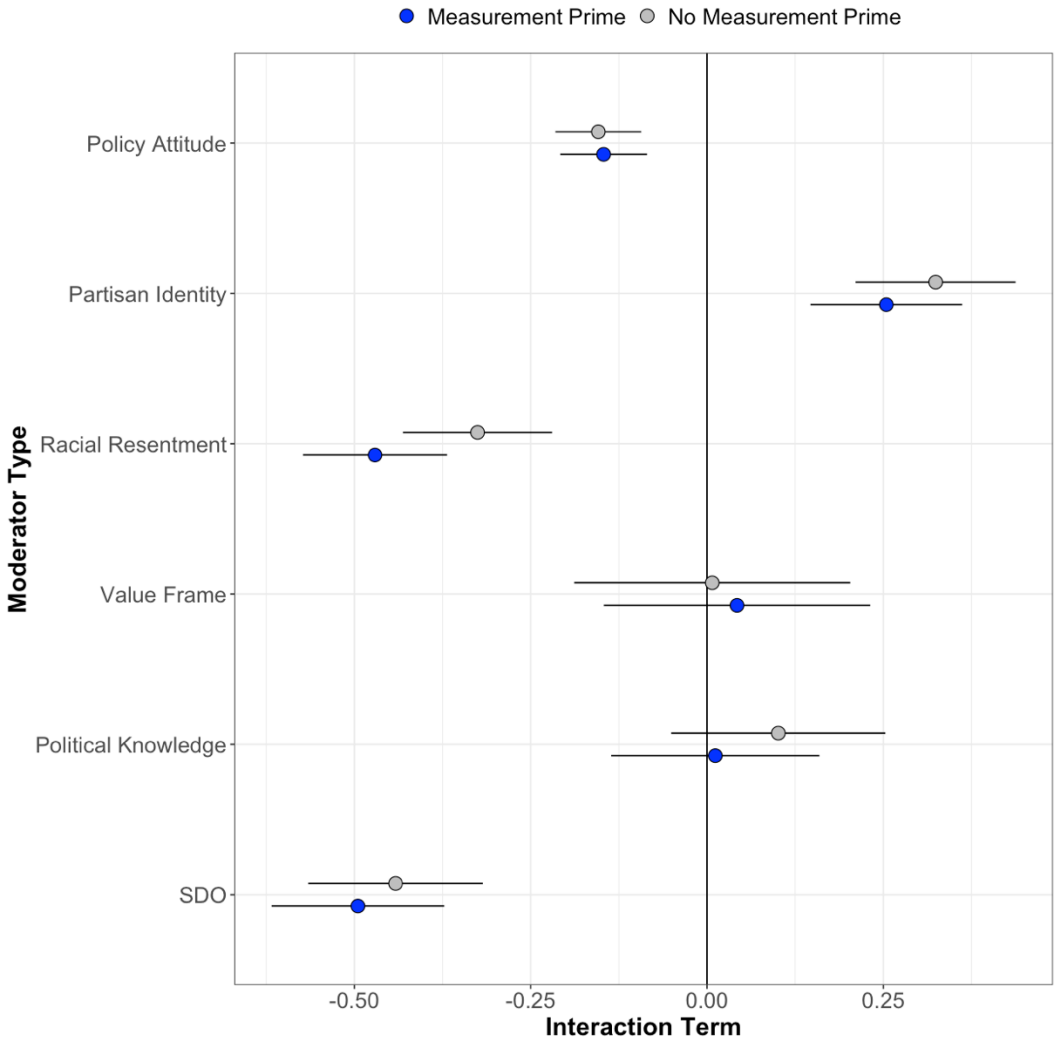
Our primary analysis focuses on the average treatment moderation effect (ATME) and whether this interaction differs depending on whether the moderator was also measured in wave 2 (a measurement prime). In other words, our concern is in the ATME within each design and the *difference* in ATMEs across conditions, as assessed with a three-way interaction term. For each study below, we report the ATME within each design condition, which represents the conclusion a researcher would arrive at with that design. Then, we report the difference between the ATMEs across the two design conditions, which represents the effect of the design on the substantive conclusion.

In each design condition, the ATME is estimated by regressing the outcome on the binary indicator for the treatment, the relevant moderator, and the interaction between the treatment and the moderator (all rescaled to range 0-1). We use only the wave 1 measure of the moderator to hold all else constant with the exception of any possible priming effects,[5] while estimating separate models for those that received the wave 2 measurement prime and those that did not. The difference in ATMEs is estimated by pooling both design conditions and altering the model described above by including a three-way interaction between the design condition, the treatment condition, and the moderator, as well as all constitutive terms.

Figure 1 plots the coefficient for the interaction term separately for the measurement prime and no prime designs. We discuss each experiment in turn, then offer more detail on the difference in ATMEs below.

[5] Pages 17-21 of the appendix include comparisons of the moderating effects between W1 and W2 measures.

Figure 1. ATME by Experimental Design and Presence of Measurement Prime



Note – This figure plots the ATME for participants exposed to a measurement prime (blue circles) and those who were not (grey circles). Bars represent 95% confidence intervals.

Candidate Position-Taking Experiment

The first experiment is a partial replication of a study on how candidates' issue stances affect citizens' perceptions of their character (Clifford 2014). All respondents were given a brief biography of Steve Bullock, including that at the time of the study he was the current Governor of Montana and candidate for the Democratic nomination for president (treatment = 0). In the

treatment condition, respondents were also informed that Bullock supports the death penalty (treatment = 1). Respondents were then asked whether Bullock is a "strong leader" and whether he "commands respect," each on a five-point scale. The two items are averaged together to form the dependent variable. The moderator is the respondent's own position on the death penalty, measured on a seven-point scale. To increase precision, we control for wave 1 partisanship. The sample size for these analyses is 3,011.

There is a negative, statistically significant interaction between the treatment and a respondent's views on the death penalty. Substantively, the treatment caused more positive trait assessments among those favoring the death penalty and more negative assessments among those opposing it. Crucially, the ATME was substantively identical for respondents who were exposed to the measurement prime (b = -.15, $p < 0.001$) and for those who were not (b = -.15, $p < 0.001$). Moreover, as indicated by the three-way interaction in the pooled model, the two effects are not significantly different (p = .89).

Partisan Cue Experiment

The *Partisan Cue Experiment* is a replication of a study on party cues (Bakker, Lelkes, and Malka 2020). All respondents were asked to read a paragraph on farm subsidies in which respondents were either told that Democrats support and Republicans oppose the policy (treatment = 0), or vice-versa (treatment = 1). Respondents were then asked to rate their support for the policy on a seven-point scale, with higher values indicating greater policy support. The moderator is a four-item scale of partisan social identity (Bankert, Huddy, and Rosema 2017)

ranging from strong Democratic identity to strong Republican identity.[6] These analyses include 2,072 respondents.

There is a positive and statistically significant relationship between the treatment and moderator. The positive interaction term indicates that partisans move in opposing directions depending on the party cue presented. We also find similar ATMEs regardless of whether respondents were exposed to a measurement prime ($b = .25$, $p < 0.001$) or not ($b = .32$, $p < 0.001$; *difference* = .07, $p = .38$).

Race-Targeted Policy Experiment

The third experiment is a partial replication of a study on how racial resentment moderates support for a race-targeted policy (Feldman and Huddy 2005). Respondents were asked about the extent to which they support providing college scholarships to students who score in the top fifteen percent of their class, regardless of the overall ranking of their school. In the treatment condition, the policy applied only to "Black students." Although the control condition is not explicitly racial, we expect that the issue of college scholarships is easily racialized given that the policy involves redistribution. Indeed, we provide some evidence later in the paper that some respondents explicitly viewed this policy through the lens of race and affirmative action. The dependent variable is support for the scholarship program, which is coded so that higher values indicate greater support. The moderator is a 4-item scale of racial resentment, with higher values indicating greater racial resentment. These analyses include 1,481 respondents and is a pre-post design.

[6] Pure independents were randomly assigned to either the Republican or the Democratic version of the scale.

The relationship between the treatment and the racial resentment moderator is negative and statistically significant. Respondents high in racial resentment were less supportive of the policy in the treatment condition, while those low in racial resentment showed the opposite pattern. Substantively, the results are the same across designs, though the interaction coefficient was somewhat *larger* in the presence of a measurement prime ($b = -.47$, $p < 0.001$) than in the absence of one ($b = -.33$, $p < 0.001$), though this difference is not statistically significant ($p = 0.06$).

These results stand in contrast to common concerns that the measurement prime will *reduce* the ATME by increasing the accessibility of racial considerations in all conditions. Absent the treatment, increasing the accessibility of racial considerations has little impact as respondents may be unable to connect these considerations to the policy. In the treatment condition, however, the measurement prime may have strengthened the impact of racial attitudes, thus increasing the ATME. We provide more evidence on this interpretation in a section below.

Value Framing Experiment

The *Value Framing Experiment* is a conceptual replication of a study that examined how media frames could activate different political values (Shen and Edwards 2005). All respondents read a fictional newspaper story about welfare reform that either highlighted individualist values (control) or humanitarian values (treatment). After reading the article, respondents were asked about their support for poor people and children receiving public assistance, with higher values

indicating greater support. The moderator is a six-item scale of humanitarian values, where higher values indicate greater humanitarianism. These analyses include 428 respondents.[7]

The interaction between humanitarian values and the treatment is substantively small and not statistically significant, regardless of whether a measurement prime was present (b = .04, p = .66) or not (b = .009, p = .93). Thus, we observe no evidence that a respondent's level of support for humanitarian values conditions their reaction to the value frame, however, there is also no evidence that a measurement prime altered this finding (*difference* = .03, p = .86).

Motivated Reasoning Experiment

The fifth experiment is a replication of a common motivated reasoning paradigm in which both political knowledge (Guay and Johnston 2022) and numeracy (Kahan et al. 2017) have been found to produce larger bias. Respondents were given a vignette describing a study on how concealed carry policies affect gun crimes. The numbers presented in the study were randomized such that they suggested that banning concealed carry either increased or decreased the crime rate. Respondents were then asked to report whether the study suggested that a ban would increase or decrease crime, then report their confidence on a 4-point scale. For the dependent variable, we combine these items into a single measure that ranges from confidently wrong to confidently correct. Each wave of the survey included a similar, but different 5-item measure of factual political knowledge (see Appendix pg. 8 for details). The total sample size for this experiment is 2,617.

[7] We opted not to collect more data for this study after initial results suggested that the key interaction would not replicate.

Respondents whose wave 1 gun control attitude aligned with the treatment were coded 1 while respondents who received a counter-attitudinal treatment were coded 0. Surprisingly, we find no evidence of a moderation effect whether a measurement prime was included (b = .01, p = .88) or not (b = .10, p = .19), though the sign of the interaction is in the expected direction. However, we find no evidence that the ATME differs between the measurement prime and no measurement prime conditions (*difference* = .09, p = .41).

Trade Attitudes Experiment

Our final experiment is a partial replication of Mutz and Kim (2017). Respondents were asked to read a brief paragraph describing a hypothetical trade policy between the US and another unnamed country. The policy was described as either leading to an equal gain in jobs in both countries (win-win condition) or a gain for the US and an equal loss in jobs for the trading partner (win-loss condition).[8] Policy support is measured on a 7-point scale where higher values indicate greater support. The moderator, social dominance orientation, is measured with the 8-item SDO_7 scale (Ho et al. 2015).

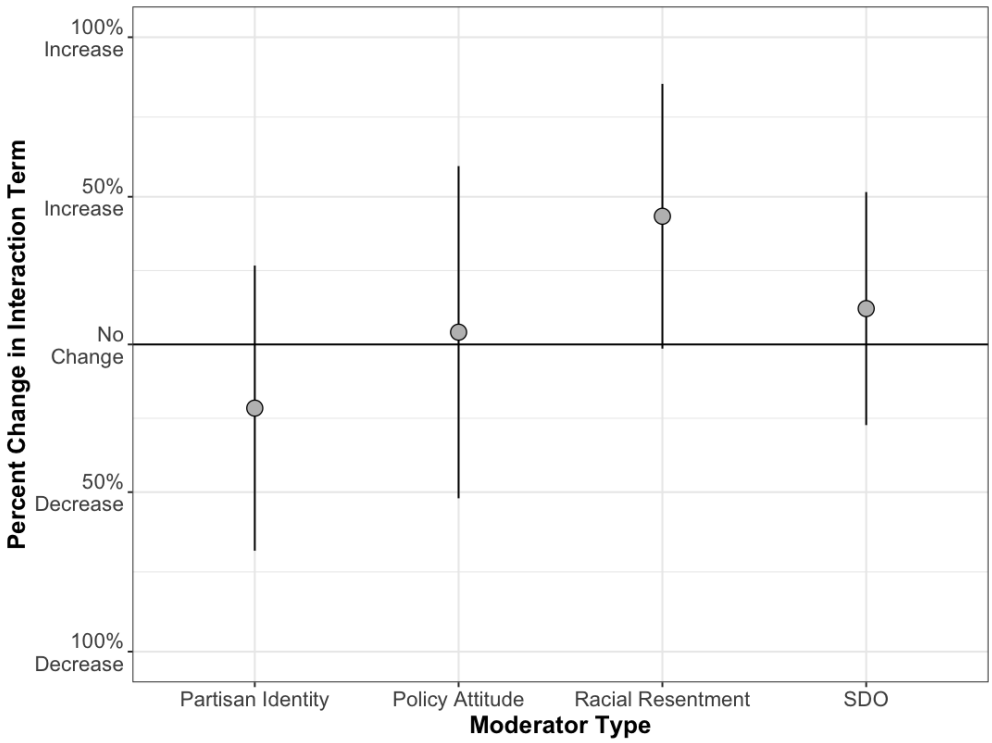
There is a negative and statistically significant relationship between the treatment and moderator. Substantively, the win-win condition increases support among those low in SDO, but has little effect on those high in SDO. The ATME is highly similar in the measurement prime condition (b = -.50, p < .001) and the no prime condition (b = -.44, p < .001), and the difference in ATMEs is not statistically significant (*difference* = .05, p = .55).

[8] On the following page, respondents were then asked about the other experimental condition for a within-subjects design. However, we observed a significant order effect, so we only report the first round of the experiment.

Overview of Results

Overall, across six different experiments, we find substantively equivalent results whether or not the moderator is measured prior to the experiment. The findings were similar for the ATMEs, as well as the marginal effects and the ATEs (see appendix pgs. 14-15). To provide a more substantive interpretation of these three-way interactions and the precision of these estimates, we rescale the three-way interaction relative to the size of the relevant two-way interaction. Specifically, we divide the three-way interaction term by the two-way interaction term, as estimated in the no measurement prime condition. This rescaling allows us to interpret the three-way interaction term as a proportional change in the size of the moderation effect. The results are shown in Figure 2, though we omit the two studies that failed to replicate a main moderation effect. The y-axis ranges from -100%, which means the measurement prime completely removed the moderation effect, to 100%, which means the measurement prime doubled the size of the moderation effect.

Figure 2. Effect of Measurement Prime on ATME.



Note – This figure plots the percent change in the interaction in the measurement prime compared to the no measurement prime conditions. Bars represent 95% confidence intervals.

Starting on the left, the three-way interaction term for the *Partisan Cues Experiment* implies that the estimated two-way moderation effect is roughly 20% smaller when the moderator is measured in the same wave as the experiment. However, the confidence intervals on this estimate range from a 69% decrease to a 27% increase. Turning to the *Candidate Position-Taking Experiment*, the estimate implies that measuring the moderator in the same wave increases the estimated moderation effect by 4%, with confidence intervals ranging from an increase of 60% to a decrease of 52%. The estimate for the *Race-Targeted Policy Experiment* suggests that measuring the moderator in the same wave increases the moderation effect by about 40%, although the confidence intervals extend from -1% to 80%, meaning that we cannot reject the null hypothesis of no effect of the measurement prime. Finally, the estimate for the *Trade Attitudes Experiment* suggests a 12% increase in the ATME, ranging from a decrease of 27% to

an increase of 52%. Overall, our evidence suggests no change in the ATME, though the evidence is suggestive in one case and there is meaningful uncertainty in all cases.[9]

Does it Matter Where the Moderator is Measured Within a Wave?

Throughout all of our studies, we included irrelevant questions between the measurement prime and the experiment, following advice 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). This raises the question of whether our results are robust to placement of the moderator *within* a survey wave, an assumption we test here.

Survey 4 (CES) included a replication of the race-targeted policy experiment. We use the same variable coding and estimation strategy as above, thus we include controls for W1 measures of party identification and the outcome. In all conditions, racial resentment was measured at the start of the survey in W2. We then randomized whether the experiment was administered immediately after this measurement (the "close" condition) or if it was administered at the end of the survey (the "distant" condition). In the distant condition, there were approximately 28 questions between the moderator and the experiment, which covered topics such as redistricting, vote counting, presidential power, issue positions, and evaluations of incumbent senators and members of Congress, and measures of affective polarization. With one exception, discussed below, the intervening questions did not explicitly measure racial attitudes

[9] Power analyses suggest that all four of these studies were well-powered to detect a difference in ATMEs when the ATME is eliminated by the measurement prime. Further, at 80% power, our studies can detect a reduction in the ATME ranging from 50% to 79%.

and thus we assume that these questions alone did not prime these considerations. These analyses include 962 respondents.

Among the policy attitudes questions, a random half of the sample was also asked about their views on whether Black Lives Matter protests should be allowed.[10] While this item complicates our test, because it was independently randomized, it also provides an additional way to assess concerns about priming. In the close condition, respondents only received the BLM question after the experiment. In the distant condition, half of respondents received the question prior to the experiment, while half did not.

We estimated the basic interactive model among three subsets of our data. The key interaction is similar among those in the close condition (and prior to BLM question; $b = -.30$, $p < .0001$), in the distant condition with the BLM question ($b = -.36$, $p < .0001$), or in the distant condition without the BLM question ($b = -.35$, $p = .0001$). Pooling both distant conditions to maximize power, we also do not find a significant difference between the far and close ($p = .50$). This suggests that our results are robust to design choices regarding the distance between the moderator and the experiment, as well as the inclusion of questions that might prime relevant attitudes.

Testing the Mechanism

[10] The other half of the sample was asked about Covid-19 protests. Roughly 19 questions were asked after the Black Lives Matter item. These questions focused on measures of partisan identity and candidate support in the 2020 election.

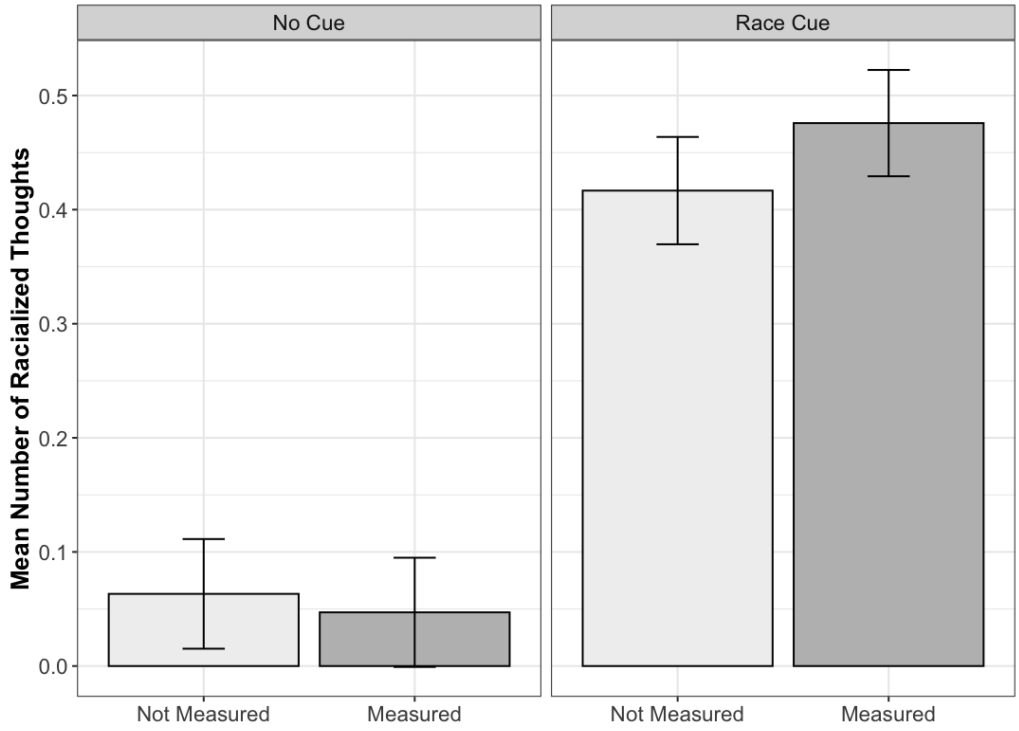
Our evidence suggests it is unlikely that the measurement of a moderator will affect the results of an experiment. However, we cannot rule out the possibility that measuring a moderator changes the effect size to a moderate degree, particularly for the racial resentment experiment. In this section, we directly test the mechanism by analyzing whether the measurement prime increased the likelihood that racial considerations were brought to bear on the outcome (see Appendix p. 22 for evidence regarding political knowledge).

In Study 5, after completing the race-targeted policy experiment, all respondents were asked what considerations came to mind. Recall that half of respondents were asked about scholarships for top students while the other half were asked about scholarships for the top Black students. Additionally, only half answered the racial resentment questions in the same wave as the experiment. Thus, we can examine whether the mere measurement of racial resentment affects the accessibility of race-related thoughts in the open-ended responses.

To analyze the open-ended responses, we coded all responses that explicitly mentioned concepts related to race, ethnicity, or discrimination as race-related, and all other responses as not race-related. To ensure the reliability of our coding, the two authors independently coded a random 99 responses and reached 100% agreement. One of the authors then coded the remaining responses.

To test for priming, we use OLS to model the likelihood of racialized considerations being mentioned as a function of experimental treatment condition and the presence of a measurement prime. We also control for wave 1 racial resentment. Figure 3 displays the modeled mean level of racialized thoughts for each condition with 95% confidence intervals.

Figure 3. Mean Number of Racialized Thoughts by Race Cue Experimental Condition and Moderator Measurement



Note – This figure plots the mean number of racialized thoughts mentioned by participants in each experimental condition. Lines represent 95% confidence intervals.

As expected, the race cue treatment condition has a large effect, increasing the probability of mentioning race by 40 percentage points ($p < .001$). In contrast, measuring racial resentment increased the probability of a racial consideration by roughly two percentage points, but this effect is not statistically significant ($p = .46$). However, any racial priming effect may occur only within one condition. Contrary to some concerns, we find no evidence that a measurement prime may activate the concept within the control condition. For these respondents, the probability of a thought having a racial consideration was .06 in the control condition and .04 in the measurement prime condition ($b = -.02$, $p = .59$). Within the treatment condition, the measurement prime increased racial considerations from 42% to 47%, though this difference is

not statistically significant ($b = .05$, $p = .11$). Nonetheless, this pattern is consistent with the slightly larger ATME observed in the measurement prime condition. Thus, the measurement of racial resentment does not seem to have a meaningful effect on the likelihood of raising explicit racial considerations, though it is possible that it modestly amplified the effect of racial attitudes in the treatment condition. Our analysis does have limitations, of course, as it is possible that our open-ended measure could not capture coded racial language or implicit effects.

Conclusion

Survey experiments involving the estimation of conditional average treatment effects are increasingly common in political science (Kam and Trussler 2016). With this expanding focus, there has also been a growing tension in the methodological literature as to when a researcher should measure a moderator. Central to this debate are concerns that measuring a moderator prior to an experiment could alter treatment effects. Though these concerns are frequently expressed in the literature and clearly influence design choices, there has been little direct empirical evidence on the topic.

Across six different experiments including the most commonly used moderators in political science, we found no evidence that measuring a moderator prior to an experiment influences the results. In all cases, we reached the same substantive results, in terms of sign and significance, regardless of whether we measured the moderator in a prior wave or shortly before the experiment. And in none of these cases could we reject the null hypothesis of no effect of the placement of the moderator. Of course, we cannot rule out moderate changes in the *size* of the effect, but we did not find any consistent pattern in the direction of any possible priming effect.

In each of our studies we sought to maximize the distance between the moderator and the experiment when they were measured in the same wave. This, we believe, is a common precaution intended to reduce the likelihood of any priming effects, though it may not be an available option in all cases (e.g., in very short surveys). However, in our one experimental test of whether the placement *within* a survey matters, we find no evidence that it does. Thus, while we still encourage researchers to separate the moderator and experiment when possible, this design choice also seems unlikely to matter. All told, the experimental results examined here seem remarkably robust to these important design choices.

Of course, we should be cautious in generalizing our findings to the wide variety of studies run by political scientists. Our findings are necessarily limited to these six moderators and the corresponding designs. The clearest concerns about measurement priming effects have been raised in the context of identity experiments, but we found no evidence of priming effects in our experiments that involved identity. However, like most experiments in political science, all of our experiments involved some informational component. We do expect that measuring a moderator prior to a pure priming experiment (i.e., with no informational component) would alter the results (e.g., Klar 2013; Transue 2007). Thus, while our results likely generalize to many common studies in political science, there are clear bounds on that generalizability and we encourage researchers to further explore this question empirically.

For researchers designing moderation experiments, we recommend first considering the nature of the design. If the treatment is expected to work to a large degree through priming, then measurement priming may reduce the ATME. However, designs of this sort are typically aimed at testing theory rather than identifying a particular effect size (Druckman et al. 2006). Thus, a reduced ATME is primarily a problem of reduced statistical power, which could be offset by

improved experimental designs, such as a pre-post design. More generally, we emphasize the importance of considering statistical power in the design stage and using reliable measures of the moderating variable (Bakker and Lelkes 2018). Overall, however, our studies suggest that – for a variety of common types of survey experiments – measurement priming poses little threat to the results. Thus, our results help clarify a debate over the design of experiments involving observed moderators, allowing researchers to avoid costly panel designs as well as the potential bias from post-treatment measurement of moderators.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3): 512–29.

Albertson, Bethany, and Stephen A. Jessee. 2022. "Moderator Placement in Survey Experiments: Racial Resentment and the 'Welfare' versus 'Assistance to the Poor' Question Wording Experiment." *Journal of Experimental Political Science*.

Aronow, Peter M., Jonathon Baron, and Lauren Pinson. 2019. "A Note on Dropping Experimental Subjects Who Fail a Manipulation Check." *Political Analysis* 27(4): 572–89.

Bailey, Jack. 2022. "Political Surveys Bias Self-Reported Economic Perceptions." *Public Opinion Quarterly* 85(4): 987–1008.

Bakker, Bert N., and Yphtach Lelkes. 2018. "Selling Ourselves Short? How Abbreviated Measures of Personality Change the Way We Think about Personality and Politics." *The Journal of Politics* 80(4): 1311–25.

Bakker, Bert N., Yphtach Lelkes, and Ariel Malka. 2020. "Understanding Partisan Cue Receptivity: Tests of Predictions from the Bounded Rationality and Expressive Utility Perspectives." *Journal of Politics* 82(3): 1061–77.

———. 2021. "Reconsidering the Link Between Self-Reported Personality Traits and Political Preferences." *American Political Science Review*.

Bankert, Alexa, Leonie Huddy, and Martin Rosema. 2017. "Measuring Partisanship as a Social Identity in Multi-Party Systems." *Political Behavior* 39: 103–32.

Banks, Antoine J. 2014. "The Public's Anger: White Racial Attitudes and Opinions Toward Health Care Reform." *Political Behavior* 36: 493–514.

Bansak, Kirk. 2021. “Estimating Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184(1): 65–86.

Bishop, George F. 1987. “Context Effects on Self-Perceptions of Interest in Government and Public Affairs.” In *Social Information Processing and Survey Methodology*, eds. Hans-J Hippler, Norbert Schwarz, and Seymour Sudman. New York: Springer-Verlag, 179–99.

Bishop, George F., Robert W. Oldendick, and Alfred J. Tuchfarber. 1982. “Political Information Processing: Question Order and Context Effects.” *Political Behavior* 4(2): 177–200.

Brutger, Ryan et al. 2022. “Abstraction and Detail in Experimental Design.” *American Journal of Political Science*.

Caruso, Eugene M., Oren Shapira, and Justin F. Landy. 2017. “Show Me the Money: A Systematic Exploration of Manipulations, Moderators, and Mechanisms of Priming Effects.” *Psychological Science* 28(8): 1148–59.

Cassino, Dan, and Cengiz Erisen. 2010. “Priming Bush and Iraq in 2008: A Survey Experiment.” *American Politics Research* 38(2): 372–94.

Clifford, Scott. 2014. “Linking Issue Stances and Trait Inferences: A Theory of Moral Exemplification.” *The Journal of Politics* 76(03): 698–710.

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.

Coppock, Alexander. 2019. “Avoiding Post-Treatment Bias in Audit Experiments.” *Journal of Experimental Political Science* 6(1): 1–4.

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.

Druckman, James N., and Justin W Holmes. 2004. "Does Presidential Rhetoric Matter? Priming and Presidential Approval." *Presidential Studies Quarterly* 34(4): 755–78.

Druckman, James N., and Thomas J. Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4): 875–96.

Federico, Christopher M., and Monica C. Schneider. 2007. "Political Expertise and the Use of Ideology: Moderating Effects of Evaluative Motivation." *Public Opinion Quarterly* 71(2): 221–52.

Fowler, Erika Franklin, and Sarah E. Gollust. 2015. "The Content and Effect of Politicized Health Controversies." *The ANNALS of the American Academy of Political and Social Science* 658(1): 155–71.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. "The Logic of the Survey Experiment Reexamined." *Political Analysis* 15(01): 1–20.

Guay, Brian, and Christopher Johnston. 2022. "Ideological Asymmetries and the Determinants of Politically Motivated Reasoning." *American Journal of Political Science*.

Hainmueller, Jens, and Daniel J. Hopkins. 2015. "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants." *American Journal of Political Science* 59(3): 529–48.

Hart, Austin, and Joel A. Middleton. 2014. "Priming under Fire: Reverse Causality and the Classic Media Priming Hypothesis." *Journal of Politics* 76(2): 581–92.

Hauser, D. J., and N. Schwarz. 2015. "It's a Trap! Instructional Manipulation Checks Prompt

Systematic Thinking on 'Tricky' Tasks." *SAGE Open* 5(2): 1–6.

Healy, Andrew J., Neil Malhotra, and Cecilia Hyunjung Mo. 2010. "Irrelevant Events Affect Voters' Evaluations of Government Performance." *Proceedings of the National Academy of Sciences* 107(29): 12804–9.

Heath, Oliver, Patten Smith, Nicholas Gilby, and Finn Hoolahan. 2015. "Partisan Priming and Subjective Evaluations of the Economy: Evidence from a Survey Experiment." *Journal of Elections, Public Opinion and Parties* 25(4): 530–42.

Ho, Arnold K. et al. 2015. "The Nature of Social Dominance Orientation: Theorizing and Measuring Preferences for Intergroup Inequality Using the New SDO$_7$ Scale." *Journal of Personality and Social Psychology* 109(6): 1003–28.

Horiuchi, Yusaku, Zachary D. Markovich, and Teppei Yamamoto. 2021. "Does Conjoint Analysis Mitigate Social Desirability Bias?" *Political Analysis*.

Huber, Gregory A., and John S. Lapinski. 2008. "Testing the Implicit-Explicit Model of Racialized Political Communication." *Perspectives on Politics* 6(1): 125–34.

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

Johnston, Christopher D., Howard Lavine, and Christopher M. Federico. 2017. *Open versus Closed: Personality, Identity, and the Politics of Redistribution*. Cambridge University Press.

Kahan, Dan M., Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. 2017. "Motivated Numeracy and Enlightened Self-Government." *Behavioural Public Policy* 1(1): 54–86.

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. 2013. "The Influence of Competing Identity Primes on Political Preferences." *Journal of Politics* 75(4): 1108–24.

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

Klar, Samara, and Alexandra McCoy. 2021. "Partisan-Motivated Evaluations of Sexual Misconduct and the Mitigating Role of the #MeToo Movement." *American Journal of Political Science* 65(4): 777–89.

Klein, Richard A. et al. 2014. "Investigating Variation in Replicability." *Social Psychology* 45(3): 142–52.

———. 2019. "Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement."

Lenz, Gabriel S. 2012. *Follow the Leader?: How Voters Respond to Politicians' Policies and Performance*. Chicago: University of Chicago Press.

Levendusky, Matthew S. 2018. "Americans, Not Partisans: Can Priming American National Identity Reduce Affective Polarization?" *Journal of Politics* 80(1): 59–70.

Lodge, Milton, and Charles S. Taber. 2005a. "The Automaticity of Affect for Political Leaders, Groups, and Issues: An Experimental Test of the Hot Cognition Hypothesis." *Political Psychology* 26(3): 455–82.

———. 2005b. "The Automaticity of Affect for Political Leaders, Groups, and Issues: An Experimental Test of the Hot Cognition Hypothesis." *Political Psychology* 26(3): 455–82.

Maio, Greg, and Kerry Rees Maio. 2009. "Changing, Priming, and Acting on Values: Effects via

Motivational Relations in a Circular Model." *Journal of Personality and Social Psychology* 97(4): 699–715.

Mendelberg, Tali. 2008a. "Racial Priming: Issues in Research Design and Interpretation." *Perspectives on Politics* 6(1): 135–40.

———. 2008b. "Racial Priming Revived." *Perspectives on Politics* 6(1): 109–23.

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.

Morris, Michael W., Erica Carranza, and Craig R. Fox. 2008. "Mistaken Identity: Activating Conservative Political Identities Induces 'Conservative' Financial Decisions: Research Article." *Psychological Science* 19(11): 1154–60.

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., and Eunji Kim. 2017. "The Impact of In-Group Favoritism on Trade Preferences." *International Organization* 71(4): 827–50.

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.

Newman, Benjamin J, and Neil Malhotra. 2019. "Economic Reasoning with a Racial Hue: Is the Immigration Consensus Purely Race Neutral?" *Journal of Politics2*.

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

Schwarz, Norbert, and Howard Schuman. 1997. "Political Knowledge, Attribution, and Inferred Interest in Politics: The Operation of Buffer Items." *International Journal of Public*

Opinion Research 9(2).

Slothuus, Rune. 2008. “More Than Weighting Cognitive Importance: A Dual-Process Model of Issue Framing Effects.” *Political Psychology* 29(1): 1–28.

Tormala, Zakary L., and Richard E. Petty. 2001. “On-Line Versus Memory-Based Processing: The Role of ‘Need to Evaluate’ in Person Perception:” *Personality and Social Psychology Bulletin* 27(12): 1599–1612.

Transue, John E. 2007. “Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force.” *American Journal of Political Science* 51(1): 78–91.

Transue, John, Daniel J. Lee, and John H. Aldrich. 2009. “Treatment Spillover Effects across Survey Experiments.” *Political Analysis* 17(2): 143–61.

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.

Valenzuela, Ali A., and Tyler Reny. 2021. “The Evolution of Experiments on Racial Priming.” In *Advances in Experimental Political Science*, eds. James Druckman and Donald P. Green. Cambridge University Press.

Waytz, Adam, James Dungan, and Liane Young. 2013. “The Whistleblower’s Dilemma and the Fairness–Loyalty Tradeoff.” *Journal of Experimental Social Psychology* 49(6): 1027–33.

Westwood, Sean J., and Erik Peterson. 2020. “The Inseparability of Race and Partisanship in the United States.” *Political Behavior* (0123456789).