No Need for a Panel Study: Measuring Moderators Doesn't Alter Treatment Effects

Geoffrey Sheagley
Assistant Professor
University of Georgia

Scott Clifford
Associate Professor
University of Houston

Abstract. As survey experiments have become more common in political science, so too have efforts to identify who is most responsive to a treatment. These moderation experiments frequently rely on observed, rather than manipulated moderators, such as partisan identity or racial attitudes. 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. Each design choice has its downsides and detractors. Measuring a moderator after the treatment opens the possibility of posttreatment bias. Measuring it prior to the experiment may create priming effects. And panel studies are costly and sometimes infeasible. We contribute to this debate by systematically studying whether measuring moderators prior to an experiment affects the results. Across four different experiments involving four of the most commonly used moderators, we find no evidence of priming effects. In an additional experiment, we find no evidence that the distance between a moderator and an experiment within a survey affects the results. Our findings thus help resolve the debate, suggesting that the pretreatment measurement of a moderator often poses little threat to the inferences drawn from an experiment.

As experiments have become more popular in public opinion research, scholars are increasingly interested in *who* is most responsive to a treatment. For example, a researcher may want to investigate who is most likely to follow a partisan cue, such as strongly identified partisans (Bakker, Lelkes, and Malka 2020), providing more insight into the nature of elite influence over public opinion. Studies such as these frequently rely on a measured moderator, such as partisan identity. 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. Recent research makes clear that measuring the moderator post-treatment can introduce bias of unknown size and direction into treatment effect estimates (Montgomery, Nyhan, and Torres 2018). Yet, others have argued that measuring the moderator pretreatment may alter the treatment effect by increasing the salience of the moderating variable (Klar, Leeper, and Robison 2020; Valenzuela and Reny 2021). 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 (Hainmueller and Hopkins 2015). But these designs are more costly, sometimes infeasible, and introduce concerns about attrition.

Taken together, the literature does not provide clear a recommendation to researchers, but instead three different design choices, each with its own downsides. In this paper, we aim to clarify this debate by providing systematic evidence on whether measuring moderators pretreatment does in fact alter estimated treatment effects. Using a series of panel studies, we replicate well-established survey experimental paradigms that use moderators, while

randomizing the placement of the moderator. Our studies include the most commonly used moderators, including policy attitudes, political values, racial resentment, and partisan identity. Additionally, we rely on a variety of samples, including those drawn from undergraduate subject pools, Mechanical Turk, and the Cooperative Election Study (CES).

Overall, across four different experiments, the estimated treatment effects are remarkably similar regardless of the placement of the moderating variable. We also find some evidence that even the distance between the moderator and the experiment *within* a survey wave does not alter the substantive results of the study. Finally, we directly test the priming hypothesis using a thought-listing task and find no evidence that measuring a moderator increases the salience of the concept. These findings suggest that in many cases researchers need not invest in costly panel designs and potentially introduce problems with attrition. We conclude with practical guidance for researchers on how to design and conduct moderation experiments.

Theory and Practice in Moderation Experiments

The core strength of an experiment is that it allows researchers to clearly identify whether and to what degree an independent variable generates a change in a dependent variable. In more recent years, survey experiments have become increasingly popular vehicles for experimental research, especially for research on political behavior (Mutz 2011). The rise of survey experiments, and their corresponding demographically diverse participants and large sample sizes, allows researchers to study what Kam and Trussler (2016, 790) call second generation questions, which are studies that, "go beyond the average treatment effect" of an experiment. Examining heterogeneity in treatment effects is valuable for a variety of reasons. First, understanding treatment effect heterogeneity can be useful for theory testing. For example,

testing for moderation in treatment effects allows researchers to examine, "the boundaries of a given theory—the kind of people for whom it is true" (Mutz 2001, 98). Second, knowing who is most responsive to treatments can help improve the design of policy interventions. For example, while pro-vaccination campaigns tend to be persuasive, on average, there is evidence that these persuasive messages can backfire among some people, making them less likely to intend to get the vaccine (Nyhan and Reifler 2014).

Of course, relying on moderators that are measured, rather than manipulated, raises a number of inferential challenges (Green and Kern 2012; Kam and Trussler 2016). Yet, many of the most common moderators cannot be easily manipulated within a survey. For example, Kam and Trussler (2016) find that the most common observed moderators are party identification, racial attitudes, political knowledge, race, ideology, and gender. Clearly, none of these variables can be easily manipulated in a survey, if at all, forcing many researchers to rely on observed moderators in their experimental designs.

As discussed above, researchers face three choices when designing such moderation experiments: posttreatment measurement, pretreatment measurement, or measurement in a prior wave of a panel study. Each design choice poses potential risks and benefits, which we review in more detail below.

The Threat of Posttreatment Bias in Moderation Experiments

Posttreatment bias refers to the potential bias induced by conditioning on posttreatment outcomes, whether in the form of posttreatment moderators, mediators, manipulation checks, or responses that are only observed among a subset of respondents (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 recent research on the topic comes to a clear conclusion: "conditioning on post-treatment variables should be avoided in all cases" (Coppock 2019, 3).

While political methodologists documented this problem long ago, a recent systematic literature review revealed that 47% of experimental studies in top journals conditioned on post-treatment variables in some way, most commonly by including a post-treatment variable as a control or a moderator (Montgomery, Nyhan, and Torres 2018). But this is not simply due to a lack of awareness. Some scholars have argued that it is acceptable to measure such moderators posttreatment when guided by past evidence and theory (Klar, Leeper, and Robison 2020; Valenzuela and Reny 2021). Supporting this view, most common observed moderators are stable traits, dispositions, or demographics that are not easy to manipulate within a survey. They are also unlikely to be affected by experimental treatments or other features of a survey. For example, partisan identity is notoriously difficult to manipulate (for discussion, see Gerber, Huber, and Washington 2010).

Researchers sometimes provide empirical evidence to support their use of posttreatment moderators. For example, Walter and Redlawsk (2019) argue that because they observed very little shared variance between the treatment and posttreatment moderator ($R^2 = .0025$), this implies no bias in their estimates. Yet, as pointed out by Montgomery, Nyhan, and Torres (2018), the absence of a statistically significant effect of the treatment on the moderator is insufficient to rule out the possibility of posttreatment bias. To do so, a researcher would have to provide evidence for the sharp null of no effect on any unit. But standard hypothesis tests cannot rule out small average treatment effects or heterogeneous effects that average out to zero. Even small deviations from this strong assumption can lead to potentially large bias (Aronow, Baron,

and Pinson 2019). Thus, the absence of posttreatment bias remains an assumption even when the treatment does not significantly affect the moderator.

Ultimately, posttreatment bias remains a threat to any experiment measuring a moderator posttreatment. Some scholars have been willing to accept the possibility of bias, suggesting it is unlikely in certain cases in which the moderator is extremely stable. But this claim remains an untestable assumption that some scholars are unwilling to grant. And given that the central advantage of a randomized experiment is that it allows researchers to identify unbiased treatment effects, the decision to measure a moderator posttreatment may undermine confidence in the results of the study.

Concerns About Priming Effects

The primary reason for measuring moderators posttreatment is the possibility that the measurement of the moderator might itself affect the experimental results. This argument has been made most clearly in the context of research on identity. Klar et al. (2020) cite a number of examples that involve priming a particular identity (e.g., partisanship, race, national identity) by randomizing whether a question about that identity is measured prior to the outcome. Drawing on this literature, they argue that measuring moderators prior to the treatment "may change the definition of the causal parameter being estimated from the effect of the treatment when identity is non-salient to the effect when it is salient, which may not be the effect the experiment is interested in detecting."[1] Others go further in the context of racial attitudes, arguing that

[1] Others have suggested that measuring the moderator prior to an experiment might cause differential measurement error (Hainmueller and Hopkins 2015).

measuring racial attitudes before an experiment carries a risk of "washing out any differences between treatment and baseline groups" (Valenzuela and Reny 2021).

This same line of argument has also been extended to moderators that do not directly involve identity. For example, Walter and Redlawsk (2019) argue that measuring their moderators (moral values and partisanship) "could not be done before treatment, given the very real risk that doing so would prime participants in their responses" to the experiment (Walter and Redlawsk 2019, 1082).

Overall, there are clear concerns that measuring a moderator prior to a treatment may alter or undermine the experimental results. This has led some authors to argue that "experimentalists should not always measure identities pretreatment either. Instead, researchers must base this decision on case-specific theory regarding the relationship between the treatment and measure of identity, and an explicit trade-off of the risks of posttreatment bias and priming effects" (Klar, Leeper, and Robison 2020). Similarly, Valenzuela and Reny (2021) suggest that "post-treatment measurement of racial attitudes is a safe approach."

Panel Studies as an Imperfect 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 wave prior to the wave containing the experiment, 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." Valentino, Neuner, and Vandenbroek (2018) went so far as to randomize the placement of the moderating variable, including in a prior wave of a panel survey.

While panel studies seem to resolve this dilemma, and have been adopted in a number of studies, they are not without limitations. First, and 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. Part of this cost is due to collecting larger samples in the initial survey wave to address the inevitable attrition between interviews. 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. 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 a number of downsides and limitations.

Revisiting the Evidence Base for Priming Effects

The challenges of measuring moderators in experimental studies is clear. Yet, in spite of the commonly cited concerns of priming effects, there is little direct evidence that priming systematically alters experimental effects. For example, Montgomery, Nyhan, and Torres (2018) admit the possibility of priming and call for further research on the topic. In the absence of direct evidence, this may be a case in which researchers' fears are overblown, and experimental results may be unaffected by the larger context of the survey (Clifford, Sheagley, and Piston 2020; Mummolo and Peterson 2019). We review the available evidence in more detail below.

We are aware of only one direct test that experimentally tests whether the placement of a moderator affects the results of a moderation experiment (Valentino, Neuner, and Vandenbroek 2018). In this study, the authors randomized whether the moderator, racial resentment, was

measured in a prior wave, just before the experiment in the same wave, or just after the experiment in the same wave. The authors find that "the timing of racial attitude measures had no impact" (767). Of course, this is just a single test, and an unusual one in that the authors expected (and found) no interaction between the moderator and the treatment.

While the direct evidence is quite limited, Klar et al. (2020) draw on a number of studies to provide indirect evidence that the pretreatment measurement of common moderators may influence experimental results. They cite several studies that use question order manipulations as a means to prime an identity (e.g., gender) and show that these primes affect political attitudes. While this is clear evidence that measuring identity can affect political attitudes, the implications for moderation experiments are less clear. Moderation experiments often involve a treatment that provides the connection between a measured identity or disposition and a political attitude. For example, consider common party cue studies. The threat is that measuring partisan identity prior to the experiment might heighten the salience of the identity, increasing its impact on the dependent variable in the control condition. The heightened salience of partisanship may reduce any impact of the treatment, relative to the control. Yet, absent the partisan cue, respondents in the control condition presumably lack the information required to make a connection between their identity and the outcome variable.[2] For respondents in the treatment condition, the partisan cues make the identity salient regardless of the measurement of the moderator. Thus, it's unclear

[2] Of course, if many respondents are pretreated (e.g., already aware of party positions on the issue), it's possible that an identity prime might activate pre-existing knowledge (cf., Druckman and Leeper 2012).

whether the existence of a priming effect necessarily affects many designs involving moderated treatment effects.

Overall, there is little direct evidence as to whether measuring moderators pretreatment will affect the results of common moderation experiments. Absent that evidence, researchers may be unnecessarily risking posttreatment bias or investing in more costly panel studies. In several studies below, we seek to provide a broad body of evidence on the topic.

Experimental Studies

To test whether measuring moderators alters treatment effects, we fielded 4 surveys that included a total of 5 different experimental designs. To select studies to replicate, we started with Kam and Trussler's (2016) systematic review of the experimental literature. They found that the most common observed moderators were partisan identity, political/racial attitudes, and political awareness. Based on this information, we decide to replicate studies involving partisan identity, racial attitudes, issue attitudes, and political values as moderators. We chose not to include political awareness as moderator because, although it is commonly used, it seems the least likely to be sensitive to measurement timing. In this sense, we have selected the topics that are the most commonly studied and also most likely to be subject to bias.

As for the studies themselves, we selected experiments that we expected would replicate and would produce a strong moderation effect, making it easier to find differences between designs. Additionally, we limited our selection to experiments that could be easily replicated in a survey experiment without special samples or measurement strategies (e.g., the Implicit Associations Test). We discuss each study in more detail below, but they all share a common structure. Each experiment includes two experimental conditions and the treatment involves

varying some piece of information (e.g., a party cue, policy detail, or candidate's stance) that helps respondents link the moderator (e.g., partisan identity) to the dependent variable (e.g., candidate support). While this general structure is quite common in political science, it surely does not encompass all possible designs, a point we revisit in the conclusion.

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. 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. To increase precision, we also make use of pretreatment covariates, including measures of the dependent variable embedded in the prior wave, as available (Clifford, Sheagley, Piston 2020).

These studies were fielded using a diverse array of samples. Table 1 provides an overview of each sample, and our discussion below contains additional information about each. Additionally, Table 1 lists the experiments that were included in each sample, along with the relevant moderator. Study designs are discussed in more detail in the following section.

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	526	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

Survey 1 – Forthright Sample

Survey 1 was conducted from the Forthright panel, which is a national online sample maintained by Bovitz Inc. While not a probability sample, the sample was matched to United States Census benchmarks for race, education, age, gender, and region. The first wave was fielded in June 2019 and Wave 2 was fielded one month later. 1,500 respondents completed Wave 1 and 998 completed Wave 2. Survey 1 included Study 1, which is a candidate position taking experiment.

Survey 2 - MTurk

Survey 2 was conducted on a sample recruited from Amazon's Mechanical Turk using the Cloud Research platform. Respondents were required to reside in the United States, have completed at least 100 HITs and have an approval rate of at least 95%. Additionally, we used the Cloud Research quality filters to block suspicious IPs and duplicate IPs, which have been associated with poor data quality (Kennedy et al. 2020). The first wave was fielded between

August 31 and September 1, 2020 and the second wave one week later (Sept 8-11, 2020). Of the 1,200 completing the first wave, 1,001 also completed the second wave.

Survey 2 included two studies. The first was a replication of Study 1, the candidate position-taking experiment. Study 2 is a partisanship experiment. Because Study 1 included a partisan politician in the design, Study 2 was placed prior to Study 1 in the survey for all respondents.

Survey 3 – Undergraduate

Survey 3 was administered on a sample of undergraduate students from a large university in the South. Students were recruited to the study through large sections of a course on introduction to American politics. The first wave was fielded in mid-October 2020 and the re-interview took place roughly one month later in mid-November. 995 respondents completed Wave 1 and 526 respondents completed both survey waves.

Survey 3 included Study 3 – a race cue experiment- and Study 4, which focused on framing and political values. Because Study 4 contained information about a social welfare program, it was administered after Study 3. Additionally, a series of questions on political participation and personality characteristics were asked between Study 3 and Study 4.

Survey 4 – Cooperative Election Survey (CES)[3]

Survey 4 was conducted on a module administered through the 2020 Cooperative Election Study (CES; Schaffner, Ansolabehere, and Luks 2021). The CES is a large, national survey administered to a sample of over 60,000 respondents asked a common set of questions and modules containing questions asked by researchers. The data for this paper are drawn from

[3] The CES is the new name for the Cooperative Congressional Election Study (CCES).

the post-election survey module, which contained 962 responses.[4] Our focus is on estimating treatment effects and to maximize precision, we do not use the provided sample weights in our analyses (Miratrix et al. 2017).

Survey 4 included a partial replication of Study 3. Rather than relying on a panel survey to measure the moderator in a prior survey wave, we manipulated the placement of the race cue experiment in a single survey wave.

Sample 5 - MTurk

Survey 5 was conducted on a sample recruited from Amazon's Mechanical Turk using the same inclusion criteria as Survey 2. The first wave was fielded on August 3, 2021 and the second wave approximately one week later (Aug. 10-11). Of the 1,303 completing the first wave, 1,050 also completed the second wave.

Survey 5 included replications of the partisanship experiment (Study 2) and the race cue experiment (Study 3). Because the survey involved two studies, it required manipulating the measurement of two moderating variables in the second wave. If they were independently manipulated, the length of the survey would vary by up to eight questions. To avoid this problem, we simply manipulated *which* moderator was measured. This allows us to test both of

[4] The sample size is larger than found in a traditional CES module (which has n=1,000 on the pre-election wave and roughly 800 on the post-election wave) because it includes data from the participants recruited into the survey but who were later discarded when YouGov matches their sample to their sample frame. These "unmatched" participants are discarded from the final survey releases. Results are the same if we restrict our analysis to the "matched" YouGov sample.

our hypotheses, but entails the assumption that the measurement of partisan identity does not affect the race experiment and that the measurement of racial resentment does not affect the partisan cue experiment.

Results

Our primary analyses focus on testing whether the conditional average treatment effect (CATE) of an experiment varies by whether the moderator was measured in the same wave as the experiment or in a prior wave of the survey. In our analysis, we use only the wave 1 measure of the moderator to hold all else constant with the exception of any possible priming or demand effects. Alternatively, we could compare the effects of the wave 1 moderator to the effects of the wave 2 moderator. However, any differences could be due to priming effects caused by the measurement of the moderator in wave 2 or by differences in how the moderator was measured (e.g., temporal instability). Relying on only the wave 1 measure allows us to isolate the effects of how measurement of the moderator in the second wave affects responses to the dependent variable.

We begin by summarizing the relationship between the moderator and the treatment effect separately for respondents for whom the moderator was measured in the prior survey wave vs. the same survey wave that contained the experimental manipulation.[5] For brevity, we refer to these two conditions as the "prior-wave" and "same-wave" designs, though it should be noted

[5] Results are substantively the same when we generate these estimates from a single model with a three-way interaction between treatment, moderator, and when the moderator was measured. We report results from these models later in the paper.

that the moderator was measured in the prior wave for all respondents regardless of experimental condition. For each study, we regress the outcome (rescaled to run from 0-1) on the binary indicator for the treatment, the relevant moderator (rescaled to run from 0-1), and the interaction between the treatment and the moderator. In order to increase precision, we also include relevant pre-treatment covariates for each study (Bowers 2003; Clifford, Sheagley, and Piston 2020), although our results hold when these additional covariates are omitted from our models.

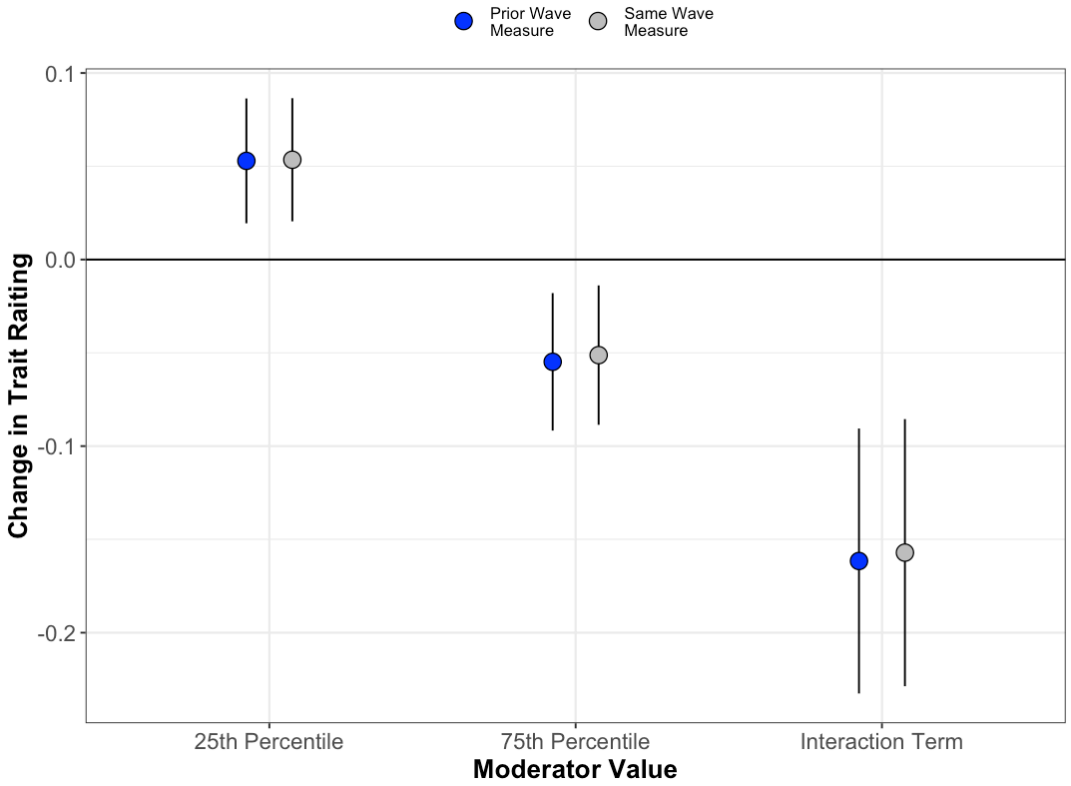
We present the results in two ways for each study. The first is by plotting the coefficient for the interaction between the moderator and the treatment separately for each design. Substantively, the interaction term represents the difference in treatment effects at the minimum and maximum values of the moderating variable. We complement these analyses by presenting the estimated treatment effect at the 25th and 75th percentile value of the moderator.

Study 1 – Issue Attitudes Moderate Effects of Candidate Stances

Study 1 is a conceptual 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 to assess 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 2,008.

The moderator is coded so that higher values correspond to greater *opposition* to the death penalty. The outcome corresponds to more *positive* trait assessments of the candidate. Thus, we would expect that the treatment should lead people with favorable views of the death penalty to have more positive assessments of the candidate, compared to when they are unaware of his position. Figure 1 plots the results from the experiment.

Figure 1. Conditional Relationship Between Policy Position, Treatment Effect, and When the Moderator Was Measured



There is a negative, statistically significant interaction between the treatment and a respondent's views on the death penalty. Substantively, the significant interaction between treatment and death penalty attitudes shows that respondents with relatively favorable views of

the death penalty had more positive trait assessments when exposed to treatment and those who were opposed to the death penalty had more negative trait assessments when exposed to treatment. For example, the effect of treatment is positive for respondents in the 25th percentile of the moderator (which corresponds to positive death penalty views) and negative for those in the 75th percentile (which corresponds to negative death penalty views).

Figure 1 also shows that the interaction between treatment and death penalty attitudes was substantively identical for respondents whose opinions about the death penalty were measured in the same (b = -.16, $p < 0.001$) or prior (b = -.16, $p < 0.001$) survey wave. In short, when the moderator was measured did not alter this conditional relationship.

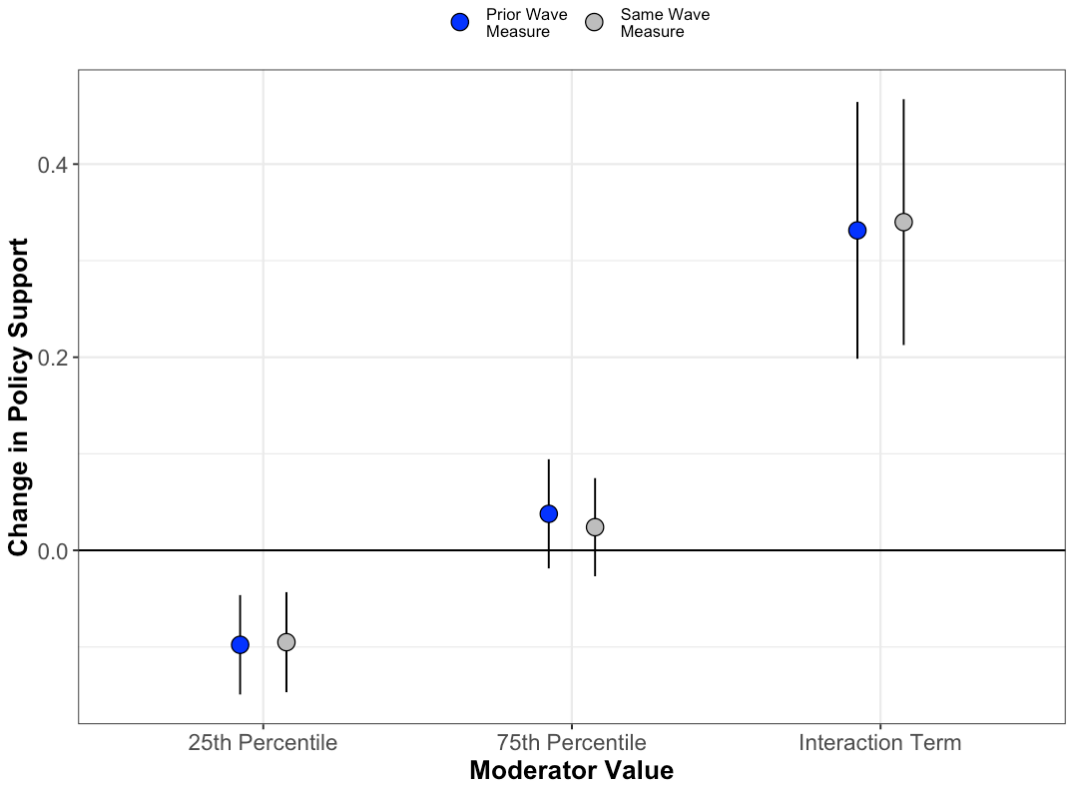
Study 2 – Partisan Identity Moderates Party Cue Effects

Study 2 is a replication of a study on party cues (Bakker, Lelkes, and Malka 2020). All respondents were asked to read a paragraph describing arguments for and against farm subsidies. Respondents were randomly assigned to a version in which Democrats supported and Republicans opposed 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, which is coded so that higher values map onto greater policy support. The moderator is a four-item scale of partisan social identity (Bankert, Huddy, and Rosema 2017), which we recode to run from 0 (Strong Democratic identity) to 1 (Strong Republican identity).[6] Substantively, we expect a positive

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

interaction between the treatment and the moderator. These analyses include 2,072 respondents. Results are displayed in Figure 2.

Figure 2. Conditional Relationship Between Partisan Identity, Treatment, and When the Moderator Was Measured



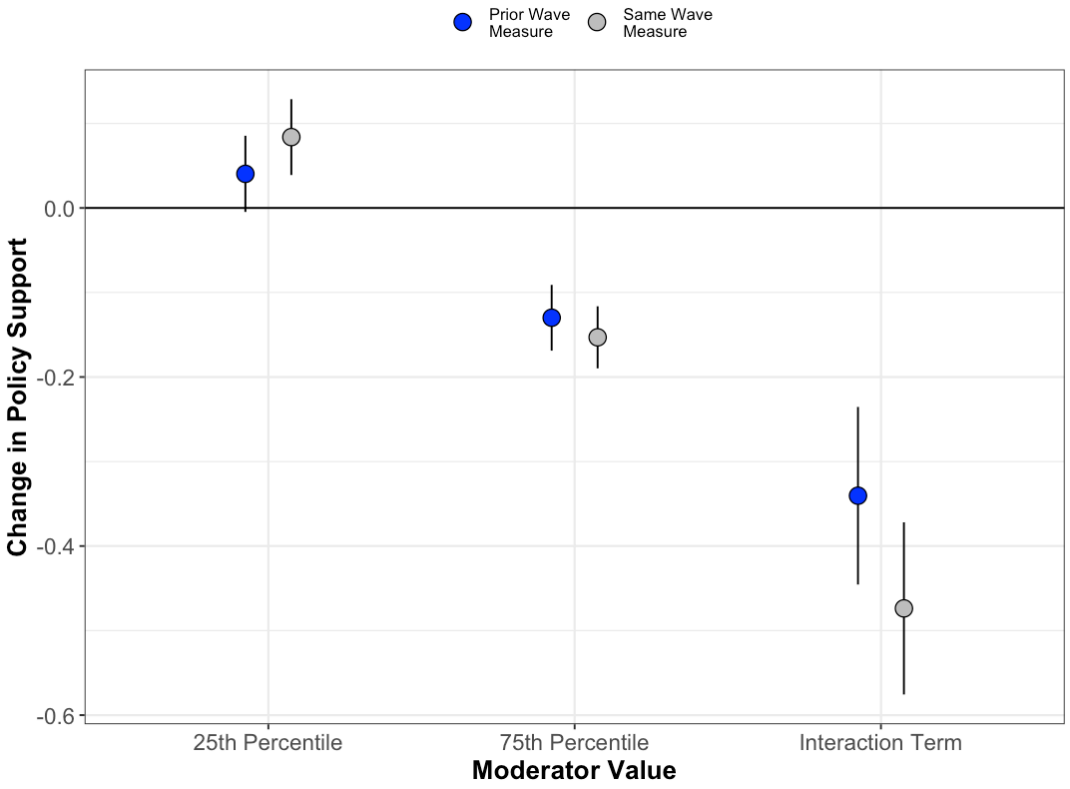
There is a positive and statistically significant relationship between the treatment and moderator. The positive interaction term reflects the fact that policy support is higher when a respondent with a strong Republican identity evaluates a policy supported by Republicans compared to a respondent with an equally strong Republican identity evaluates a policy supported by Democrats. We also find remarkable consistency in the magnitude and significance

of the interaction regardless of whether the moderator was measured in the same ($b = .34$, $p < 0.001$) or prior ($b = .33$, $p < 0.001$) survey wave.

Study 3 – Racial Resentment Moderates Race-Targeted Policy Support

Study 3 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. Respondents were randomly assigned to a version that applied the policy to all students (treatment = 0) or specifically to "Black students" (treatment = 1). Policy support serves as the dependent variable, which is coded so that higher values correspond to greater support for the scholarship program. The moderator is a 4-item scale of racial resentment, with higher values corresponds to greater racial resentment. We also measured the version of the question included in the control condition on the first survey wave and include it in our analysis to improve precision (Clifford, Sheagley, and Piston 2020). These analyses include 1,481 respondents. Figure 3 contains the results.

Figure 3. Conditional Relationship Between Racial Resentment, Treatment, and When the Moderator Was Measured



The relationship between the treatment and the racial resentment moderator is negative and statistically significant. Respondents who were higher in racial resentment were less supportive of the policy that benefited Black students compared to a policy that did not include the racial cue. For respondents in the 25th percentile of racial resentment, there was no difference in policy support between respondents in the treatment and in the control condition. For those in the 75th percentile, the difference was negative and statistically significant.

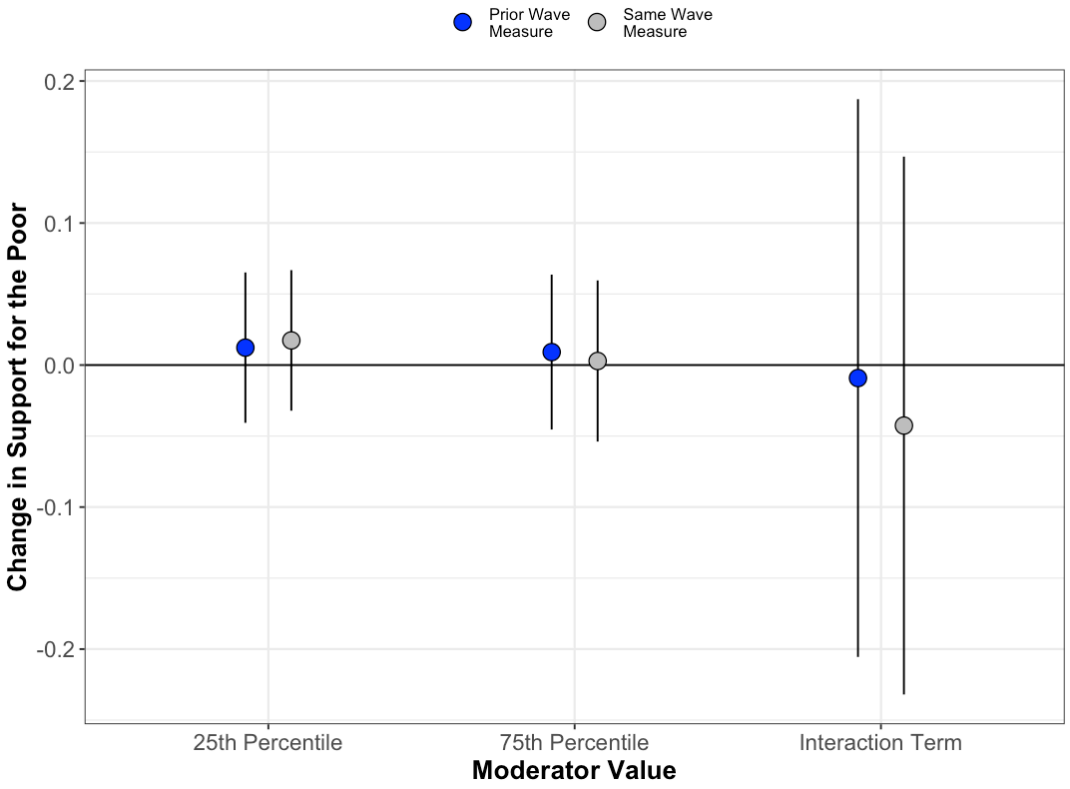
The interaction between racial resentment and treatment was negative and statistically significant regardless of where the moderator was measured. However, the interaction coefficient was a bit larger in the same wave condition (b = -.47, $p < 0.001$) than in the prior wave condition (b = -.34, $p < 0.001$). While the *difference* in these quantities is marginally significant (diff = .14, $p = 0.07$), the direction of this difference is also contrary to common concerns about the effect of

measuring racial attitudes directly before treatment. For example, Valenzuela and Reny (2001) note that measuring racial attitudes immediately before treatment, "carries a significant risk of making racial considerations salient and priming racial attitudes among all study participants, thereby washing out any differences between treatment and baseline groups." While marginally significant, the pattern of results here suggests the opposite: potentially a stronger effect of the treatment when racial resentment is measured in the same wave as the experiment.

Study 4 – Political Values Moderate Framing Effects

Study 4 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. One version of the story was designed to activate humanitarian values (treatment = 1) and the other focused on individualism (treatment = 0). After reading the article, respondents were asked about their support for whether government "should make sure the poor and children receive public assistance." These responses serve as our outcome, with higher values indicating greater support for helping the poor. The moderator is a six-item scale designed to measure humanitarian values. Humanitarianism is coded such that higher values correspond to greater value endorsement. We include a control for a respondent's party identification and pre-treatment measure of the outcome. These analyses include 526 respondents. Results are displayed in Figure 4.

Figure 4. Conditional Relationship Between Humanitarian Values, Treatment, and When the Moderator Was Measured

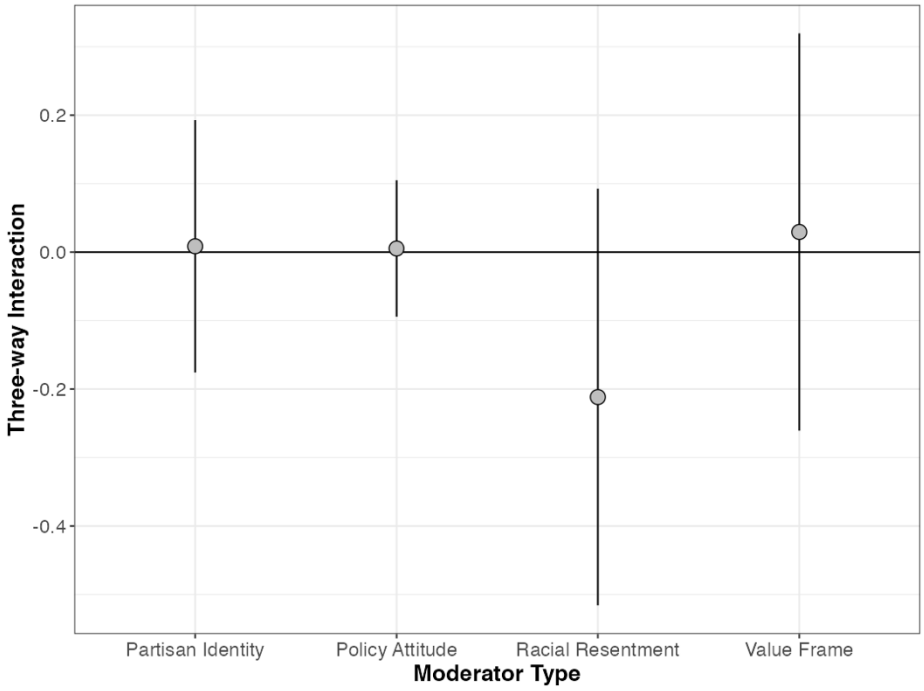


The interaction between humanitarian values and the treatment is not statistically significant and is substantively small. Thus, we observe no evidence that a respondent's level of support for humanitarian values conditions their reaction to a treatment that did vs. did not cue those values. We also observe no evidence that measurement of the moderator alters this relationship. Coefficients on the interaction term were small and non-significant in both the same-wave condition (b = -0.06, p = 0.55) and the prior wave condition (b = -0.02, p = 0.82). Thus, we found no evidence that priming humanitarian values by measuring them in the same survey wave as the experiment altered the findings from the experiment.

Overall, across four different experiments, we find substantively equivalent results when measuring the moderator in the same wave as the experiment or in a prior wave. The findings were similar for both the interaction terms and the marginal effects, and also extended to an

experiment with a null interaction. To formally test for differences between designs, for each study we pooled the two experimental conditions and estimated a three-way interaction between treatment, moderator, and placement of the moderator, along with all constituent terms and relevant covariates. The three-way interaction terms are plotted in Figure 5, which represent the difference in the two-way interactions across the two placements of the moderator. As is clear, none of the three-way interaction terms are statistically significant, and three of the four are quite close to zero.

Figure 5. Three Way Interaction Term Between Moderator, Treatment, and Moderator Measurement Condition.

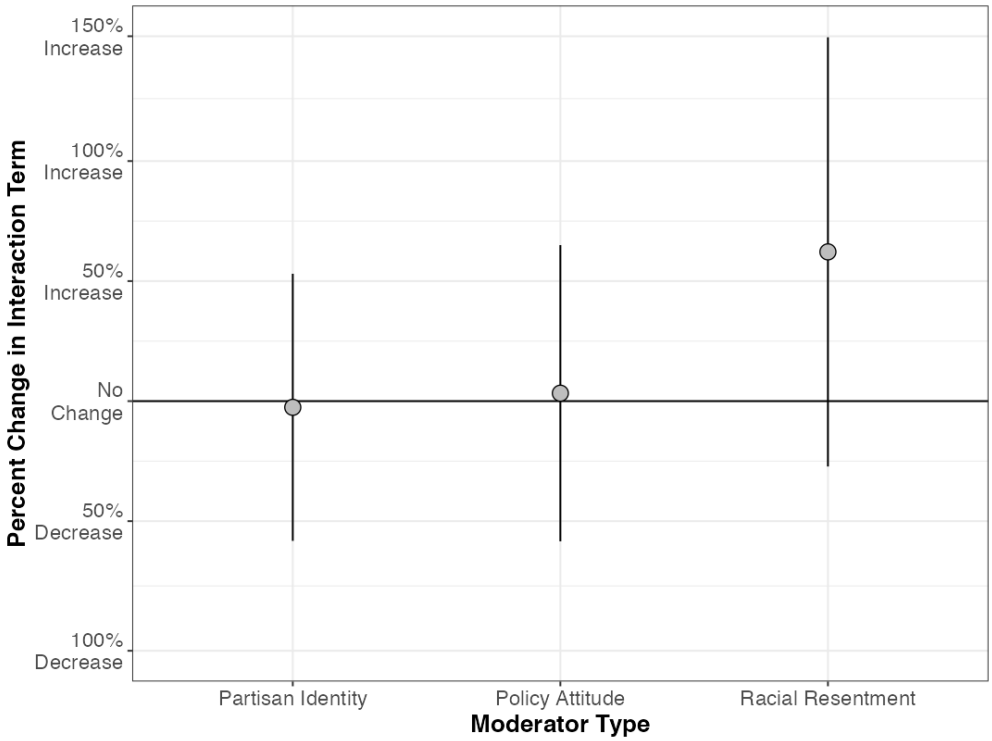


To provide an easier interpretation of these interaction terms and their confidence intervals, we rescale them relative to the size of the relevant two-way interaction. Specifically,

we divide the three-way interaction term by the two-way interaction term, as estimated in the prior wave condition. This rescaling allows us to interpret the three-way interaction term as a proportional change in the size of the moderation effect. Rescaling in this way also yields more comparable effects, as we expect any effect of the moderator placement to operate relative to the original effect size, rather than having a constant absolute effect size. The results are shown in Figure 6, though we omit the value framing study because neither condition yielded a significant moderation effect. The y-axis ranges from -100%, which means that the same wave condition completely removed the moderation effect, to 150%, which means that the same wave condition more than doubled the size of the moderation effect.

Starting on the left, the three-way interaction term for the partisan identity experiment implies that the estimated two-way moderation effect is less than 1% smaller when the moderator is measured in the same way as the experiment. However, the confidence intervals on this estimate range from a 58% decrease to a 52% increase. Turning to the policy attitude experiment, the estimate implies that measuring the moderator in the same wave increases the estimated moderation effect by 3%, with confidence intervals ranging from an increase of 65% to a decrease of 58%. Finally, the estimate for the racial resentment experiment suggests that measuring the moderator in the same wave increases the moderation effect by about 62%, although the confidence intervals extend from -27% to 150%, meaning that we cannot reject the null hypothesis of no effect of moderator placement. Notably, while this effect deviates the most from zero, most scholars have expressed concern that measuring racial resentment in the same wave of an experiment would *decrease* the moderation effect, contrary to what we find here.

Figure 6. Three Way Interaction Term Between Moderator, Treatment, and Moderator Measurement Condition.



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

To this point, we have offered evidence that whether a moderator is measured in the same survey wave as an experiment or a prior survey wave does not substantively alter the nature of the related conditional treatment effect. However, a related concern is whether the distance between the moderator and the experiment *in the same survey wave* alters a treatment effect. Conventional wisdom holds that researchers should include other relevant questions between a pre-treatment covariate, like a moderator, and the relevant experiment. For example, Montgomery, Nyhan, and Torres (2018, 773) recommend that researchers "carefully separate pretreatment questions from their experiment and outcome measures to avoid inadvertently

affecting the treatment effects they seek to estimate." Consistent with this advice, in all of our previous studies we sought to provide distance between the moderator and the experiment in each survey. For example, for study 1, the wave 2 moderator was measured at the beginning of the survey instrument and the experiment was administered near the end. In this section, we test this assumption with an experiment included in survey 4, the CES.

This sample included a conceptual replication of the race cue experiment used in study 3. Variable coding and the estimation strategy mirror those used in the related experiment. These analyses also include covariates for a respondent's party identification and pre-treatment value on the dependent variable, which was measured in a prior wave of the CES.[7] In all conditions, racial resentment was measured at the start of the survey. We then randomized whether the experiment was administered immediately after this measurement (the close measurement condition) or if it was administered at the end of the survey (the distant measurement 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, which we discuss below, the intervening questions did not explicitly measure racial attitudes and thus we assume that these questions alone did not prime these considerations. These analyses include 962 respondents.

[7] The CES employs a panel design, with the pre-election survey wave fielded in October of 2020 and the post-election wave in late November 2020.

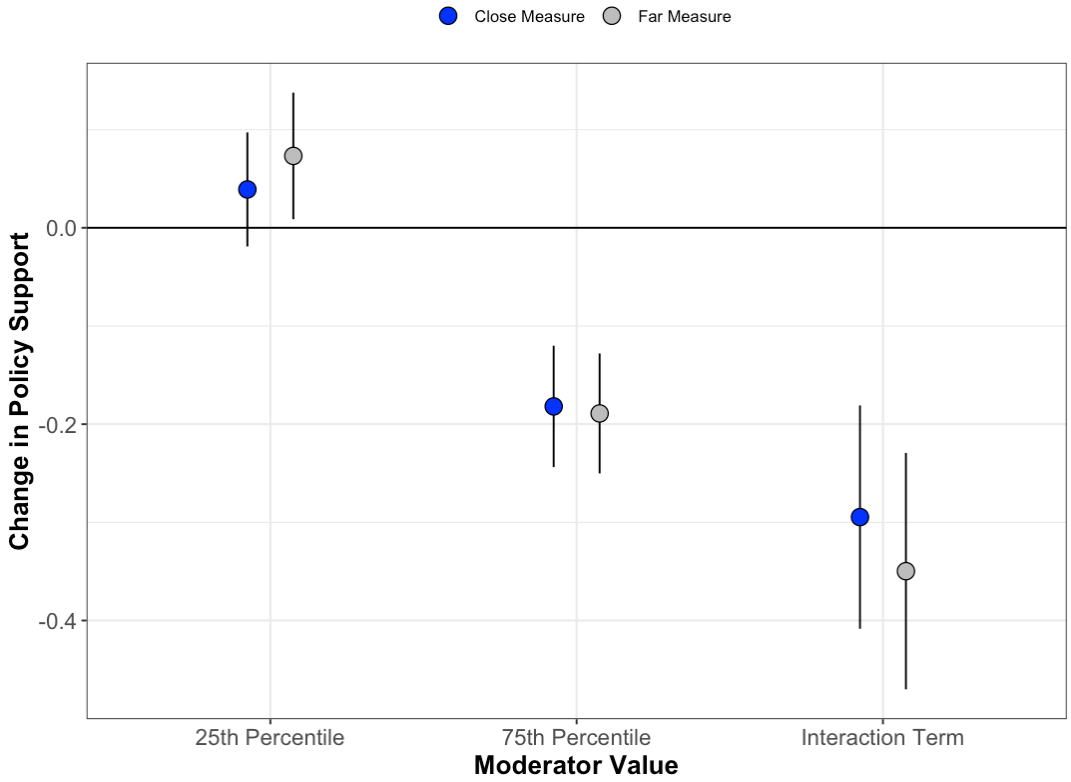
Among the policy attitudes questions, a random half of the sample was also asked about their views on whether Black Lives Matter protests should be allowed.[8] While this item complicates our test, because it was independently randomized, it also provides an additional way to assess concerns about priming. 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 begin by ignoring this feature of the survey, then by analyzing the BLM question as an additional possible prime.

Figure 6 displays the marginal effects and interaction terms for both designs (close and distant) while ignoring the randomized question on BLM. The conclusions from this experiment are the same as before: racial resentment moderates the treatment effect. Substantively, higher levels of racial resentment translate to lower levels of policy support when that policy is described as benefiting Black students compared to just students (i.e., no race cue). We observe substantively similar and statistically significant interaction effects when the experiment is administered immediately after we measure racial resentment (b = -.29, $p < 0.001$) and when it is more distant from the racial resentment measure (b = -.35, $p < 0.001$). The coefficient on the three-way interaction between treatment, measurement distance, and racial resentment is small and not statistically significant (0.07, $p = 0.45$), meaning that we cannot reject the null

[8] The other half of the sample was asked about Covid-19 protests. This portion of the survey also included policy questions that measured opinions about funding for schools, free speech, border control, COVID-19, and police funding. 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.

hypothesis that the moderating effect of racial resentment is unaffected by the distance between the moderator and the experiment.

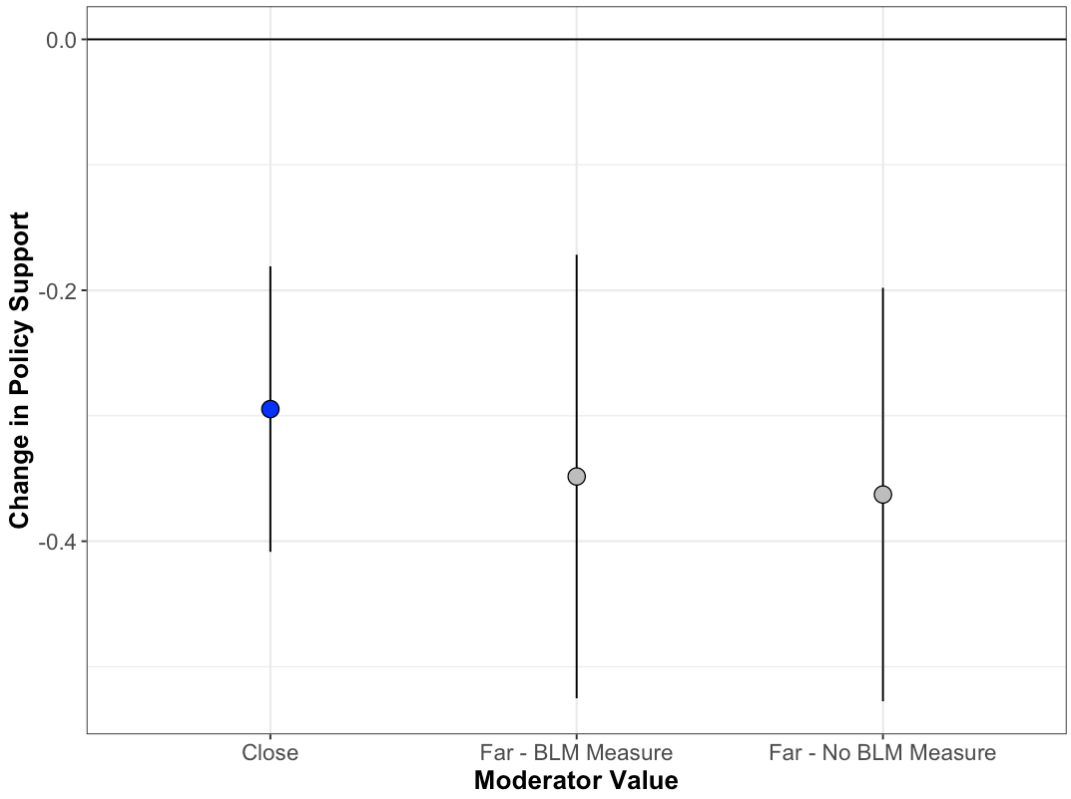
Figure 6. Conditional Relationship Between Racial Resentment, Treatment, and Measurement Location in Same Survey



We also estimated the basic interactive model among three subsets of our data: 1) when the experiment was placed close to the moderator (and prior to the BLM question), 2) when the experiment was placed distant from the moderator and the BLM was not asked, and 3) when the experiment was distant from the moderator and the BLM question was asked. The results are shown in Figure 5. As is clear, the coefficient on the interaction term is substantively identical across all four subsets of the data. When the moderator was measured close to the experiment, we find a strong negative interaction (b = -.30, p < .0001). When the moderator is measured far

from the experiment, we find similarly strong results regardless of whether the BLM question was (b = -.36, p < .0001) or was not (b = -.35, p = .0001) measured prior to the experiment. Thus, 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.

Figure 7. Conditional Relationship Between Racial Resentment, Treatment, Measurement Location in Same Survey, and BLM Measurement



Of course, we cannot rule out the possibility that the other intervening questions affected the experimental results. We find this unlikely. However, in practice, we expect that most researchers will also be choosing between placing a moderator before or after a set of political questions, rather than a set of non-political questions or tasks. Thus, our test here is

representative of the type of design choice that researchers typically face. Our findings suggest that this choice matters little.

Testing the Mechanism

Our evidence suggests that is unlikely that the measurement of a moderator will affect the results of an experiment. However, we could not rule out the possibility that measuring a moderator changes the effect size to a moderate degree, particularly for the racial resentment experiment. Research in this area, in particular, has raised the concern that measuring a moderator might affect the results by priming the measured concept and thus increasing the accessibility of that concept. We directly assess that mechanism here.

In Study 5, after completing the racial resentment experiment, all respondents were asked what considerations came to mind. Recall that half of respondents were asked about scholarships for the top 15% of students, while the other half were asked about scholarships for the top 15% of 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 race, ethnicity, discrimination, or related concepts as race-related, and all other responses as not race-related. To ensure the reliability of our coding, two authors independently coded a random 99 responses and reached 100% agreement. One of the authors then coded the remaining responses.

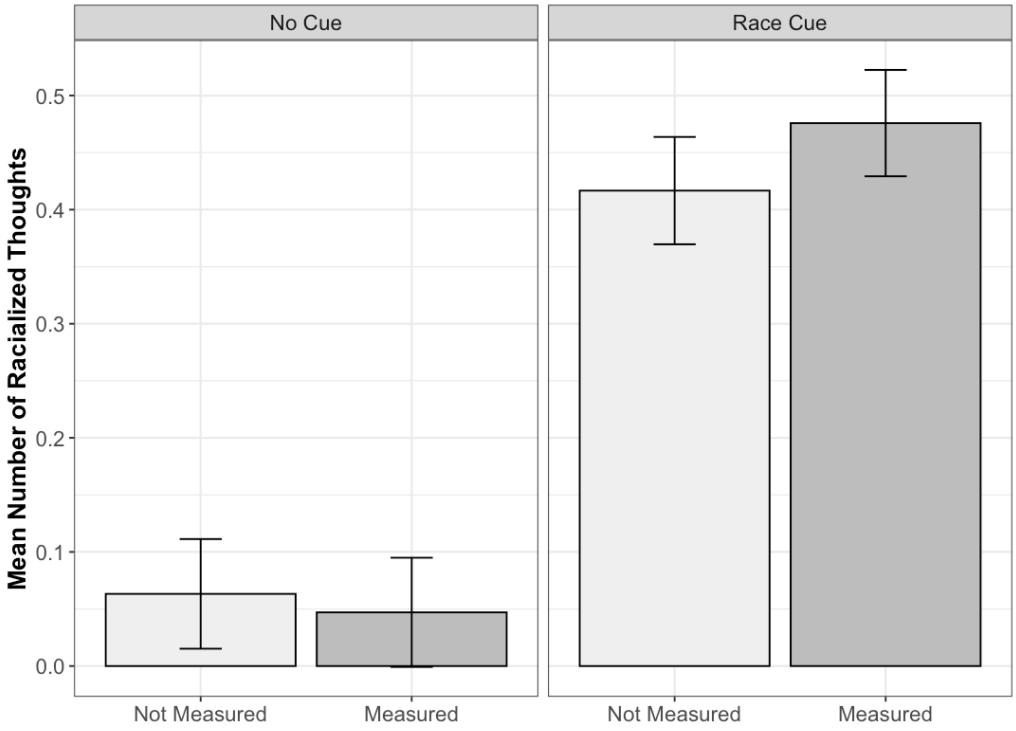
To test for priming, we use OLS to model the accessibility of racialized considerations as a function of an indicator of the treatment condition and an indicator of whether racial resentment was measured in wave 2. We also control for wave 1 racial resentment. Figure 8

displays the mean level of racialized thoughts within the four experimental conditions. Lines through each bar correspond to 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 ($p = .458$). The concern raised in the literature is that any priming effect should occur primarily within the control condition (cf., Valenzuela and Reny 2021). To test this possibility, we estimate the model described above, but restricted to only the control condition. In contrast to common concerns, measuring racial resentment in the same wave *reduces* the change of a racial consideration by about two percentage points, but this effect is not statistically significant ($p = .590$).[9] We also do not find evidence of interactions between treatment condition and moderator measurement, nor with levels of racial resentment. Thus, measurement of racial resentment does not seem to have a meaningful effect on the likelihood of raising racial considerations.

Figure 8. Mean Number of Racialized Thoughts by Race Cue Experimental Condition and Moderator Measurement

[9] Within the race cue condition, the difference between the not measured and measured conditions is roughly 5% and not statistically significant ($p = .119$).



Consequences of Using a Panel Design

So far, we've provided evidence that measuring moderators within the same wave of a survey experiment is unlikely to affect the results. Of course, some researchers may still prefer using a panel design out of caution. In this section, we discuss the implications of that design choice, particularly in terms of attrition and attitude stability. Starting with attrition, across our four experimental studies, we retained between 53% and 83% of respondents across waves. In other words, due to the panel design, we lost an average of about one-third of our sample size, which raises challenges for statistical power. If attrition is not random, then it also raises concerns about generalizability. For example, in our first MTurk panel, attrition was significantly related to partisan identity, the moderator in one of our studies. Thus, an experiment embedded in the second wave of a panel study may be less representative of the population than an experiment embedded within a single wave.

Panel designs may even yield different results if the moderator is temporally unstable. However, observed moderators tend to be durable identities and dispositions, like partisanship, that change little over time. In our own data, test-retest reliability for our moderators tended to be quite high (e.g., $r = .92$ for partisan social identity and $r = .87$ for death penalty attitudes). As a result, instability in a moderator is unlikely to be a problem for inferences. Our data allows us to test this possibility. In our main analyses above, we used the Wave 1 moderator for all experimental conditions to focus on the causal effect of exposure, rather than differences in measurement. Here we focus instead solely on measurement, restricting our analysis to only respondents who were exposed to the moderator in both waves. We then estimate two sets of models for each study – one in which we use the Wave 1 moderator and one in which we use the Wave 2 moderator. To summarize, we find no significant differences in moderation effects depending on which measure we use (see Appendix for full results). Thus, in the context of these studies, the moderators prove durable enough over a short period of time to yield the same results.

Overall, the use of a panel design seems unlikely to substantively affect results, given that common observational moderators tend to be highly stable over time. However, given substantial rates of attrition, researchers will have to recruit significantly more respondents in panel designs to offset attrition and maintain the same level of statistical power.

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 research

should measure a moderator. Central to this debate are concerns that measuring a moderator prior to an experiment could alter treatment effects. In spite of these concerns being frequently expressed in the literature and that these concerns clearly have influenced design choices, there is little direct empirical evidence on the topic.

Across four different experiments including four of the most commonly used moderators, 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. While we focused on four of the most commonly used moderators in experimental political science, our studies included only four different experiments. In general, we agree with the advice of Klar et al. (2020) that researchers should

rely on theory and evidence to inform their design choices. The clearest concerns about priming effects have been raised in the context of identity experiments, but we found no evidence of priming effects in either of our experiments that involved identity. However, we are not prepared to generalize from these results to all identity experiments. Both of our relevant experiments involved providing contextual information that allowed respondents to connect either their partisanship or racial attitudes to the outcome variable. In the case of partisan cues, the treatment was necessary for many respondents to connect their partisan identity to the policy. In the racial resentment experiment, the policy became relevant to racial attitudes when the treatment stipulated that the policy only applied to Black students. Thus, we see our findings as most likely to generalize to designs with a similar structure.

The partisan and racial identity experiments we used for our studies differ in form from some common identity experiments. Consider, for example, an experiment in which the policy outcome and relevant information are held constant, but an identity prime is expected to cause respondents to be more likely to draw on identity-relevant considerations in evaluating the policy. And this effect will be moderated by the particular identity the respondent holds. In this case, measurement of the moderator may well affect the experimental results and we caution against extrapolating our results to this type of design. However, we do expect that our results extend to a variety of other designs that involve manipulating the presence of information or cues that allow respondents to connect their identities or dispositions to the outcome variable.

In sum, our studies suggest that – for a variety of common types of survey experiments – designs that measure a moderator within a single wave, prior to the treatment, yield the same substantive results as designs that measure a moderator in a prior wave of a survey. The proximity between the moderator and the experiment seems not to matter either. 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

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. http://www.ssrn.com/abstract=2683588 (February 7, 2018).

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*. https://www.dropbox.com/s/gbw56kdk8d0aoa1/expressivecues.pdf?dl=0.

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

Clifford, Scott. 2014. "Linking Issue Stances and Trait Inferences: A Theory of Moral Exemplification." *The Journal of Politics* 76(03): 698–710. http://journals.cambridge.org/abstract_S0022381614000176 (September 2, 2014).

Clifford, Scott, Geoffrey Sheagley, and Spencer Piston. 2020. *Increasing Precision in Survey Experiments Without Introducing Bias*. https://preprints.apsanet.org/engage/apsa/article-details/5ed292bbf6cc080019a3505b.

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

Gerber, Alan S., Gregory A. Huber, and Ebonya Washington. 2010. "Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment." *American Political Science Review* 104(04): 720–44. http://www.journals.cambridge.org/abstract_S0003055410000407 (July 13, 2016).

Green, D. P., and H. L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491–511. http://poq.oxfordjournals.org/content/76/3/491 (January 22, 2016).

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. http://doi.wiley.com/10.1111/ajps.12138 (March 14, 2019).

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. http://link.springer.com/10.1007/s11109-016-9379-z (December 30, 2016).

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.

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.

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

Nyhan, Brendan, and Jason Reifler. 2014. "Does Correcting Myths about the Flu Vaccine Work?

An Experimental Evaluation of the Effects of Corrective Information." *Vaccine* 33(3): 459–64. http://www.sciencedirect.com/science/article/pii/S0264410X14015424 (December 16, 2014).

Schaffner, Brian F., Stephen Ansolabehere, and Samantha Luks. 2021. *Cooperative Election Study Common Content, 2020*.

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. https://www.journals.uchicago.edu/doi/10.1086/694845 (December 9, 2018).

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.

Walter, Annemarie S., and David P. Redlawsk. 2019. "Voters' Partisan Responses to Politicians' Immoral Behavior." *Political Psychology* 40(5): 1075–97. https://onlinelibrary.wiley.com/doi/abs/10.1111/pops.12582 (February 24, 2020).