Political Science Research and Methods (2025), 1–22
doi:10.1017/psrm.2025.10017

ORIGINAL ARTICLE

Where to place sensitive questions? Experiments on survey response order and measures of discriminatory attitudes

Amanda Sahar d'Urso[1], Tabitha Bonilla[2] and Genni Bogdanowicz[3]

[1]Department of Government, Georgetown University, Washington, DC, USA; [2]Human Development and Social Policy, Political Science, Institute for Policy Research, Northwestern University, Evanston, IL, USA and [3]Northwestern University, Evanston, IL, USA
Corresponding author: Amanda Sahar d'Urso; Email: amanda.durso@georgetown.edu

*Author order has been reverse alphabetized. Amanda Sahar d'Urso and Tabitha Bonilla have been involved in the project from the beginning and have both worked to theorize the paper, analyze the data, and write the paper. Genni Bogdanowicz has worked to analyze data.

(Received 26 December 2023; revised 20 January 2025; accepted 27 January 2025)

Abstract
In survey experiments, should all covariates be administered before the experimental treatment? Some scholars argue that post-treatment items should never be used as covariates because the treatment could bias the measurement of those items and disrupt experimental randomization. Other scholars argue certain items—specifically sensitive questions measuring prejudice—should not be administered pre-treatment. They argue if asked pre-treatment, these items may prime respondents in ways that will influence how they engage with the experiment treatment, thereby affecting the overall outcome of the experiment. Using evidence from four studies (two original collections) that vary the placement of sensitive items—pre-treatment, post-treatment, or in a separate wave—we find little evidence that the placement of sensitive items influences the measurement of those items, the experimental outcomes, nor heterogeneously affects the outcome conditional on the treatment. However, we find the placement of sensitive items inconsistently affects the experimental outcome by interacting with both the measurement of the items and the experimental treatment condition. Overall, we find these measures to be robust to where they are administered. It may be best to place items pre-treatment to preserve randomization. If researchers have reason to include sensitive moderators post-treatment, they should transparently discuss this choice and the anticipated trade-offs.

Keywords: survey experiments; question order; senstive questions; survey measurement; survey moderators; measuring prejudice; racial attitudes gender attitudes

1. Introduction

Where should researchers place items that measure prejudice and bias in experimental studies? Despite recent work addressing where to include moderators in surveys (Montgomery *et al.*, 2018; Coppock, 2019; Klar *et al.*, 2020; Albertson and Jessee, 2023; Sheagley and Clifford, 2025), collective guidance is still unclear with respect to the measurement of prejudicial beliefs. Currently, experimental tests suggest that if researchers intend on conditioning experimental analyses on a given variable, that variable should be measured before the administration of the experimental treatment due to concern that experimental treatments could alter the measurement of items administered post-treatment

(Montgomery *et al.*, 2018; Coppock, 2019). However, there is concern that some covariates may have the potential to alter the outcomes of an experiment if these items are administered pre-treatment because the moderators might influence responses to the experimental treatment (Klar *et al.*, 2020). While some have empirically tested these concerns with commonly used moderators (Albertson and Jessee, 2023; Sheagley and Clifford, 2025), we extend this work with the unique case of sensitive questions—those that measure prejudice and group bias. Sensitive questions may prime respondents to think about groups (e.g., by ethnicity, race, religion, or gender) differently if asked pre-treatment, causing respondents to interact with the experimental treatment in ways that alter the experimental outcome (e.g., Williams *et al.*, 2008; Benjamin *et al.*, 2010). This paper provides much-needed empirical guidance on the matter for scholars studying prejudice and attitudes toward marginalized groups, particularly in (but not limited to) the subfield of racial and ethnic politics.

To examine these arguments, we empirically test whether the placement of sensitive items in an experiment (1) affects the measurement of those items, (2) affects the experimental outcome, (3) heterogeneously affects the experimental outcome conditional on the experimental treatment, and (4) heterogeneously affects the experimental outcome based on the experimental treatment conditional on the placement of the sensitive items.[1] We provide evidence from five experiments across four studies. These studies include two original experiments—using racial resentment and Muslim American resentment (MAR)—and two previously published studies—varying placement of symbolic racism and symbolic sexism.

Across these four studies, we find that the placement of the sensitive items (1) does not change the measurement of the sensitive items themselves, (2) does not change the measurement of the outcome, (3) largely does not affect the experimental outcome by interacting with the treatment condition, but (4) inconsistently affects the experimental outcome by interacting with the measurement of the items themselves and the experimental treatment condition. However, even when we find differences, these differences rarely result in researchers drawing a different conclusion from the experiment based on the placement of sensitive items. Thus, these results indicate that asking individuals about prejudicial beliefs does not seem to influence how respondents engage with the treatment itself. As such, we conclude that placing these items post-treatment is very unlikely to change the measurement of these items; at the same time, placing sensitive items pre-treatment may also not affect the experimental outcome. In most cases, we recommend that researchers follow Montgomery *et al.* (2018) because post-treatment measurement could still disrupt experimental randomization. However, if researchers remain concerned about the internal validity of the experimental treatment, they should discuss their decision-making process and how they may address concerns about bias introduced by weakening assumptions of randomization.

2. State of the field: where are scholars placing sensitive items in their experiments?

Researchers have long considered best practices around measuring sensitive questions in survey research. This research began as a way to examine how survey researchers may address variability from known outcomes to lower-than-expected outcomes on surveys (e.g., voting) (Tourangeau and Smith, 1996). Overall, sensitive questions have three criteria: (1) they are intrusive, (2) they increase the risk of identification, and (3) the social desirability of a particular outcome (Tourangeau *et al.*, 2000). Each factor can cause bias in the item estimate (Rasinski *et al.*, 1999; Tourangeau and Yan, 2007) and research innovations attempt to address these issues (e.g., Tourangeau and Yan, 2007; Kreuter *et al.*, 2008; Näher and Krumpal, 2012; Lehrer *et al.*, 2019). Political science work, particularly race and ethnic politics research, often considers questions that measure group bias. This includes

[1]Sheagley and Clifford (2025) show that placement of moderating covariates does not alter the moderating effect of that variable. However, while Sheagley and Clifford (2025) "focus on commonly used moderators," we focus specifically on sensitive items critical to understanding the causes and consequences of prejudice. Indeed, the one moderator that did seem to differ as a result of placement was racial resentment (Sheagley and Clifford, 2025, p. 11).

discussing the difficulty of measurement (e.g., Berinsky *et al.*, 2012; Mo and Bonilla, 2020) as well as potential solutions (Blair, 2015; Blair *et al.*, 2020).

Given the posturing of research that engages in how to best measure public opinion, increasing use of experiments in research (Robison *et al.*, 2018), and the need to understand best practices in how to measure sensitive attitudes, an underlying tension has come to the forefront. Current studies demonstrate how conditioning on variables measured post-treatment could potentially disrupt the measurement of the conditioning variable (King and Zeng, 2006; Montgomery *et al.*, 2018; Coppock, 2019). Montgomery *et al.* (2018) primarily argue conditioning results on post-treatment variables disrupt experimental randomization because once a group is divided into treatment and control conditions, they are not answering a post-treatment question from the same starting point. But Montgomery *et al.* (2018) warn of "treatment spillovers like racial resentment," which "should be measured pre-treatment" (Montgomery *et al.*, 2018, p. 771). However, Klar *et al.* (2020) voice concern that these items can also cause spillover on the outcome of the experiment.

Arguments around measurement spillover and order effects around the placement of measures of discrimination are not new. For example, Huber and Lapinski (2006), in criticism of Mendelberg (2001), place the racial resentment battery pre-treatment, and find that implicit racial priming is no more effective than explicit racial appeals, undermining a key finding in Mendelberg (2001) and many others (e.g., Entman and Rojecki 2001; Valentino *et al.* 2002). Mendelberg (2008) critiques the research design on the grounds that pre-treatment placement of the racial resentment battery mutes the effects of following treatments, as respondents likely noticed the racial appeals in the implicit appeals treatment since they were just previously primed to think explicitly about their racial attitudes. Published research reflects the lack of conformity to best practices. We identified every experiment published in top, general field journals in political science between 2010 and 2018 (i.e., following the Mendelberg-Huber debate) to determine what type of guidance one might receive from extant literature on the placement of sensitive questions aimed at measuring prejudice.[2] In total, we found 19 articles from 2010 to early 2018 that used experiments and racial prejudice items. Of these 19 articles, seven (36.8%) use the sensitive items post-treatment, five (26.4%) use the items pre-treatment, and seven (36.8%) use the items in a multi-wave manner, often separating the measurement of sensitive items (most commonly racial resentment) from the treatments by a number of weeks.[3] Most often, researchers justify measuring racial attitudes post-treatment by arguing that pre-treatment measurement could result in pre-treatment effects that biased the experimental results (McConnaughy *et al.*, 2010; Baker, 2015; Hassell and Visalvanich, 2015).

3. Where *should* sensitive items be administered?

Consistent with the evidence of conflicting placements, guidance from research on best practices is also in disagreement. Montgomery *et al.* (2018) argue that by using post-treatment variables as covariates, researchers can introduce bias in two ways: (1) the treatment can affect the measurement of sensitive items, which could lead to (2) conditioning on post-treatment variables can "ruin" experimental randomization. Therefore, when conditioning on the post-treatment variable, the differences between treatment and control group are no longer based solely on the treatment, but on how the treatment influences the measurement of the covariate.[4] Though Montgomery *et al.* (2018) model the latter, the question of the former remains untested with respect to sensitive measures and is not

[2]We did not search sub-field journals to restrict the scope of the search and because one of the top sub-field journals in race and ethnic politics was not active for most of the observation period.

[3]Our coding and findings can be found in Appendix Table A.1.

[4]See example from Montgomery *et al.* (2018). If an experiment is testing the efficacy of a civic education program and political interest (binary either high or low) is measured post-treatment, then the comparison between the treatment and control conditional on political interest is actually a comparison between those with low (high) political interest *after* receiving the civic engagement treatment and those with low (high) political interest who did not receive the treatment.

always supported by additional data collection (e.g., Schiff *et al.*, 2022). On the other hand, Klar *et al.* (2020) argue that placing sensitive items before the experimental treatment might affect the internal validity of an experiment by "[changing] 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" (p. 3). In particular, some studies have demonstrated that measuring sensitive items pre-treatment can induce changes in subsequent items, either by moderating results as with partisan identities (Klar, 2013) or directly affecting attitudes as with measurement of racial prejudice (Hutchings and Jardina, 2009; Steele, 2011; Hussey and De Houwer, 2018) and ethnic identities (Jackson, 2011; Ostfeld and Pedraza, nd).

In response, recent scholarship has attempted to empirically test these concerns. Sheagley and Clifford (2025) test whether the placement of "commonly used moderators" immediately pre-treatment or in a separate wave prior to the experimental treatment influences the magnitude of the treatment effect across the moderating variable. They find that the placement of the covariate relative to the treatment *does not* change the average treatment moderation effect in nearly all cases—except for racial resentment—a measure we include in our definition of sensitive items. Thus, of all the tests, the study with racial resentment is the only case in which Sheagley and Clifford (2025) "reject the null hypothesis of no effect of the measurement prime" (p. 11). Albertson and Jessee (2023) directly test concerns by Montgomery *et al.* (2018) and Klar *et al.* (2020) using racial resentment. They find that the placement of racial resentment does not alter the measurement of racial resentment, nor does it interact with the substantive treatment to heterogeneously influence the experimental outcome when measured pre- or post-treatment.

We expand on the empirical strategies of these existing studies by administering sensitive items at three distinct times in the study: pre-treatment, post-treatment, and in a separate wave prior to the experimental treatment and outcomes.[5] This matches the variation we see in the analysis of published work using measures of prejudice and across the empirical designs we see represented in research. Importantly, as we discuss later, there are important methodological concerns that are made when measuring moderators at each point in a survey experiment.

We also consider a broader, theoretically motivated set of moderators beyond anti-Black racism—e.g., racial resentment and symbolic racism (Kinder and Sears, 1981). Because both Klar *et al.* (2020) and Montgomery *et al.* (2018) mention concerns about questions with the potential for spillover effects, including but not limited to measures of anti-Black prejudice, it is important to test a broader category of moderators. We include other moderators aimed at measuring prejudice, including MAR (Lajevardi and Oskooii, 2018; Lajevardi, 2020) and symbolic sexism (Pingree *et al.*, 1976; Benokraitis and Feagin, 1995).

Despite different concerns about when to measure moderators, both Sheagley and Clifford (2025) and Klar (2013) are concerned about priming, albeit on different measures. Priming in an experimental context occurs when exposure to a stimulus—in this case, an experimental treatment—alters how individuals respond to subsequent questions (e.g., Iyengar, 2008; Cassino and Erisen, 2010; Iyengar and Kinder, 2010). This occurs because the priming stimulus makes certain information more relevant than others by activating brain activities (Fiske *et al.*, 1993). We argue that sensitive items aimed at measuring prejudice are theoretically important to test this methodological question because these survey questions may elicit strong, emotional reactions. This is of particular importance to research on race, which has focused specifically on the effects of priming race and racial prejudice. Messages on policies can be framed in ways that prime respondents to increase the weight they put on racial considerations and potentially shift how respondents answer questions on the bias that are often used to moderate these experiments (Hutchings and Jardina, 2009; Chong and Junn, 2011, p. 320). For example, studies comparing in-group identity relative to out-group animosity show that out-group

[5]Sheagley and Clifford (2025) do not measure racial resentment post-treatment and Albertson and Jessee (2023) do not measure racial resentment in a separate wave.

attitudes are stronger predictors of public opinion and political behavior (Buyuker *et al.*, 2021; Jardina, 2021; Rathje *et al.*, 2021; Cuevas-Molina, 2023). This may also help explain why sensitive items would be more susceptible to variation relative to the treatment rather than a measure like partisanship. It follows that asking respondents about their views on race (or gender or identity broadly) may shift responses on either the experimental outcome (if bias is measured first) or the sensitive measure (if bias is measured second). Whereas Sheagley and Clifford (2025) discuss the interaction between the placement of the moderator, the measurement of the moderator, and the treatment on the experimental outcome, Klar *et al.* (2020) are more concerned with the effect the placement of the moderator will have on the measurement of the experimental outcome. However, both questions are important to examine, especially when considering items that are known to be sensitive to priming.

Importantly, the measurement of sensitive items rests not just on concerns about the measurement of these items based on when they are administered (Montgomery *et al.*, 2018) but also on concerns that social desirability may bias how individuals respond to these sensitive items (Kreuter *et al.*, 2008). There has long been debate about whether social desirability causes some respondents to answer explicit prejudice questions more favorably than their attitudes truly are (Mo, 2015), which suggests that these attitudes may be susceptible to priming when combined with research demonstrating the influence of measures of bias on political attitudes (Valenzuela and Reny, 2021; Gothreau *et al.*, 2022). In the first case, respondents with more egalitarian attitudes may aim to answer items aimed at measuring their prejudice pre-treatment and may respond to treatments with more heightened awareness to check their bias. Conversely, respondents with less egalitarian attitudes may be primed to be particularly negative toward interventions that mention targets of their bias. As a result, the priming treatment may work simultaneously to exacerbate differences between treatment and control, or to minimize differences between treatment and control.

However, if the experimental treatment is asked first, respondents may be primed to respond differently to the sensitive question. For instance, treatments that may cause those with high levels of bias to act differently than treatments that do not While some past research has demonstrated that sensitive measures can be stable to question order, other research has demonstrated differences in the estimates of these batteries which suggests that these attitudes may not always be stable (Kam and Burge, 2019; Smith *et al.*, 2020). This is particularly true for survey batteries that may be sensitive to other demographic influences (DeSante and Watts Smith, 2020; Banda and Cassese, 2022).

4. Hypotheses

Based on the abovementioned discussion, we proffer four hypotheses for how the placement of sensitive items relative to an experimental treatment might affect measurement. As we have no *a priori* notion of direction, all hypotheses are two-tailed. Hypotheses 1–3 are listed in our pre-analyses plans in the Appendix (see Figures B.1 and C.4). We first focused on these questions given the primary concerns introduced by Montgomery *et al.* (2018) and Klar *et al.* (2020). However, we did not include a preregistered hypothesis for a three-way interaction between the placement of the sensitive item, the measurement of the sensitive item, and the experimental treatment (Hypothesis 4)—the primary concern introduced by Sheagley and Clifford (2025). Though not preregistered, we include this test in the analyses.

$$
\begin{aligned}
&\textit{Let},\\
&\quad D = \textit{Experimental treatment}\\
&\quad Y = \textit{Experimental outcome}\\
&\quad S = \textit{Sensitive item}\\
&\quad T = \textit{Placement of the sensitive item } (\textit{e.g.}, \textit{ pre-treatment}, \textit{ post-treatment}, \textit{ or in two waves})
\end{aligned}
$$

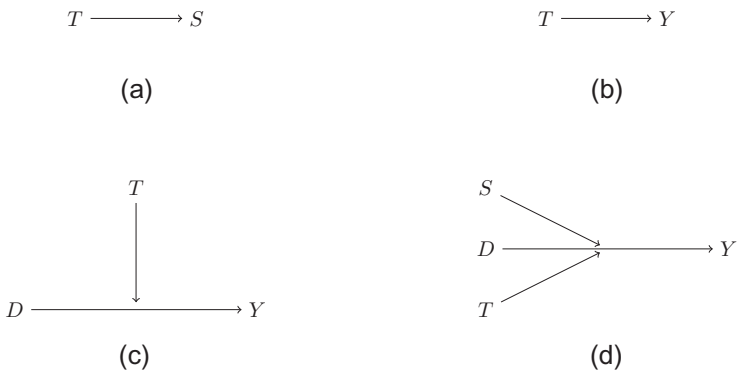

Figure 1. Directed acyclic graph of hypotheses. (a) Hypothesis 1 H_1: Placement of the sensitive items (T) could influence the measurement of the sensitive items (S). (b) Hypothesis 2 H_2: Placement of the sensitive items (T) could influence the measurement of the experimental outcome (Y). (c) Hypothesis 3 H_3: Placement of the sensitive items (T) moderates (or interacts with) the effect of the experimental treatment (D) on the experimental outcome (Y). (d) Hypothesis 4 H_4: Placement of the sensitive items (T), the measurement of the sensitive item (S), and the experimental treatment (D) could heterogeneously affect the experimental outcome (Y).

Our first hypothesis addresses the first concern of Montgomery *et al.* (2018): whether the placement of the sensitive items changes estimates of the sensitive items themselves. Consider a case where S, our sensitive item, captures the effect of the placement of that item (T). If $S_i = \beta_0 + \beta_1 T_i + \epsilon$, and $\beta_1 T_i \neq 0$, one concern in Montgomery *et al.* (2018) is realized. However, if $\beta_1 T_i = 0$, then there is less concern about using the sensitive items as a post-treatment moderator.[6] Therefore,

H_1: The placement of the sensitive items will affect the measurement of the sensitive items (Figure 1a).

Next, we address whether the placement of the sensitive items prior to the experimental treatment could potentially affect the results of the experiment (i.e., Klar *et al.*, 2020). Consider the placement of the sensitive items T and a main outcome variable Y, where $Y_i = \beta_0 + \beta_1 T_i + \epsilon$. If $\beta_1 T_i \neq 0$, then pre-treatment administration undermines the internal validity of the experiment itself. But, if $\beta_1 T_i = 0$, the placement of the sensitive item appears to have no effect on the experimental outcome. Therefore,

H_2: The placement of the sensitive items will affect the measurement of the experimental outcome (Figure 1b).

Third, we investigate if there is an interaction between the administration of the sensitive items relative to the experimental treatment and the experimental treatment itself on the experimental outcomes. Even if $\beta_1 T \neq 0$, it is still possible for pre-treatment items to have minimal effects on an experimental outcome *if* the experimental outcome is simply measuring average differences between groups. If the treatment produces the same point-estimate change across each group, the effect may be less worrisome because it results in the same average treatment effect.[7] However, the placement of sensitive items could heterogeneously affect the experimental outcome by interacting with the treatment conditions. Our third test then investigates this possibility by estimating the interaction between the timing of sensitive items and experimental treatment on the results of the experimental outcome:

[6]The concern of losing the benefits of randomization from Montgomery *et al.* (2018) remains, but we are first addressing the concern about the internal validity. In Hypothesis 2 and 3, we address the concern raised by Klar *et al.* (2020) measuring moderators prior to treatment may introduce bias in the treatment randomization as well.

[7]For studies that use absolute outcomes to validate experimental results, this would be concerning.

$Y_i = \beta_0 + \beta_1 T_i + \beta_2 D_i + \beta_3 T_i \cdot D_i + \epsilon$. Here, if $\beta_3 \neq 0$, then the placement of sensitive items pre-treatment is especially concerning because it completely undermines the results of the experiment. If $\beta_3 = 0$, then placement of the sensitive question pre-treatment is of no concern. Therefore, we hypothesize:

H_3: The interaction between the placement of the sensitive items and experimental treatment will affect the experimental outcome (Figure 1c).

Fourth, we investigate if there is an interaction between the administration of the sensitive items relative to the experimental treatment, the experimental treatment itself, and the measurement of the sensitive item itself on the experimental outcome. Even if $\beta_1 T \neq 0$, it is still possible for pre-treatment items to have minimal effects on an experimental outcome *if* the experimental outcome is heterogeneously affected by the sensitive item themselves and the experimental treatment groups. This is especially important given the rationale behind measuring sensitive items in these experiments is to test heterogeneity between those items and the experimental treatment on the experimental outcome. We investigate this possibility by estimating the interaction between the administration of sensitive items relative to the treatment, the measurement of the sensitive item, and the experimental treatment on the results of the experimental outcome: $Y_i = \beta_0 + \beta_1 T_i + \beta_2 D_i + \beta_3 S_i + \beta_4 T_i \cdot D_i \cdot S_i + \epsilon$. If $\beta_4 \neq 0$, then the placement of sensitive items pre-treatment undermines the results of the experiment. If $\beta_4 = 0$, then placement of the sensitive question pre-treatment is of no concern. Therefore, we hypothesize:

H_4: The interaction between the placement of the sensitive items, the measurement of the sensitive item, and the experimental treatment will affect the experimental outcome (Figure 1d).

5. Data

We use data from four studies—one with two experiments—to test the potential effects of the placement of sensitive items on both the measurement of the sensitive items and the experimental outcome. We collect original data from two studies and examine data from two studies fielded by others who randomized the placement of sensitive items relative to the experimental treatments (Table 1). In two experiments, respondents are randomly assigned to either respond to sensitive items pre- or post-treatment, and then are *independently* randomly assigned to the substantive experimental treatment or control group. Subjects in three additional experiments are treated similarly; however, the sensitive items are incorporated in either a two-wave condition (preceding the experimental treatment by a week), immediately pre-treatment, or immediately post-treatment.[8] The two-wave conditions provide the ability to separate the recency of the treatment from the sensitive items. The data are briefly described later, with each study discussed in more detail in Appendices B.1–B.4, C.1–C.4, D.1–D.3,[9] and E.1–E.4.

The selection of these studies is intentional; we test our hypotheses on a combination of studies that have been successfully published and that are novel to help inform researchers on types of decision-making around the placement of sensitive moderators. We also intentionally leverage a variety of sensitive questions across the studies. Finally, while most designs are 2x2 messaging tests, because of the increase in conjoint experiments, we test this format as well. This collection of studies suggests that the findings approximate and inform a variety of work that researchers facing questions about the placement of sensitive moderators may have.

[8]We do not focus solely on cases with two-wave administrations for two important reasons. First, many researchers do not have the funds to be able to field multi-wave survey data collections, and we need to understand single-wave collections. Second, multi-wave studies induce additional issues such as attrition between waves which may present a challenge for randomization.

[9]There were no sample demographics available for Study 3 in either the replication file or the original study and Appendix.

https://doi.org/10.1017/psrm.2025.10017 Published online by Cambridge University Press

Table 1. Summary of studies

Study	Source	Summary	Sensitive items (*S*)	Treatment (*D*)	Outcome (*Y*)	Placement (*T*)
1	Original Data	Understanding White perceptions of #BLM	Racial Resentment	-Nationalist -Feminist	-Support for #BLM Goals	-Pre -Post
2	Original Data	Understanding White attitudes toward Muslims of different races	Muslim-American Resentment	-Conjoint	-Immigrant Picked	-Two-wave -Pre -Post
3a	Valentino *et al.* (2018)	Racial rhetoric on public opinion	Symbolic Racism	-Implicit -Explicit	-Health Care -Leader	-Two-wave -Pre -Post
3b	Valentino *et al.* (2018)	Racial rhetoric on public opinion	Symbolic Racism	-Implicit -Explicit	-Health Care -Leader	-Two-wave -Pre -Post
4	Mo (2015)	Evaluation of judicial candidates based on gender and quality	Symbolic Sexism	-Candidate Strength	-Candidate Score	-Pre -Post

5.1. Study 1: Black Lives Matter experiment

We follow Bonilla and Tillery (2020) in fielding a study asking respondents about Black Lives Matter (BLM) and intersectionality. We extend their study by asking how White subjects perceive BLM. Importantly, while this study replicates the instrument of Bonilla and Tillery (2020), we field the survey to a White sample instead of a Black sample. We hypothesize that a White sample may be more responsive to a treatment discussing Black subgroups (Black women) over a treatment that speaks to Black unity (Black nationalism). In particular, we hypothesize that respondents with higher levels of racial resentment might not differentiate between the two treatments because racial bias may make respondents less inclined to support Black movements in general. Those with lower levels of racial resentment are expected to be more opposed to the Nationalist treatment compared to the Feminist treatment.

We fielded the experiment in May 2019 to 885 White Americans on Lucid (Coppock and McClellan, 2019) and assigned participants to one of four experimental conditions. Half of the respondents received the racial resentment battery pre-treatment, and the other half post-treatment. Within those conditions, half were assigned to one of two substantive experimental conditions. The experimental control gives a description of BLM that strongly emphasizes both the distinctness of the Black experience and presents a unifying call for Black people as a whole. The experimental treatment group received a treatment that strongly emphasized the particular experience of Black women in regard to violence. Since we argue the BLM treatment highlighting the role of Black women in the movement will affect how individuals perceive BLM, we estimate this effect by asking respondents how much they support BLM's goals as the outcome.

5.2. Study 2: Muslim green cards and MAR

We ask how White American respondents understand intersections of religion and race, particularly as it relates to Muslim and Middle Eastern and North African (MENA) identities. Muslim and MENA individuals are often conflated, despite key differences between religious and ethnoracial groups (Beydoun, 2013; Lajevardi, 2020; Aziz, 2021; d'Urso and Bonilla, 2023; d'Urso, 2024). In our experiment (d'Urso and Bonilla, 2023), we assess how White Americans evaluate immigrants when they are forced to consider Muslim and MENA identities both together and separately. Thus, we examine whether religion, race, or both inform White Americans' decisions of which migrants belong in the United States.[10] 590 White American participants were recruited from Bovitz, Inc. The conjoint experiment presented is forced-choice; participants are asked to select one of two presented applicants to whom they would prefer to grant a green card. Conjoint experiments of this kind are increasingly common in the social sciences, particularly to study attitudes toward migrants (e.g., Bansak *et al.*, 2021; Lajevardi, 2020; Clayton *et al.*, 2021; Denney and Green, 2021; d'Urso and Bonilla, 2023). Here, we randomly vary the applicants' education, gender, English language proficiency, religion, and country of origin. Participants were exposed to five different selection tasks wherein they were asked the forced-choice question, "Which immigrant do you think the US should give a green card to?"

The sensitive item in this study is MAR. This battery has been used and validated in recent studies of attitudes toward Muslim Americans (Lajevardi and Oskooii, 2018; Lajevardi and Abrajano, 2019) and was developed to capture anti-Muslim affect among participants similar to the purpose of capturing racial resentment toward Black Americans. The sensitive items were placed in a separate survey wave a week ahead of treatment, or in the same survey as treatment and measured before or after the experiment was presented. To ensure that all respondents received the substantive treatment of the immigrant profiles and to preserve randomness in our treatment assignment, we worked with the survey provider to randomly assign those in their pool of respondents to one of our three

[10]This study was preregistered at AsPredicted and included in Figure C.2.

placement conditions.[11] Those who received the two-wave condition were invited to fill out a basic questionnaire that included MAR one week prior to a survey that included the migrant attitudes experiment. Invitations for the two groups receiving MAR immediately pre-treatment and immediately post-treatment were sent at the same time as the second wave. Thus, everyone received profiles during the same time period.

Importantly, because the experiment embedded the two-wave design, we can report attrition, unlike many multi-wave studies. Wave 1 collected 303 responses while wave 2 collected 213 in the two-wave condition. This means approximately 29.7% of the two-wave treatment group did not respond to the second wave of the experiment. We report the balance table across the placement groups in Appendix Table C.11. Across age, gender, partisanship, and ideology, we see no differences in our sample. However, respondents report a significantly lower level of income between the two-wave group ($p = 0.034$) and the pre-treatment sample, but not in any other comparisons.[12]

5.3. Study 3: implicit and explicit messages and racial resentment

We use data from Valentino *et al.* (2018), which examines how U.S. adults respond to explicit and implicit racial appeals and form attitudes on health care and social welfare policy. The original study uses four different experiments to test their hypotheses; of these, two of the studies vary the order in which the moderating question is asked in the order of the study. These studies feature a candidate messaging experiment where the candidate explicitly or implicitly primes race. In one version of the message, the candidates then discuss the Affordable Care Act (ACA) or social welfare policy.

They field four surveys in their paper, and we leverage two of these studies because they randomize where symbolic racism (similar to racial resentment) was presented in the experiment relative to the treatment. For both Experiment 1 and 4, the authors measure symbolic racism one week prior to the remainder of the experiment or in the same wave before and after the experiment. However, since this was not the initial purpose of the study, we do not have information about respondent attrition between waves. Both experiments are 2x2 messaging experiments based on real newspaper stories from Hartford, Connecticut, varying race (Black and White) and message type (explicit and implicit). Study 1 uses racially coded language to refer to the city versus the suburbs in the implicit condition and "Blacks" and "Whites" in the explicit condition. Study 4 uses criticism of social welfare legislation with "the poor" for implicit language and "Blacks" in the explicit condition. Respondents are then asked about their support for health policy and the ACA or about social welfare policies generally.

5.4. Study 4: gender, implicit and explicit sexism, and candidate qualifications

For our final study, we use data from Mo (2015) and Mo and Bonilla (2020), which examines how voters in Florida use information about experience and gender on whether they select female judicial candidates. In this experiment, Mo (2015) is interested in understanding the effect of gendered bias on candidate selection. The experiment leverages non-partisan judicial races where voters are told the candidate's level of experience, gender, and extraneous information, and are then asked to select a candidate.

The data includes 390 Florida residents surveyed from September to October 2008. This experiment functioned similarly to a conjoint, with profiles created for two judicial candidates. While there were five features with several attributes, the primary attributes studied here were candidate gender

[11]This requires the assumption that there are no differences within each of the three divisions of the survey provider's respondent pool that would alter who accepts a survey invitation. Because the survey pool was itself randomly divided to receive different sets of invitations, any differences respondents have in accepting invitations should be randomly distributed between the three placement groups.

[12]As we discuss in the conclusion, we believe the high rate of attrition suggests an important drawback to assuming that a two-wave sample is necessarily an optimal way to measure sensitive questions.

and candidate experience—signaled by the American Bar Association rating, which scored the candidate as either "strong" or "weak." The outcome of interest is whether respondents selected the female candidate. Here, the sensitive items are symbolic sexism, which was presented in the same wave and is composed of 14 questions that ask about attitudes toward women.

6. Results

6.1. Hypothesis 1

First, we test Hypothesis 1. For all tests, we transform variables to a 0–1 scale so results can be read as $\beta * 100$ percentage-point changes. Figure 2 displays these estimates.[13] Across all four studies, the measurement of the sensitive items does not statistically significantly change as a result of the placement of those items—either pre- or post-treatment ($p_1 = 0.488; p_2 = 0.084; p_{3a} = 0.060; p_{3b} = 0.67; p_4 = 0.22$). Importantly, for Studies 2 and 3, which feature a second wave where sensitive items were measured 1-2 weeks prior to the experiment, there are no significant differences between the measurement of the sensitive items in a separate wave relative to pre-treatment ($p_2 = 0.580; p_{3a} = 0.760; p_{3b} = 0.405$) nor between the separate wave and post-treatment ($p_2 = 0.280; p_{3a} = 0.126; p_{3b} = 0.692$). Placing the sensitive items before or after treatment does not result in a statistically significantly different measure of the sensitive items. Moreover, placing the sensitive item in a separate wave before the treatment does not result in a statistically significant difference in the measurement of those sensitive items. Thus, we fail to reject the null for Hypothesis 1.

6.2. Hypothesis 2

Next, we test Hypothesis 2. Figure 3 displays the differences between the estimates of the substantive dependent variable for the experiment of each study by the placement of the sensitive items. Note that in Study 3, the authors also use two different outcome variables for both Experiment 1 and Experiment 4.[14] Finally, while three of the studies can be read as means comparisons across the different treatments, in Study 2, the experimental outcome is read differently. The conjoint experiment tests which attributes affect the decision to grant Green Cards to one of two immigrant profiles. As a result, to interpret the experimental outcomes, we look for differences from 0.5—which means that an attribute significantly affects a respondent's decision-making. We consider attributes scoring above 0.5 as favorable attributes, and those scoring below 0.5 as unfavorable attributes. For Hypothesis 2 only, we vary our experimental test here because this is a conjoint experiment. If we use a forced-choice variable of Green Card choice for Y, we will always estimate 0.5 because there is a 0 coded for the profile not chosen, and a 1 coded for the profile chosen. Since our goal is to determine whether the placement of the sensitive variable causes a difference in the estimate of the experimental outcome regardless of treatments, we use a different variable. Drawing from d'Urso and Bonilla (2023), we use cultural assimilation—how likely the immigrant is perceived to be able to assimilate into American culture. We believe this is a conservative test, since this secondary dependent variable is likely to be influenced by attitudes toward sensitive measures in other studies (e.g., Lajevardi, 2020).[15]

Across all four studies, the dependent variable of the experiment does not significantly change based on whether the sensitive items were introduced pre-treatment or post-treatment ($p_1 = 0.094; p_2 = 0.601; p_{3a-health} = 0.755; p_{3a-leader} = 0.939; p_{3b-health} = 0.116; p_{3b-leader} = 0.506; p_4 = 0.823$). Asking the sensitive items in a separate wave also does not shift results relative to pre-treatment placement ($p_2 = 0.583; p_{3a-health} = 0.804; p_{3a-leader} = 0.876; p_{3b-health} = 0.942;$

[13] We provide the full linear models in the Appendix (Tables B.4, C.12, D.25, and E.36).

[14] We display the full linear models in the Appendix (Study 1: Table B.4, Column 2; Study 2: Table C.12, Column 2; Study 3: Table D.26; Study 4: Table E.36, Column 2).

[15] Although we only use this dependent variable for this study for Hypothesis 2, we have included the full linear models for Hypotheses 3 and 4 using this dependent variable in Tables C.16, C.17, C.21, and C.22.

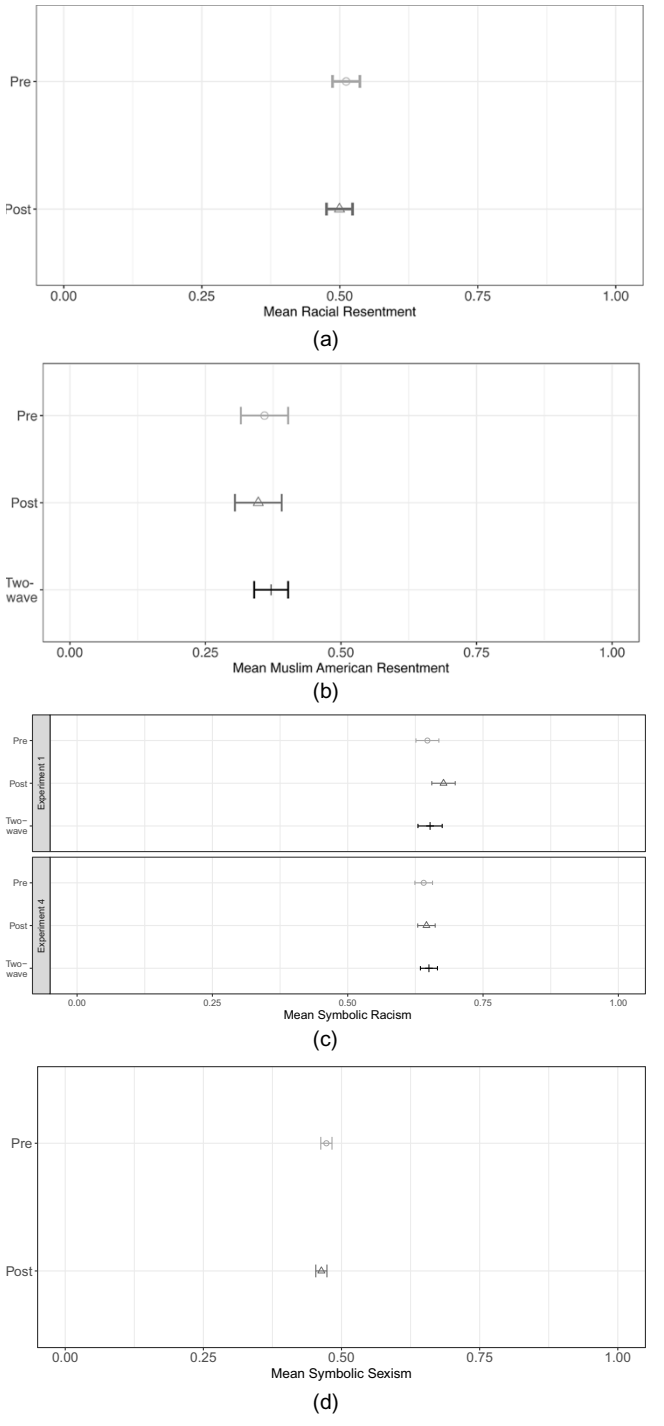

Figure 2. Effect of placement of sensitive items on the measurement of those sensitive items. a) Study 1. b) Study 2. c) Study 3. d) Study 4.

Notes: Each figure reports the regression coefficient and the 95% confidence interval associated with that test. For all studies, the x-axis indicates the placement of the sensitive items (*T*) and the y-axis indicates the estimate of the sensitive item (*S*). For Study 3, please note the two experiments presented, indicated by different colored and shaped points. Comparisons should be made across corresponding colored and shaped points.

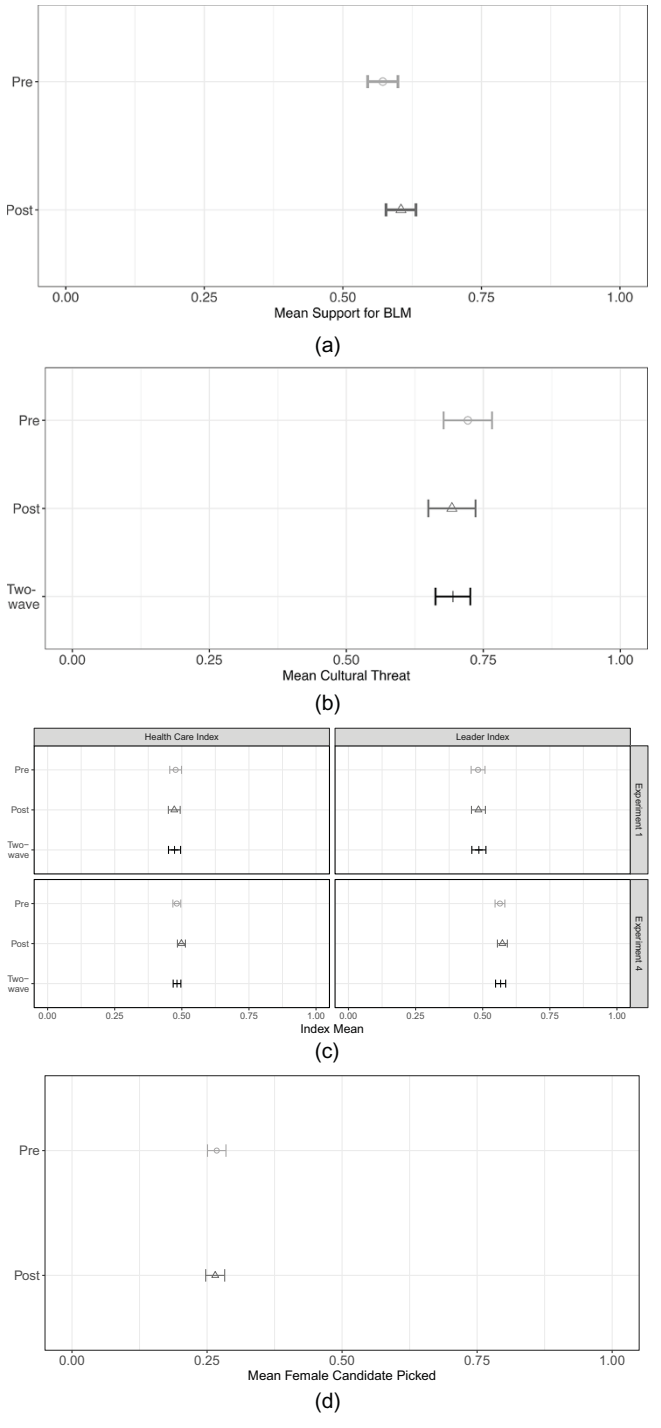

Figure 3. Effect of placement of sensitive items on the measurement of the study's dependent variable. a) Study 1. b) Study 2. c) Study 3. d) Study 4.

Notes: Each figure reports the regression coefficient and the 95% confidence interval associated with that test. For all studies, the x-axis indicates the placement of the sensitive items (*T*) and the y-axis indicates the experimental outcome (*Y*). For Study 3, please note the two experiments presented indicated by different colored points; here, the shapes represent the different dependent variables. Comparisons should be made across corresponding colored and shaped points.

$p_{3b-leader} = 0.855$) nor relative to post-treatment placement ($p_2 = 0.283$; $p_{3a-health} = 0.953$; $p_{3a-leader} = 0.936$; $p_{3b-health} = 0.128$; $p_{3-b-leader} = 0.634$). We conclude that the measurement of the outcome variables does not significantly differ based on where the sensitive items are measured in the survey. As a result, we fail to reject the null of Hypothesis 2; the placement of sensitive items relative to the treatment will not statistically significantly alter the measurement of the dependent variable.

6.3. Hypothesis 3

Third, we test Hypothesis 3 and display the results in Figure 4.[16] Figure 4a presents the results for Study 1. We compare the two different treatments—"Nationalist" and "Feminist," and we find no significant differences in the experimental outcome based on placement in either the control ($p_{Nationalist} = 0.134$) or treatment ($p_{Feminist} = 0.331$).

In Study 2, the experimental treatment is the conjoint task displayed in Figure 4b. The x-axis presents the marginal mean, while the y-axis includes each conjoint attribute and level. We find no significant differences in the measurement of the experimental outcome for religious or racial characteristics across the placement of MAR (all $p > 0.05$). However, we do interpret the results of the experiment differently based on the placement of MAR—where the outcome is statistically distinguishable from 0.5 in some placement conditions but not others. Regardless of whether MAR was placed pre-treatment, post-treatment, or in a wave prior to the experiment, we see no differences in favorability based on the immigrants' religion based on the placement of MAR. On the race attributes, respondents only tended to favor immigrants who are Middle Eastern less ($\bar{x} = 0.464$; $p = 0.025$) if MAR was placed pre-treatment but not if MAR is placed post-treatment ($\bar{x} = 0.484$; $p = 0.213$) or in a separate wave ($\bar{x} = 0.475$; $p = 0.080$). On the other hand, only when MAR is placed post-treatment do respondents view South Asian immigrants favorably ($\bar{x} = 0.543$; $p = 0.015$) but not when MAR is measured pre-treatment ($\bar{x} = 0.497$; $p = 0.877$) or in the two-wave ($\bar{x} = 0.516$; $p = 0.389$). Overall, this study yields some evidence that the placement of the sensitive measure may interact with treatment conditions to yield different experimental outcomes.[17]

Study 3 includes two different experimental outcomes: measuring respondents' attitudes toward health care policies and social welfare policies, as well as evaluating various political leaders' stances on health care policies in Figure 4c. Across the relevant tests, we find no statistically significant differences in the placement of the sensitive item (symbolic racism) on the experimental outcomes (health care or leader index) based on the experimental treatment (implicit or explicit prime).[18]

We present the results of Study 4 in Figure 4d. We find no significant differences between the placement of the sensitive items and the estimates of the dependent within the treatment conditions: a strong female candidate ($p = 0.158$), a strong male candidate ($p = 0.133$), or both strong candidates ($p = 0.954$). Respondents are more likely to select the female candidate when only the female candidate is strong, but are less likely to select the female candidate if both candidates are strong. And, these experimental conclusions remain the same across the timing of the measurement of sensitive questions. We see no significant differences in the experimental outcome, nor do we draw different conclusions from the experiment based on where symbolic sexism is administered.

Across all four studies, we see no statistically significant interaction effects between the placement of the sensitive measure and the experimental treatments themselves, and the placement causes no shifts in the outcome of the experimental conclusions. In Study 2, we see some evidence that may suggest an interaction effect between the placement of MAR and the immigrant's race on whether

[16] All linear models and individual difference of means tests can be found in the Appendix (Tables B.5–B.6, C.13–C.15, D.27–28, and E.37–E.38.

[17] This experiment also includes control attributes including education, gender, and English fluency, which we present the results with controls in Appendix Figure C.2.

[18] See Appendix Tables D.27 and D.28.

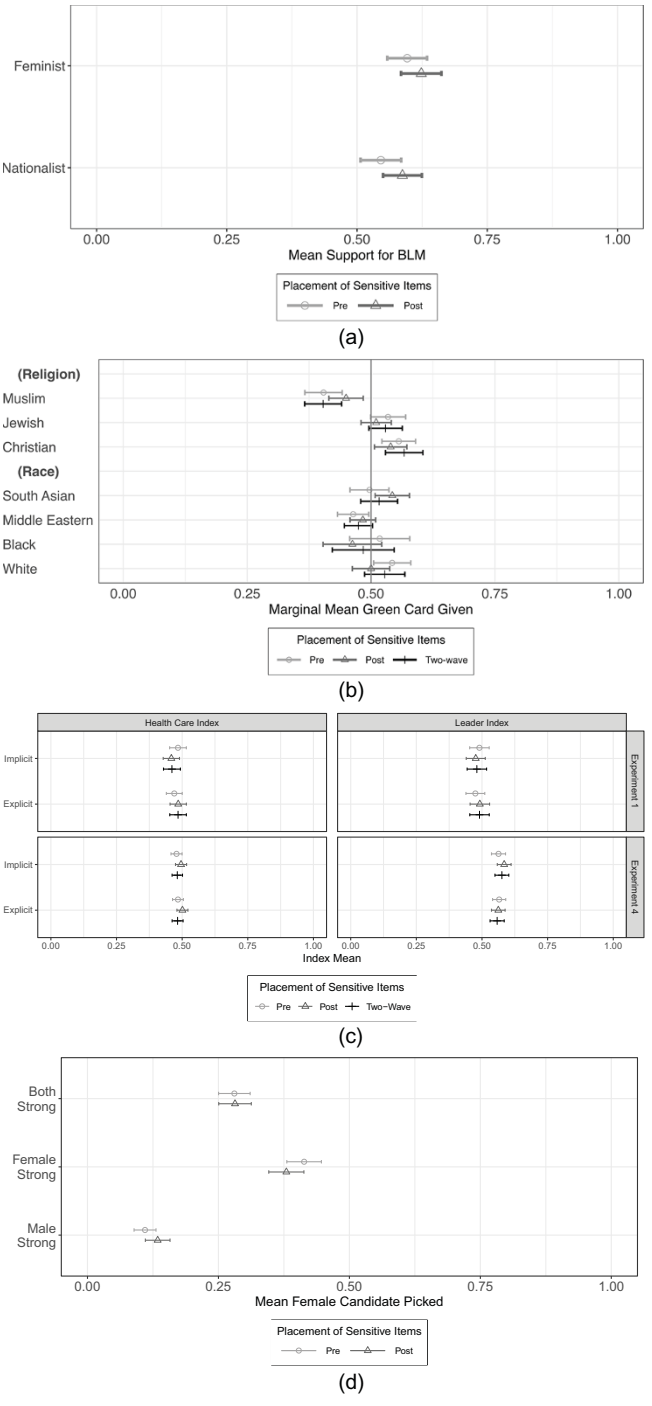

Figure 4. Interaction of placement of sensitive items and experimental treatment on experimental outcome. a) Study 1. b) Study 2. c) Study 3. d) Study 4.

Notes: Each figure reports the regression coefficient and the 95% confidence interval associated with that test. For Studies 1, 3, and 4, the x-axis indicates the placement of the sensitive items (*T*) and the y-axis indicates the experimental outcome (*Y*). Studies 1 and 4 distinguish between experimental treatment (*D*) by color and shape. Study 3 indicates the experiment by color, treatment condition (*D*) by shape, and experimental outcome (*Y*) in separate subplots. Comparisons should be made between the corresponding color and shape across the placement of the sensitive item (*T*). Study 2 presents the marginal mean of the experimental outcome (*Y*) on the x-axis instead of the y-axis. The y-axis provides the conjoint features, which can be thought of as experimental treatments (*D*). The placement of the sensitive items (*T*) is indicated by color and shape.

they were favored for a green card, but no interaction related to the immigrant's religion. Overall, we conclude that the placement of the sensitive items does not appear to interact with the experimental treatments to alter the measurement of the experimental outcome. Therefore, we fail to reject the null for Hypothesis 3 in most of the studies.

6.4. Hypothesis 4

Lastly, we test Hypothesis 4. For ease of interpretation with a three-way interaction effect, we create a dichotomous measurement of our sensitive items for those who score high (≥ 0.5) on the left and those who score low (< 0.5) on the right. Figure 5 displays the results.[19]

In Study 1, we find that racial resentment itself interacts with the treatment on support for the goals of BLM. Those with higher levels of racial resentment are less likely to support the goals of BLM. In certain instances, the placement of racial resentment and the measured level of racial resentment interact with the experimental treatment to produce different experimental conclusions. In two instances, researchers may draw different conclusions when looking at heterogeneity by the sensitive items based on where those items were placed. For those scoring low on racial resentment, there is a statistically significant difference in how they rate the goals of BLM under the Black Feminist framing ($\bar{x} = 0.776$) compared to the Black Nationalist framing ($\bar{x} = 0.707, p < 0.05$) when racial resentment is measured post-treatment. However, there is no difference when racial resentment was measured pre-treatment. In contrast, for those scoring high on racial resentment, there is a statistically significant difference in how they rate the goals of BLM under Black Feminist framing ($\bar{x} = 0.491$) relative to the Black Nationalist framing ($\bar{x} = 0.408, p < 0.05$) when racial resentment is measured pre-treatment. Again, there is no difference when racial resentment was measured post-treatment. Therefore, it may be important to consider that respondents' level of racial resentment may influence how they interact with the experimental treatment, conditional upon when racial resentment was administered in the experiment.

For Study 2, Figure 5b presents the results. Among those who score low on the MAR scale, there are no differences in pre-treatment, post-treatment, or two-wave administration of MAR on the measurement of the outcome based on the different immigrant characteristics (religion and race). However, we do see some differences in the experimental conclusions researchers make based on the placement of MAR. Those who score low on MAR and receive MAR pre-treatment are less likely to award Green Cards to immigrants identifying as Christian ($\bar{x} = 0.542; p = 0.038$), Muslim ($\bar{x} = 0.450; p = 0.015$), or Middle Eastern ($\bar{x} = 0.465; p = 0.060$). However, these differences are not significant for post-treatment or two-wave treatment. Those who scored low on MAR also rate the South Asian immigrant more favorably when MAR was measured post-treatment ($\bar{x} = 0.548; p = 0.023$) but not when measured pre-treatment or in a separate wave. As a result, some evidence suggests that the administration of MAR in the pre-treatment may have primed individuals who scored lower on the scale to evaluate Muslim and Middle Eastern migrants less favorably.

Further differences occur among those who score high on the MAR scale. Moreover, respondents high on MAR rate Christian immigrants ($\delta\bar{x} = 0.092; p = 0.032$) higher and Jewish immigrants ($\delta\bar{x} = -0.110; p = 0.020$) lower if MAR placement occurs in the two-wave condition over pre-treatment. Christian immigrants are rated higher ($\delta\bar{x} = 0.110; p = 0.004$) and Muslim immigrants lower ($\delta\bar{x} = -0.091; p = 0.039$) when MAR is asked in a separate wave than post-treatment. Respondents rate Muslim immigrants more favorably when MAR is measured in post-treatment compared to pre-treatment ($\delta\bar{x} = 0.107; p = 0.029$). This could suggest that asking MAR pre-treatment may have influenced evaluations of Muslim immigrants. Next, we consider whether the experimental outcome differed based on the placement of MAR among those who scored high on the scale of items (i.e., distinguishable from 0.5). Across all three placements, we see that those who were high on MAR

[19] All linear models and individual difference of means tests can be found in Appendix Tables B.7–B.8, C.18–C.20, D.29–D.30, and E.39–E.40.

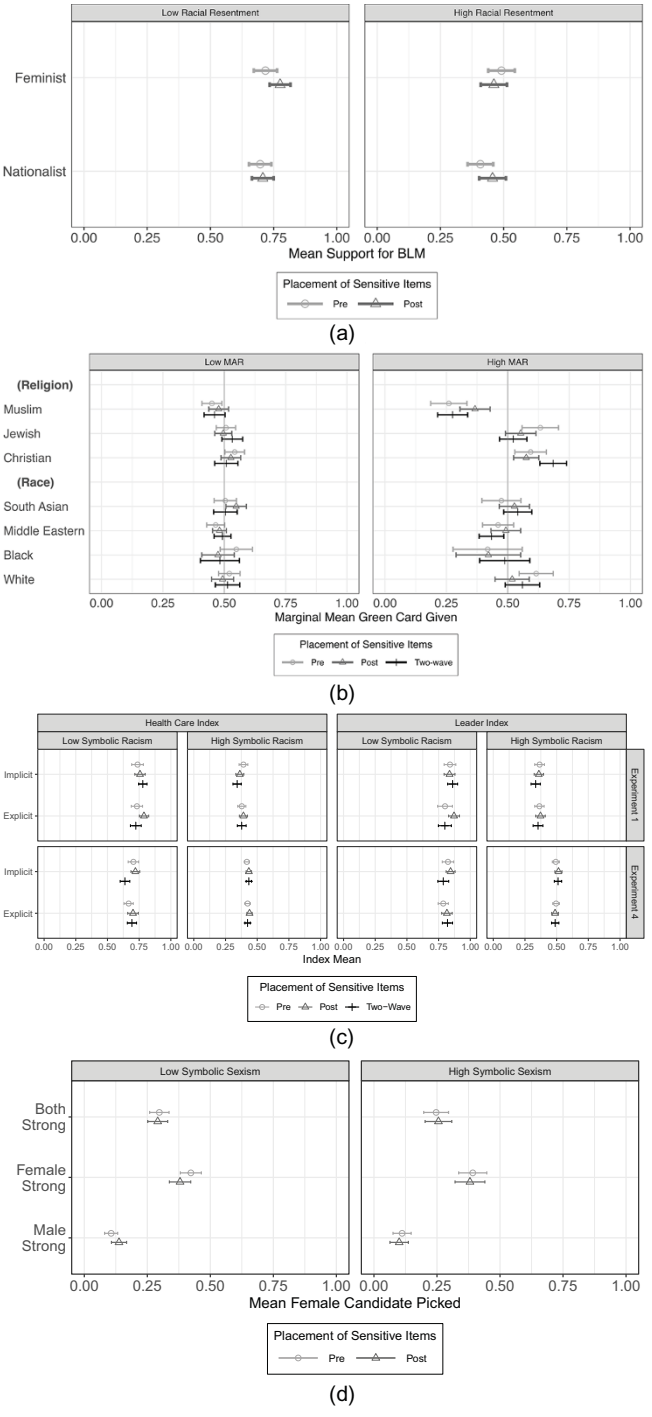

Figure 5. Interaction of placement of sensitive items, measurement of sensitive items, and experimental treatment on the experimental outcome. a) Study 1. b) Study 2. c) Study 3. d) Study 4.

evaluated the Christian immigrants more favorably ($p < 0.05$ for all) and Muslim immigrants less favorably ($p < 0.05$ for all) relative to the baseline of 0.5. However, among the other attributes, there were some differences based on the placement of MAR. Those who received MAR pre-treatment rated Jewish immigrants ($\bar{x} = 0.620$; $p = 0.002$) and White immigrants ($\bar{x} = 0.606$; $p = 0.005$) favorably, while those who received MAR in a separate wave prior to the experiment rated Middle Eastern immigrants ($\bar{x} = 0.434$; $p = 0.010$) unfavorably.[20] For those low on MAR, when MAR is asked pre-treatment, Muslims are evaluated less favorably; however, this experimental conclusion did not emerge when MAR was measured post-treatment or in a second wave prior to the experiment. On the other hand, respondent attitudes appear more stable among those scoring high on MAR; their evaluations of Muslims remained consistently unfavorable regardless of where MAR was placed. This suggests some stability among those who score higher on the scale than those scoring low on MAR, as these items are specifically aimed at measuring prejudice toward Muslims.

In Study 3 (Figure 5c), we find some evidence of statistically significant differences ($p < 0.05$) in the placement of the sensitive item on the experimental outcomes based on the experimental treatment. However, these differences occur only among those low on symbolic racism. First, we see no statistically significant differences in outcomes when comparing pre- and post-treatment measurement of symbolic racism in either experimental study. Only one statistically significant difference occurs when comparing two-wave versus pre-treatment administration. Those in the implicit priming condition in Experiment 4 rated the health care policies *higher* when symbolic racism was measured pre-treatment ($\bar{x} = 0.705$) relative to two-wave ($\bar{x} = 0.638$, $p < 0.05$). However, this difference was not statistically significant for the leader outcome. There were also no other differences in Experiment 1, nor for any explicit treatments. The biggest differences we see are between the two-wave and post-treatment measurements. In Experiment 1, those in the explicit racial prime condition rated both outcome variables statistically significantly higher when symbolic racism was measured post-treatment relative to a separate wave. In Experiment 4, those in the implicit racial prime condition rated both outcome variables higher when symbolic racism was measured post-treatment relative to a separate wave. Finally, while there are differences in the mean outcomes conditional on the placement of the sensitive items within experimental treatments, the overall experimental conclusions do not change based on when sensitive items are measured. In Experiment 4, there is a significant difference between explicit ($\bar{x} = 0.693$) and implicit ($\bar{x} = 0.638$, $p < 0.05$) primes on the health care outcome when symbolic racism is measured in a separate wave. However, all other comparisons between explicit and implicit racial primes across, holding the placement of symbolic racism constant, are not statistically significant for any outcome measure in either experimental study. Thus, among those who are lower on symbolic racism, we see some evidence to support the concern that the experimental treatment itself could influence the measurement of the moderator. We do not see enough evidence to support the concern of priming, as pre-treatment measurement and two-wave measurement are largely not statistically distinguishable.

Next, we consider those who are high on symbolic racism. Across all comparisons, there is only one instance where there is a statistically significant difference between the point estimate as a result of the placement. In Experiment 1 in the implicit priming condition, those who were higher on symbolic racism rated the health care policies *higher* when symbolic racism was administered pre-treatment ($\bar{x} = 0.392$) relative to in a separate wave ($\bar{x} = 0.342$, $p < 0.05$).

Finally, we analyze Study 4 presented in Figure 5d. This means that across all treatments, respondents selected strong female candidates more frequently than strong male candidates in the condition with both strong candidates. The experimental conclusions hold across the timing of the measurement of the experimental treatment and between moderated treatment groups. Therefore, we find no evidence in Study 4 of a three-way interaction.

[20] We include p-value adjustments because conjoint analyses involve multiple comparisons, which increases the likelihood of a false positive (Liu and Shiraito, 2023) in Appendix Table C.20. When accounting for this adjustment, none of the statistically significant findings remain among those who scored low on MAR. Among those who scored high on MAR, only the difference between evaluation of Christians when MAR is asked in two waves relative to post-treatment remains statistically significant.

Across all hypotheses, we fail to reject the null in most cases that the *average* causal effect of placement of sensitive items is zero, though we cannot do that across all studies. To test the reliability of our null findings, we present a formal equivalence test in Appendix Tables B.9, C.23, D.31, and E.41 to assess whether the data are equivalent across our analyses of the four hypotheses (Hartman and Hidalgo, 2018). Ultimately, we detect a small to medium effect size for analyses involving null results.

7. Conclusion

There is no clear guidance on when to measure prejudice in an experimental setting. We empirically test two competing arguments about experimental design involving where to place questions that are vulnerable to spillover because they are sensitive items. Across these four studies, we find that the placement of sensitive items has no significant effect on the estimate of the sensitive items (e.g., racial resentment/symbolic racism, MAR, or symbolic sexism), even when the sensitive item was measured on a separate question wave, a major concern of Montgomery *et al.* (2018). We also find no significant effects on the experimental outcomes of the experiment, addressing the concerns of Klar *et al.* (2020). Finally, we find only inconsistent evidence that the placement of sensitive items affects the experimental outcome conditional on the type of experimental treatment. However, even in these cases, these differences rarely changed the overall conclusion of the experiment.

These results yield instructive criteria for how we may weigh conflicting concerns, particularly for scholars working on topics involving prejudice toward marginalized groups—including racial and ethnic prejudice, Islamophobia, and sexism. As scholarship has expanded to consider other forms of prejudice, these findings are relevant to studies of classism (ambivalent classism), ableism (symbolic ableism), ageism (ambivalent ageism), and beyond. On one hand, we do not find evidence that the order of the moderator on average changes the measurement of the moderator itself. On the other hand, we do not find consistent evidence that the order of the moderator affects the measurement of the experimental outcome or the conclusions of the experiment itself. As a result, we do recommend that, absent additional evidence, researchers should follow Montgomery *et al.* (2018) and measure moderators prior to the experiment, but we wish to make two critical caveats.

First, we want to address the assumption that measuring sensitive items in a two-wave format necessarily avoids the issues addressed by Montgomery *et al.* (2018) and Klar *et al.* (2020). In Study 2, where we incorporated a two-wave placement treatment, nearly one-third of the treatment arm discontinued the study between waves. Attrition invariably means studies cost more as researchers must pay not just for two surveys over one, but it also means they pay for respondents who do not complete the study. Perhaps more importantly, it may be worth investigating whether attrition is not evenly spread across the sensitive measure, thereby affecting randomization. If key differences are found by demographic, block randomization may help alleviate these differences.[21] Future research should seek to understand the role of attrition in these scenarios. Given our work to divide the survey firm's sampling pool in designing this study, we believe this is possible, however, this is beyond the scope of this paper.

Second, our findings do not mean that we should discount the concerns raised by Klar (2013) and research that is concerned about placing moderators pre-treatment. Nor should researchers discount previously published work on prejudice toward marginalized groups solely due to where sensitive questions are administered in experiments. While we do not find consistent evidence that moderating items may prime respondents to think differently about treatments and affect experimental conclusions, we also acknowledge that, at most, we can only fail to reject the null hypothesis in most cases. Moreover, when evaluating how respondents interact with the treatment conditional on the placement of these sensitive items, we find more stability in null effects among those who fall higher on these measures of prejudice. This suggests that researchers may wish to be especially careful if they

[21] This may be useful if there are treatment differences across groups as well, even when measuring moderators pre-treatment. See, for instance, Bonilla and Mo (2018).

are hypothesizing how those who are lower on these scales might interact with the experimental treatment based on when these items are administered. Even in the best of circumstances, experiments are always subject to a bad draw, and all researchers should check for balance between their treatment groups (Mutz *et al.*, 2019). We show that most of these measures are robust to pre-treatment placement in most scenarios. Thus, researchers may not need to be especially concerned when placing measures of prejudice pre-treatment. If researchers choose to measure sensitive moderators post-treatment, they should transparently discuss how this choice and the anticipated trade-offs associated therein. Ultimately, the placement of sensitive items is a design choice that should be made thoughtfully and transparently, and researchers must directly acknowledge that where they choose to place these items may each lead to specific and different potential biases.

Supplementary material. The supplementary material for this article can be found at https://10.1017/psrm.2025.10018. To obtain replication material for this article, https://doi.org/10.7910/DVN/D8VODI.

Funding. This project was funded by the Center for the Study of Diversity and Democracy at Northwestern University, the Institute for Policy Research, and the School of Education and Social Policy at Northwestern University. Special thanks to S.R. Gubitz for his work in the early stages of the project. Jamie Druckman made space for this project to happen in his lab at Northwestern University; we thank Jamie and the lab participants who gave feedback to this project at different stages. Similarly, Al Tillery (through the CSDD) ensured this work would receive funding, for which we remain grateful. We would like to thank Fabian Neuner who directed us toward data for Study 3, and Cecilia Mo who shared her study data with us for Study 4. Finally, we would like to thank Brendan Nyhan and Jacob Montgomery for reading drafts of this paper and offering feedback.

References

Albertson B and Jessee S (2023) Moderator placement in survey experiments: Racial resentment and the "welfare" versus "assistance to the poor" question wording experiment. *Journal of Experimental Political Science* **10**, 448–454.

Aziz SF (2021) *The Racial Muslim: When Racism Quashes Religious Freedom*. Oakland, California: University of California Press.

Baker A (2015) Race, paternalism, and foreign aid: Evidence from US public opinion. *American Political Science Review* **109**, 93–109.

Banda KK and Cassese EC (2022) Hostile sexism, racial resentment, and political mobilization. *Political Behavior* **44**, 1317–1335.

Bansak K, Hainmueller J, Hopkins DJ, and Yamamoto T (2021) Conjoint Survey Experiments. In Druckman, JN, and Green, DP (eds), *Advances in Experimental Political Science*. Cambridge, United Kingdom: Cambridge University Press.

Benjamin DJ, Choi JJ and Joshua Strickland A (2010) Social identity and preferences. *American Economic Review* **100**, 1913–1928.

Benokraitis NV and Feagin JR (1995) *Modern Sexism: Blatant, Subtle, and Covert Discrimination*. New Jersey: Prentice Hall.

Berinsky AJ, Huber GA and Lenz GS (2012) Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* **20**, 351–368.

Beydoun KA (2013) Between Muslim and white: The legal construction of Arab American identity. *NYU Ann. Surv. Am. L.* **69**, 29.

Blair G (2015) Survey methods for sensitive topics. *Comparative Politics Newsletter* **12**, 44.

Blair G, Coppock A and Moor M (2020) When to worry about sensitivity bias: A social reference theory and evidence from 30 years of list experiments. *American Political Science Review* **114**, 1297–1315.

Bonilla T and Cecilia Hyunjung M (2018) Bridging the partisan divide on immigration policy attitudes through a bipartisan issue area: The case of human trafficking. *Journal of Experimental Political Science* **5**, 107–120.

Bonilla T and Tillery AB (2020) Which identity frames boost support for and mobilization in the #BlackLivesMatter movement? An experimental test. *American Political Science Review* **0**, 1–16.

Buyuker B, Jadidi D'Urso A, Filindra A and Kaplan NJ (2021) Race politics research and the American presidency: Thinking about white attitudes, identities and vote choice in the Trump era and beyond. *Journal of Race, Ethnicity, and Politics* **6**, 600–641.

Cassino D and Erisen C (2010) Priming Bush and Iraq in 2008: A survey experiment. *American Politics Research* **38**, 372–394.

Cecilia Hyunjung M (2015) The consequences of explicit and implicit gender attitudes and candidate quality in the calculations of voters. *Political behavior* **37**, 357–395.

Chong D and Junn J (2011) Politics from the perspective of minority populations. In Druckman, JN, Green, DP, Kuklinski, JH, Lupia, Arthur (eds), *Cambridge Handbook of Experimental Political science*. Cambridge, United Kingdom: Cambridge University Press, 320–335.

Clayton K, Ferwerda J and Horiuchi Y (2021) Exposure to immigration and admission preferences: Evidence from France. *Political Behavior* **43**, 175–200.
Coppock A (2019) Avoiding post-treatment bias in audit experiments. *Journal of Experimental Political Science* **6**, 1–4.
Coppock A, and McClellan OA (2019) Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics* **6**(1), 1–14.
Cuevas-Molina I (2023) White racial identity, racial attitudes, and Latino partisanship. *Journal of Race, Ethnicity, and Politics* **8**, 469–491.
Denney S and Green C (2021) Who should be admitted? Conjoint analysis of South Korean attitudes toward immigrants. *Ethnicities* **21**, 120–145.
DeSante CD and Watts Smith C (2020) Less is more: A cross-generational analysis of the nature and role of racial attitudes in the twenty-first century. *The Journal of politics* **82**, 967–980.
d'Urso AS (2024) A boundary of white inclusion: The role of religion in Ethnoracial assignment. *Perspectives on Politics* **22**, 559–576.
d'Urso AS, and Bonilla T (2023) Religion or Race? Using intersectionality to examine the role of Muslim identity and evaluations on belonging in the United States. *Journal of Race, Ethnicity, and Politics* **8**(2), 1–21.
Entman RM, and Rojecki A (2001) *The Black Image in the White Mind: Media and Race in America*. Chicago, Illinois: University of Chicago Press.
Fiske ST *et al.* (1993) Social cognition and social perception. *Annual Review of Psychology* **44**, 155–194.
Gothreau C, Arceneaux K and Friesen A (2022) Hostile, benevolent, implicit: How different shades of sexism impact gendered policy attitudes. *Frontiers in Political Science* **4**, 817309.
Hartman E and Daniel Hidalgo F (2018) An equivalence approach to balance and placebo tests. *American Journal of Political Science* **62**, 1000–1013.
Hassell HJG and Visalvanich N (2015) Call to (in)action: The effects of racial priming on grassroots mobilization. *Political Behavior* **37**, 911–932.
Huber GA and Lapinski JS (2006) The "race card" revisited: Assessing racial priming in policy contests. *American Journal of Political Science* **50**, 421–440.
Hussey I, and De Houwer J (2018) Completing a Race IAT increases implicit racial bias. https://osf.io/preprints/psyarxiv/vxsj7.
Hutchings VL and Jardina AE (2009) Experiments on racial priming in political campaigns. *Annual Review of Political Science* **12**, 397–402.
Iyengar S (2008) Priming theory. *The International Encyclopedia of Communication* **9**, 3883–3886.
Iyengar S and Kinder DR (2010) *News That matters: Television and American Opinion*. Chicago, Illinois: University of Chicago Press.
Jackson MS (2011) Priming the sleeping giant: The dynamics of Latino political identity and vote choice. *Political Psychology* **32**, 691–716.
Jardina A (2021) In-group love and out-group hate: White racial attitudes in contemporary US elections. *Political Behavior* **43**, 1535–1559.
Kam CD and Burge CD (2019) TRENDS: Racial resentment and public opinion across the racial divide. *Political Research Quarterly* **72**, 767–784.
Kinder DR and Sears DO (1981) Prejudice and politics: Symbolic racism versus racial threats to the good life. *Journal of Personality and Social Psychology* **40**, 414.
King G and Zeng L (2006) The dangers of extreme counterfactuals. *Political Analysis* **14**, 131–159.
Klar S (2013) The influence of competing identity primes on political preferences. *The Journal of Politics* **75**, 1108–1124.
Klar S, Leeper T and Robison J (2020) Studying identities with experiments: Weighing the risk of posttreatment bias against priming effects. *Journal of Experimental Political Science* **7**, 56–60.
Kreuter F, Presser S and Tourangeau R (2008) Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly* **72**, 847–865.
Lajevardi N (2020) *Outsiders at Home: The Politics of American Islamophobia*. Cambridge, United Kingdom: Cambridge University Press.
Lajevardi N and Abrajano M (2019) How negative sentiment toward Muslim Americans predicts support for Trump in the 2016 presidential election. *The Journal of Politics* **81**, 296–302.
Lajevardi N and Oskooii KAR (2018) Old-fashioned racism, contemporary islamophobia, and the isolation of Muslim Americans in the age of Trump. *Journal of Race, Ethnicity, and Politics* **3**, 112–152.
Lehrer R, Juhl S and Gschwend T (2019) The wisdom of crowds design for sensitive survey questions. *Electoral Studies* **57**, 99–109.
Liu G and Shiraito Y (2023) Multiple hypothesis testing in conjoint analysis. *Political Analysis* **31**, 380–395.
McConnaughy CM, White IK, Leal DL and Casellas JP (2010) A Latino on the ballot: Explaining coethnic voting among Latinos and the response of White Americans. *The Journal of Politics* **72**, 1199–1211.

Mendelberg T (2001) *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. NJ: Princeton University Press.
Mendelberg T (2008) Racial priming revived. *Perspectives on Politics* **6**, 109–123.
Mo CH and Bonilla T (2020) Predicting biased behavior with implicit attitudes. In Krosnick JA, Stark TH and Scott AL (eds), *The Cambridge Handbook of Implicit Bias and Racism*. Cambridge University Press, Cambridge.
Montgomery JM, Nyhan B and Torres M (2018) How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science* **62**, 760–775.
Mutz DC, Pemantle R and Pham P (2019) The perils of balance testing in experimental design: Messy analyses of clean data. *The American Statistician* **73**, 32–42.
Näher A-F and Krumpal I (2012) Asking sensitive questions: The impact of forgiving wording and question context on social desirability bias. *Quality & Quantity* **46**, 1601–1616.
Ostfeld M and Pedraza F (2015) Ethnic filter questions: The political implications of priming Latino identity Interdisciplinary Workshops on Politics and Policy University of Michigan, May 20, 2015.
Pingree S, Parker Hawkins R, Butler M and Paisley W (1976) A scale for sexism. *Journal of Communication* **26**, 193–200.
Rasinski KA, Willis GB, Baldwin AK, Yeh W and Lee L (1999) Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* **13**, 465–484.
Rathje S, Van Bavel JJ and Van Der Linden S (2021) Out-group animosity drives engagement on social media *Proceedings of the National Academy of Sciences*. **118**(26), 1–9.
Robison J, Stevenson RT, Druckman JN, Jackman S, Katz JN and Vavreck L (2018) An audit of political behavior research. *Sage Open* **8**, 2158244018794769.
Schiff KJ, Pablo Montagnes B and Peskowitz Z (2022) Priming self-reported partisanship: Implications for survey design and analysis. *Public Opinion Quarterly* **86**, 643–667.
Sheagley G and Clifford S (2025) No evidence that measuring moderators alters treatment effects. *American Journal of Political Science* **69**(1), 49–63.
Smith CW, Kreitzer RJ and Suo F (2020) The dynamics of racial resentment across the 50 US states. *Perspectives on Politics* **18**, 527–538.
Steele CM (2011) *Whistling Vivaldi: And Other Clues to How Stereotypes Affect Us (Issues of Our Time)*. New York, New York: WW Norton & Company.
Tourangeau R, Rips LJ, and Rasinski K (2000) *The psychology of survey response.* Cambridge, United Kingdom: Cambridge University Press.
Tourangeau R and Smith TW (1996) Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly* **60**, 275–304.
Tourangeau R and Yan T (2007) Sensitive questions in surveys. *Psychological Bulletin* **133**, 859.
Valentino NA, Hutchings VL and White IK (2002) Cues that matter: How political ads prime racial attitudes during campaigns. *American Political Science Review* **96**, 75–90.
Valentino NA, Neuner FG and Matthew Vandenbroek L (2018) The changing norms of racial political rhetoric and the end of racial priming. *The Journal of Politics* **80**(3), 757–771.
Valenzuela AA and Reny T (2021) Evolution of experiments on racial priming. In Druckman, JN, and Green, DP (eds), *Advances in Experimental Political Science*. Cambridge, United Kingdom: Cambridge University Press, 447–467.
Williams MT, Turkheimer E, Magee E and Guterbock T (2008) The effects of race and racial priming on self-report of contamination anxiety. *Personality and Individual Differences* **44**, 746–757.

Cite this article: d'Urso AS, Bonilla T and Bogdanowicz G (2025) Where to place sensitive questions? Experiments on survey response order and measures of discriminatory attitudes. *Political Science Research and Methods*, 1–22. https://doi.org/10.1017/psrm.2025.10017