The Perils of Self-Assessed Attitude Change

Matthew Graham and Alexander Coppock*

April 1, 2019

Prepared for the Annual Meeting of the Midwest Political Science Association, Chicago, April 4, 2019

Abstract

Surveys often ask respondents to self-report how events or information changed their attitudes. Does [event X] make you more or less likely to vote for a politician? How did [information X] affect your opinion? Would you be more supportive of a candidate who took [position X]? We show that the self-assesment question type exhibits poor measurement properties. Using 18 mini-experiments across three studies, we compare this question type—and eight alternative ways of asking subjects to assess counterfactuals—to randomized experiments. When asked to report how their attitudes change, subjects appear to frequently overestimate the magnitude of treatment effects and sometimes get the sign wrong. Our results are broadly consistent with the notion that respondents misreport their attitude change because they engage in response substitution. Self-reports of attitude *change* appear to be infected by respondents' absolute *level* of support for the candidate or policy in question.

*Matthew Grahem is a Doctoral Candidate in Political Science, Yale University. Alexander Coppock is Assistant Professor of Political Science, Yale University. We thank Jordan Farenhem for his excellent research assistance. For helpful comments on earlier versions of this paper, we are grateful to seminar participants at Yale and the American Political Science Association.

In advance of Alabama's 2017 special election for U.S. Senator, the polling firm JMC Analytics released a survey which sought to estimate the effect of sexual misconduct allegations on support for Roy Moore's candidacy. The question read, "Given the allegations that have come out about Roy Moore's alleged sexual misconduct against four underage women, are you more or less likely to support him as a result of these allegations?" Among the 575 registered Alabama voters sampled, 29% responded "more likely," 38% responded "less likely," and 33% responded "no difference." Among self-identified evangelical Christians, "more likely" outnumbered "less likely." The poll was widely discussed in the media, with many commentators decrying the apparent depravity of Alabama voters who, according to this poll, *increased* their support for a candidate when allegations of sexual assault against minors surfaced.

Survey questions like this one ask respondents to assess the causal effect of some news event on their attitudes. In this article, we make the argument that this format leads to badly biased inferences that wildly overstate changes in attitudes. The bias stems in part from the nature of the question – just as assessing causality is hard for social scientists, it is also hard for survey subjects. The bias is also due to the question format, which we argue induces subjects to engage in "response substitution" (Gal and Rucker 2011; Yair and Huber 2018), wherein respondents use the question to indicate the level of their opinion rather than the changes in it. In our view, those Alabama voters were not saying they support Roy Moore *more* because he was accused of sexual assault, they were just saying they support him anyway.

To some, it is "obvious" that the poll should not be interpreted to mean that 29% of Alabama voters literally thought better of Moore because of the scandal. Responsible analysts interpret the question to mean that 29% of Alabama voters want to express support for their preferred candidate, despite the scandal. We agree with this interpretation. However, it subtly switches the goal of the research (the estimand) from the average causal effect of

the scandal on support to the average level of support after the scandal. In our view, if researchers are interested in this second estimand, they should choose survey questions that directly measure it. In this article, however, we are squarely focused on the original, causal estimand and we will consider a series of survey techniques for estimating it, including the usual approach (randomized experimentation) and new approaches (survey questions that prompt counterfactual reflection).

Why pick on self-reported change questions? The main reason is that they are too common. With help from a research assistant, we documented nearly 200 instances of self-reported attitude change questions in state or national polls during 2017 and 2018. These questions tended to focus on the effect of candidates' policy positions on candidate support; the effects of endorsements or support for other candidates; social and economic behavior; and influences on one's own attitudes (Table 1). Reflecting the revolution in discourse over sexual misconduct taking place during this time period, the Roy Moore question was accompanied by many others touching on sex and scandal, including 15 questions about the effect of Supreme Court Justice Brett Kavanaugh's confirmation on support for politicians.

We have two main goals in this article. First, we seek to explain why and how the self-reported attitude change question generates biased inferences. Our main explanation is that it hard for subjects to compute counterfactual quantities, so they engage in response substitution. Second, we develop alternative question formats that help subjects make more informed guesses about counterfactuals. These question formats appear to give subjects a "counterfactual assist," exhibiting lower bias—as compared to experimental benchmarks—than the standard self-reported attitude change question format.

Table 1: Self-Reported Attitude Change Questions in Public-Facing Polls

Category	Polls	Qs	Example
Candidate positions	25	68	If your member of Congress voted for the health care bill currently being considered by Congress, would that make you more or less likely to vote for them in the next election, or would it not make a difference either way? — *Public Policy Polling, July 2017*
Endorsements by or support for other people	25	33	If [Claire McCaskill / Heidi Heitkamp / Joe Donnelly] votes against Brett Kavanaugh's nomination to the Supreme Court, would that make you more likely or less likely to vote for [her / him], or would it not make a difference to your vote for Senate? — *Fox News, September 2018.*
Politics and social or economic behavior	13	17	As you may know, some athletes and sports teams have begun not standing during the national anthem in order to protest police violence against the black community in the United States. Does this make you more likely, less likely or has no influence on you to watch NFL games on television? — *University of North Florida, September 2017.*
Attitudes	12	30	If you knew that the Republican tax plan would cause a significant increase in the national debt over the next 10 years, would that make you more likely to support it, less likely to support it, or would it not have an impact? — *Quinnipiac, November 2017*
Misconduct	10	13	Does the issue of sexual harassment make you more likely to vote for a [Democratic / Republican / woman] candidate, or not? — *Quinnipiac, December 2017*
Candidate attributes	9	12	Stacey Abrams has discussed being more than $200,000 in debt. Does Stacey Abrams' debt make you more likely to consider voting for her? Less likely to consider voting for her? Or does it make no difference? — *SurveyUSA, May 2018*
Political participation	6	7	Has what you've seen in Washington over the last year made you more likely to speak up and let your political views be known, less likely to speak up and let your political views be known, or has there been no change? — *CBS, October 2018*

Note: These questions were found using two searches. First, we searched the Roper Center's iPoll database using the search string: more OR less OR make% OR likely OR (change AND your) OR (would AND you AND be) OR (rate AND you%). Second, we searched Google for this same search string plus the word "poll." For both searches, we only considered polls conducted between January 1, 2017 and December 31, 2018.

Theoretical Setting

We begin with the Neyman-Rubin model of causal inference (Neyman 1923; Rubin 1974), which posits that subjects reveal different potential outcomes (PO) in different states of the world. When the relevant states are "control" and "treatment," subject i has two potential outcomes: untreated and treated, which we respectively denote as $Y_i(0)$ and $Y_i(1)$. Subject i's individual-level causal effect, τ_i, is the difference between these two states of the world ($\tau_i \equiv Y_i(1) - Y_i(0)$). This conception raises three possible threats to inference: standard self-report questions (1) ask subjects to report an unknowable quantity (2) in a format that is vulnerable to misreporting and (3) collects no information about τ_i's magnitude.

First, self-reported attitude change questions ask subjects to report on a quantity that is *unknowable* to both the researcher and the subject. This problem has famously been called the Fundamental Problem of Causal Inference (Holland 1986). Because only one state of the world ever realizes, it is possible to observe either $Y_i(0)$ or $Y_i(1)$, but not both. People can of course guess what the unobserved potential outcome would have been, and those guesses may be more or less accurate depending on the domain and the person. We draw accurate causal inferences about out own lives all the time (I broke my leg *because* I slipped), but some prior evidence gives us reason to be skeptical of peoples' causal inferences about their *attitudes*. Lord et al. (1979)'s classic study found that information about the death penalty produced large amounts self-reported attitude change: proponents report becoming more supportive and opponents report becoming more opposed. However, an experimental replication that compared treated subjects to untreated subjects found that both proponents and opponents was small and had the same sign for both groups (Guess and Coppock 2018). At least in this case, self-reports and experiments on attitude change yield different conclusions.

Second, self-reported attitude change questions are subject to *misreporting*. Although the questions ask subjects to focus on their attitude change (τ_i), subjects are often thought

answer the question they want to answer rather than the question before them. Although we cannot be certain about the underlying cognitive processes that produce misleading self-reports, we worry in particular that respondents may use state their absolute level of support for a candidate or issue rather than the change in support due to the stimulus. Gal and Rucker (2011) and Yair and Huber (2018) present evidence that subjects often engage in "response substitution," using questions that are clearly about one aspect of a person or situation to express their overall attitude toward the same. In both of these studies, giving respondents the opportunity to answer "unasked questions" focuses respondents on the intended question.

Third, standard self-reported attitude change questions ask subjects to report the sign of τ_i – did [X] make you more or less supportive of [Y]? – but not its magnitude. This complicates interpretation when analysts try to estimate the average treatment effect. The dominant practice is subtract the the percentage of respondents who said more supportive from the percentage who said less supportive. Even if subjects could faithfully report the sign of their causal effect, this procedure for estimating the average treatment effect is prone to bias. For example, imagine that among the survey respondents, a small number experience large negative effects, but a large number experience small positive effects. The differencing procedure would conclude the sign of the ATE is positive when it could easily be negative, depending on how large is large. Ignoring the magnitude of change can lead to substantial inferential errors. More generally, the fact that the standard question does not estimate τ_i itself, just its sign, means that it is difficult to compare to benchmark experimental estimates of the ATE.

Demonstrating the bias of the self-reported attitude change question is complicated by the very issue that introduces the bias in the first place. The fundamental problem of causal inference means that survey respondents do not know their own causal effects and researchers do not know them either. This makes our setting somewhat different than other validation

studies of self-report measures. For example, findings that show survey subjects overreport media exposure (Vavreck 2007; Guess 2015; Jerit et al. 2016) and voter turnout (Holbrook and Krosnick 2010) measure accuracy by using another measure of the event in question. In our setting, measuring the 'ground truth" of an individual causal effect is impossible. Our evaluation of self-report questions will therefore depend in part on comparisons of substantive takeaways about average causal effects.

Alternative Question Formats

Our goal is not just to knock down the self-reported change format, but to explore some alternatives. We consulted an assortment of survey design textbooks (Babbie 2011; Fowler 1995, 2014; Groves et al. 2009; Torangeau and Rips 2000; Sudman and Bradburn 1982). While all are valuable works, none contained specific guidance for questions that ask subjects to assess causal effects. In this section, we will propose some alternatives that may decrease bias by helping subjects to explicitly consider counterfactuals. To be clear, we will not solve the fundamental problem of causal inference. We hope the alternative formats will begin, not conclude, a conversation about improving measurement techniques for subject beliefs about causal effects.

Our nine question formats can be thought of as falling into four groups:

- The standard format, *sign (no anchor)*, asks subjects whether some piece of information changed their attitudes in a positive direction, a negative direction, or made no difference. In other words, subjects are asked to guess the sign of τ_i.

- *Anchored* formats pair the standard format with another question about one or both PO. In addition to guessing the sign of τ_i, subjects are given an opportunity to state their absolute level of support for a candidate or policy ($Y_i(0)$ or $Y_i(1)$).

- *Both PO* formats abandon the standard format altogether. Instead, subjects are asked to state both of their potential outcomes, $Y_i(0)$ and $Y_i(1)$. From this information, one can calculate the subject's guess of τ_i as well as τ_i's sign.

- The *sign (explicit PO)* format can be used with the same question wording as the standard sign (no anchor) format. However, for a binary outcome variable Y, the response options simultaneously reveal both potential outcomes and τ_i. The response options are variants of "support either way," "oppose either way," "support but would have opposed," and "oppose but would have supported."

The alternative formats are designed to mitigate the three threats to inference we ascribed to the standard format in the previous section. First, every alternative format encourages respondents to introspect about at least one of their potential outcomes, which may reduce error in correctly remembering how one's attitudes changed. Second, every alternative format also gives respondents the opportunity to state one of the alternative potential outcomes, which may discourage response substitution: giving respondents the opportunity to state $Y_i(0)$, $Y_i(1)$, or both allows them to express this attitude and, at the same time, makes it clearer that the change question is different from the level question. Third, the both PO formats allow us to infer τ_i's magnitude rather than just its sign, providing more information than the standard format. The sign (explicit PO) format also allows us to infer τ_i, but does so by eliminating the possibility of measuring attitude change on a fine-grained level.

Table 2 describes each of the nine question formats. For each format, we provide a concrete example based on Study 1's *disputed accusation* mini-experiment. In this mini-experiment, all respondents read a short description of a then-recently-resigned Republican state legislator from Minnesota, Tony Cornish. The control vignette presented only information about his personal background and policy positions, all closely paraphrased from his

website. The treatment vignette used real quotes from local newspaper coverage[1] to add one more piece of information:

> *Cornish has been accused of making inappropriate sexual comments by fellow legislator Erin Quade, a Democrat. Cornish denied the allegations, saying he was "blindsided." Quade admitted having a "cordial and collegial relationship" with Cornish but said that "doesn't excuse sexual harassment."*

Respondents assigned to the standard format were shown the treatment vignette, then asked how Quade's accusation affected their support for Cornish. Each alternative format either added anchors or replaced the standard format with separate questions about both POs. Where appropriate, respondents were explicitly asked to imagine that they did not know that Cornish was accused of sexual misconduct.

[1] Jennifer Bjorhus and J. Patrick Coolican, "Minnesota lawmaker, lobbyist accuse Rep. Tony Cornish of sexual harassment," *Star Tribune*, November 10, 2017.

Table 2: Description of Self-Assessment Question Formats

Format	Sequence	Example
Sign (no anchor)	· Treatment · Report sign(τ_i)	[Treatment vignette] · Does the fact that Cornish was accused of sexual misconduct make you more or less likely to support him in an election against a moderate Democrat?
Sign ($Y_i(0)$ anchor)	· Treatment · Report $Y_i(0)^*$ and sign(τ_i)	[Treatment vignette] · Suppose you **did not know** that Cornish was accused of sexual misconduct. If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him? · Does the fact that Cornish was accused of sexual misconduct make you more or less likely to support him in an election against a moderate Democrat?
Sign ($Y_i(1)$ anchor)	· Treatment · Report $Y_i(1)$ and sign(τ_i)	[Treatment vignette] · If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him? · Does the fact that Cornish was accused of sexual misconduct make you more or less likely to support him in an election against a moderate Democrat?
Sign ($Y_i(0)$ and $Y_i(1)$ anchor)	· Treatment · Report $Y_i(1)$, $Y_i(0)^*$, and sign(τ_i)	[Treatment vignette] · If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him? · Suppose you **did not know** that Cornish was accused of sexual misconduct. If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him? · Does the fact that Cornish was accused of sexual misconduct make you more or less likely to support him in an election against a moderate Democrat?
Sign ($Y_i(0)$ first)	· Report $Y_i(0)$ · New slide · Treatment · Report sign(τ_i)	[Control vignette] · If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him? [Treatment vignette] · If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him? · Does the fact that Cornish was accused of sexual misconduct make you more or less likely to support him in an election against a moderate Democrat?
Both PO ($Y_i(0)$ first)	· Report $Y_i(0)$ · New slide · Treatment · Report $Y_i(1)$	[Control vignette] · If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him? [Treatment vignette] · If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him?
Both PO ($Y_i(1)$ first)	· Treatment · Report $Y_i(1)$ · New slide · Report $Y_i(0)^*$	[Treatment vignette] · If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him? Suppose you had seen the same information, but **without any mention** of the fact that Cornish was accused of sexual misconduct. [Control vignette] · If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him?
Both PO ($Y_i(0)$ and $Y_i(1)$)	· Treatment · Report $Y_i(0)^*$ and $Y_i(1)$	[Treatment vignette] · If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him? · Suppose you **did not know** that Cornish was accused of sexual misconduct. If Cornish were running for Congress in your district against a moderate Democrat, how likely would you be to support him?
Sign (explicit PO)	· Treatment · Report binary τ_i	[Treatment vignette] · How does the fact that Cornish was accused of sexual misconduct change your support for him in an election against a moderate Democrat? (Support either way; oppose either way; oppose but would have supported; support but would have opposed)

Note: $Y_i(0)^*$ denotes questions that ask respondents to imagine not knowing the treatment information.

Research Design

We carried out our empirical research in a series of three surveys, each of which featured a series of mini-experiments. All three surveys were carried out on Lucid, which provides online convienence samples quota sampled to U.S. census demographic margins (Coppock and McClellan 2019).[2] In each survey, subjects were randomized to answer questions in one of the several alternative question formats described above or to participate in a standard experiment with a randomized treatment and control group, which serves as our benchmark. Each survey featured a different mix of mini-experiments and question formats. Across the three surveys, we distributed the topics of our mini-experiments in rough proportion to the topic areas we observed in real-world instances of self-reported attitude change questions (Table 1). We summarize the content of each survey below and in Table 3. The online appendix provides full details of all 18 mini-experiments.

Study 1 featured 1,074 participants who were randomized to participate in a true experiment or one of eight alternative question formats. The first five mini-experiments were conducted in a "candidate profile" format, designed to ensure that all of the information presented was novel to almost all of the subjects. Control subjects read a vignette describing one of five real state legislators. Treatment subjects read the same vignette with one additional piece of information, which became the subject of the self-reported attitude change questions. All of the information in the candidate profile experiments was gathered from campaign websites, voting records, and newspaper articles, with care to keep the phrasing as close to the real-world content as possible. The second three mini-experiments concerned issue attitudes. Control subjects directly stated their attitudes, while treatment subjects stated their attitudes after exposure to information.

Study 2 featured 3,281 participants. The survey included the standard self-report for-

[2]For evidence that Lucid respondents have similar knowledge levels and attitudes as American National Election Study respondents, see the appendices to Graham (2018) and Graham and Svolik (2019).

Table 3: Study Information

Study	N	Date	Scale	Self-Report Formats	Topic Areas
1	1,074	May 2018	7-point	· Direction, no anchor · Direction with anchor (5) · Both PO (3)	· Candidate support (5) · Attitudes (3)
2	3,281	Nov. 2018	6-point	· Direction, no anchor · Direction, explicit PO · Both PO (2)	· Candidate support (3) · Attitudes (3) · Celebrities (4)
3	1,110	Dec. 2018	Binary	· Direction, no anchor · Direction, explicit PO · Both PO (2)	· Candidate support (1) · Issue positions (3)

Note: Each study was conducted on Lucid.

mat, two *both PO* formats, and introduced the *sign (explicit PO)* format. Each subject then participated in a series of five mini-experiments. Each experiment followed the same format: control subjects answered questions without any information while treatment subjects were supplied a piece of information before they answered. For one mini-experiment, subjects were only eligible for the "supported" or "opposed" condition, depending on the position their Senator took on Brett Kavanaugh's confirmation to the Supreme Court (see below for details). This brought the total number of mini-experiments in Study 2 to six.

Study 3 featured 1,110 participants. It replicated four of Study 2's six mini-experiments using binary questions rather than Likert scales, enabling us to put the *sign (explicit PO)* format on the same scale as an experiment.

Our analysis uses three modes of inference to assess self-reported attitude change questions. First, we will examine the percentage of respondents who report attitude change under each format, regardless of the sign (direction) of the reported change. Second, for a subset of our alternative formats, it is possible to express the self-reported treatment effects in the same units as the experiment. For these formats, we will directly compare our alternative formats to the potential outcomes and average treatment effect estimates obtained from a

true experiment. Third because the standard self-report format and some of our alternatives cannot be directly compared to a standard experiment, we will compare our experiments to the conclusions that would result from interpreting the standard format as a trustworthy measure of attitude change.

For all of our analysis, we use OLS regression to estimate means and differences in means.[3] We always report robust standard errors; for all analysis that pool across multiple mini-experiments, we cluster our errors by respondent. Because each of our studies used different numbers of scale points, we standardize all experimental outcome measures by dividing by the standard deviation in the control group. Consequently, all experimental treatment effects are expressed in terms of the number of standard units by which the outcome changed due to treatment.

Results

We report three sets of results. First, we describe how the distributions of the "direction" questions change depending on format. Second, we assess the accuracy of the "both potential outcomes" formats relative to an experimental benchmark. Third, we compare the substantive takeaways from the experiments to the conclusions that would result from over-interpreting the "direction" questions.

Results I: Reducing Self-Reported Attitude Change

We begin by describing the distributions of self-reported attitude change by topic and question format. Figure 1 plots the percentage of respondents in each condition who reported that the information made them more supportive of the candidate or policy, less supportive,

[3]While some analysts prefer to analyze binary outcome variables with generalized linear models such as logit or probit, this is unnecessary. The difference-in-means is unbiased for the average treatment effect regardless of the outcome space (Gerber and Green 2012, chapter 2). As is common, none of our substantive interpretations depend on this choice.

or caused no change. In all 18 mini-experiments, the share of subjects who report attitude change (i.e., do not say "no change") tends to be largest in the *sign (no anchor)* format and successively smaller in the other formats. The most notable exception to this generalization are Republican self-reports of attitude change in the "disputed accusation" mini-experiment, who seem more willing to admit that they dislike sexual harassment when they can also voice their continued support for their co-partisan.

Pooling across the mini-experiments, Figure 2 presents a series formal tests of the claim that the alternative formats reduce self-reports of attitude change. For each question format, Figure 2 shows the the percentage of respondents reporting that their attitudes changed as a result of the treatment information. In each panel, the topmost estimate and confidence interval represent the mean percentage of respondents who report attitude change in the standard *sign (no anchor)* format.

Compared with the both PO formats, the anchoring formats resulted in relatively modest decreases in the percentage of respondents self-reporting attitude change. In Study 1, the average decrease is about 4 percentage points in the two anchored direction formats that include only $Y_i(0)$ and about 12 percentage points in the two direction formats that include $Y_i(1)$. By comparison, the *both PO* formats reduce self-reported change by an even larger magnitude, by about 23 percentage points in Study 1, about 12 to 20 percentage points in Study 2, and about 35 percentage points in Study 3. Although each study was conducted on a separate sample, we suspect that Study 3's relatively large reduction in self-reported attitude change is attributable to its use of binary outcome variables.

We interpret these results as initial, suggestive evidence that respondents in the standard *sign (no anchor)* format over-report attitude change. To sharpen this inference, our next two sets of results will compare the self-reported change questions to randomized experiments.

Figure 1: Distribution of Self-Reported Change

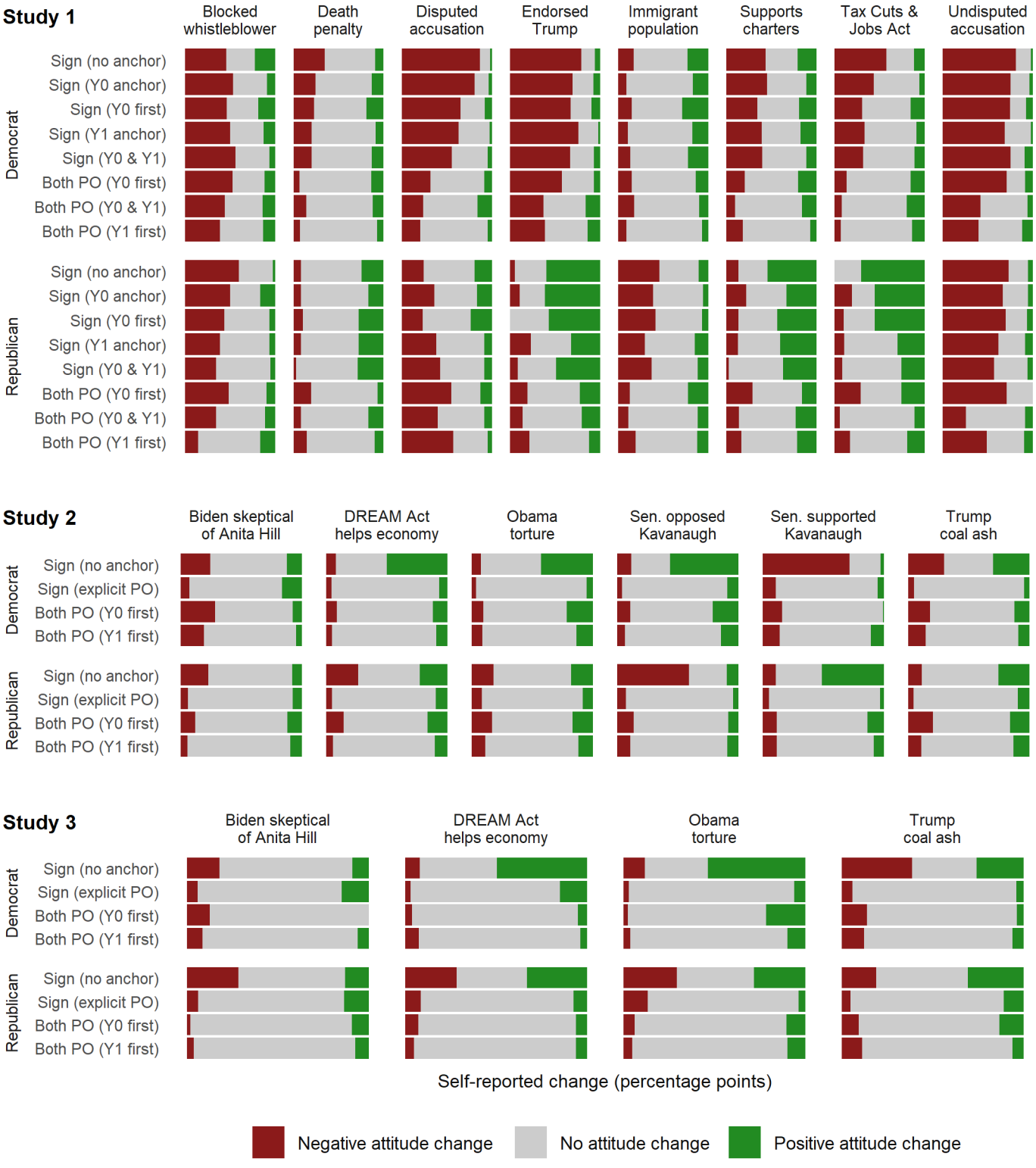
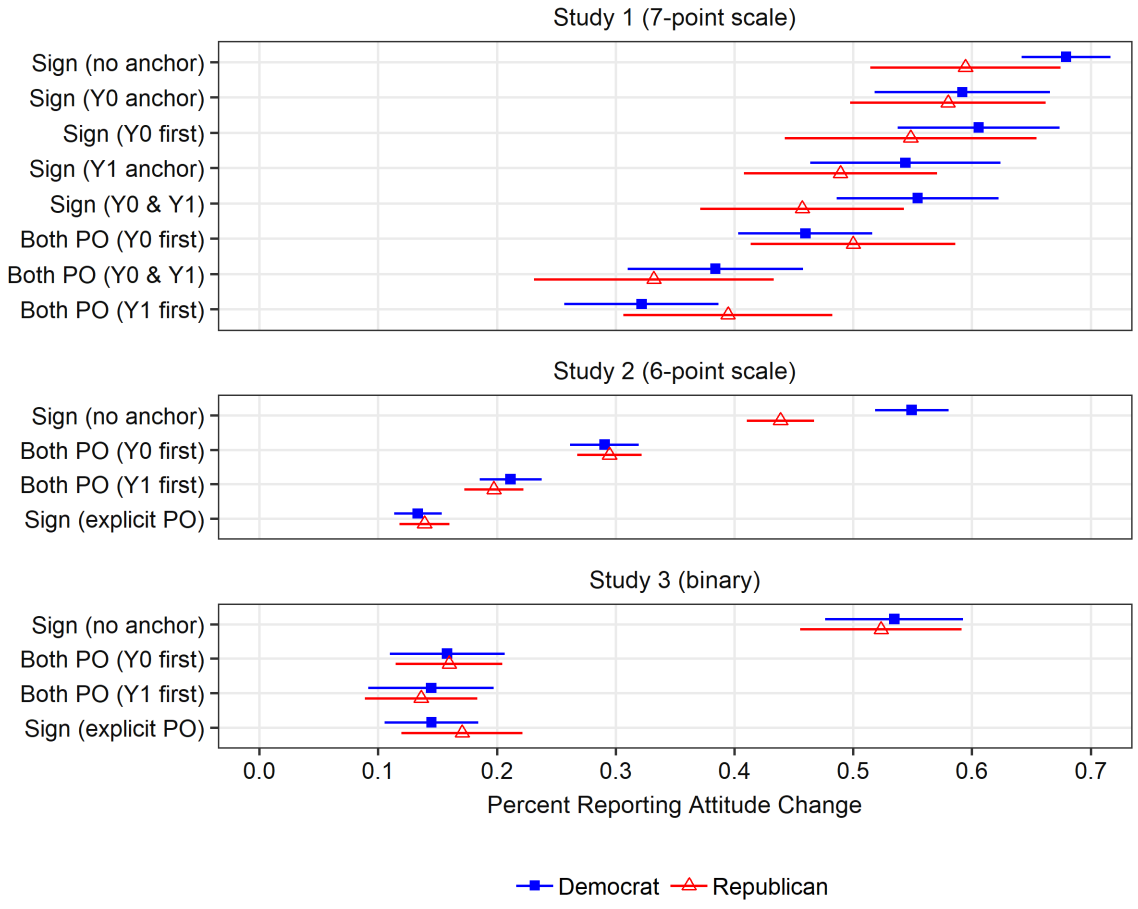


Figure 2: Percent Reporting Attitude Change by Study and Question Format



Results II: Self-Reported Treatment Effects

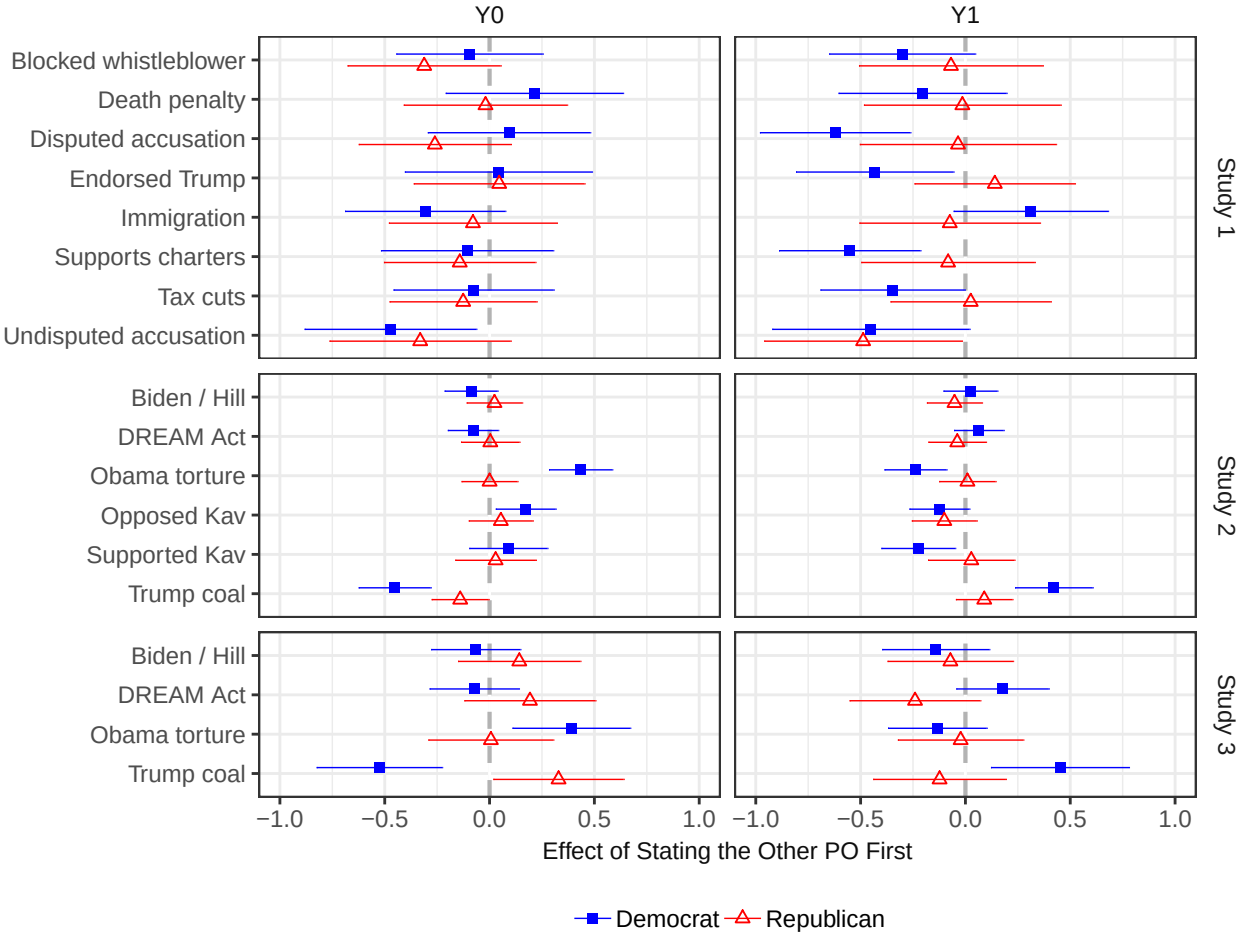
Although the standard *sign (no anchor)* format cannot be placed on the same scale as an experiment, our alternative formats that measure both POs can be. We argue above that this gives our alternative formats two measurement advantages over the standard format: they define what counts as "attitude change" and produce outcome measures that can be directly compared to an experiment. In this section, we take advantage of these properties to provide some evidence as to the accuracy of the both PO formats.

Can respondents guess their missing potential outcome?

For an initial look at the accuracy of self reports, we focus on the two alternative formats that were present in all three studies: both PO (Y0 first) and both PO (Y1 first). In these formats, respondents first stated their control or treatment outcome, and were then immediately switched to the other condition. In both PO (Y0 first), respondents stated their control outcome, then saw the treatment information and answered the question again. In both PO (Y1 first), respondents stated their treatment outcome, then were asked to pretend they did not know the treatment information and state what their control outcome would have been (see Table 2). These formats allow us to focus on the question: *when people have just reported one potential outcome, are they able to accurately guess their missing potential outcome?*

To answer this question, Figure 3 presents the average difference between the "true" measure of respondents' $Y_i(0)$ and $Y_i(1)$ and respondents' guesses about these outcomes after they have just stated the other. Estimates near zero indicate either that respondents accurately reported their missing potential outcome on average, or that we cannot detect a difference between the self-reports and the experiment. Estimates that are statistically distinguishable from zero indicate that, on average, respondents who are considering how

Figure 3: Misreporting the Missing Outcome (both PO formats)



their attitudes changed self-report a different potential outcome than they would have in an experiment.

The results suggest that when respondents are asked to report their missing potential outcome, they often systematically misreport it. Out of the 72 coefficients displayed in Figure 3—each representing a different combination of party identification (Democrat or Republican), potential outcome ($Y_i(0)$ or $Y_i(1)$), and mini-experiment (18 total)—we reject the null hypothesis of no average difference in 17 cases (23%) at the $p < 0.05$ level and in 9 cases (13%) at the $p < 0.01$ level.

These patterns of misreporting are consistent with the idea that respondents have

trouble separating their self-report of the missing outcome from the outcome they had just reported. In other words, the actual state of the world appears to cloud respondents' introspection about the counterfactual state of the world. Out of the 17 cases in which we can detect misreporting at the $p < 0.05$ level, there are 13 cases (77%) in which the misreport falls on the same side of the scale as the potential outcome the respondent just reported. This includes eight of the nine misleading cases (89%) in which respondents had already reported their $Y_i(1)$ and are then asked to state what $Y_i(0)$ would have been if they did not know the treatment information.[4]

For a concrete example of this pattern of misreporting, consider Democrats' self-reports in the *Trump coal* mini-experiment. In the both PO (Y1 first) format, Democrats learn that President Trump issued an executive order making it easier to dump coal ash in streams, then report $Y_i(1)$: their support for the order in the state of the world in which they know that Trump issued the order. On the next screen, they are asked to report $Y_i(0)^*$: what their support would have been if they did not know that Trump had issued the order. These respondents systematically report lower support than respondents who do not already know that Trump issued the order (Figure 3).

Although we cannot be certain about why respondents tend to misreport in this direction, it is consistent with the response substitution story. Our analysis below provides further support for this interpretation. In fact, the pathologies we just described appear to be worse under the standard format.

[4]For this classification, we placed all outcome measures on a 0-1 scale. If respondents reported a larger $Y_i(0)^*$ than the true $Y_i(0)$, and $Y_i(0)$ was greater than 0.5—or if respondents reported a smaller $Y_i(0)^*$ than the true $Y_i(0)$ and $Y_i(0)$ was less than 0.5—we considered the misreporting to be consistent with response substitution. We applied the equivalent rule for misreporting of $Y_i(1)$.

Can respondents accurately self-report their treatment effect?

Another way to directly compare the *both PO* formats to experiments is by examining self-reported average treatment effects. For all 22 mini-experiments, Figure 4 plots our estimate of the treatment effect from the true experiment (black diamonds) and the average self-reported treatment effect from each alternative format that features a measure of both $Y_i(0)$ and $Y_i(1)$ (grey squares, circles, and triangles).

Although the same number of respondents were randomized into each arm of our study, our estimates of the self-reported treatment effects are always more precise than the in the benchmark experiment. This is because in the benchmark experiment, we use only one piece of information per respondent, but in the alternative formats, we use two pieces of information per respondent. This highlights a tantalizing property of the self-reports: if a good self-report measure were obtained, it would be possible to estimate effects far more precisely than is possible in a randomized experiment. In this sense, the choice between the randomized experiment and our alternative formats amounts to a bias-variance tradeoff.

Unfortunately, in our judgment, the self-reports are not always accurate enough for researchers to take advantage of their greater precision. In Figure 4, the experimental treatment effects often appear substantially different than the self-reported treatment effects, including a handful of cases in which the self reports appear to get the sign wrong. The fact that experiments combine information about $Y_i(0)$ and $Y_i(1)$ make speculation about the reasons for these errors more difficult here than elsewhere in our analysis; we therefore refrain from that exercise in this subsection.

For a systematic look at the accuracy of self-reported treatment effects, Table 4 displays the mean squared error of each self-report measure's treatment effect estimate, pooling across surveys.[5] In most cases, the differences between our alternative formats are too small to be

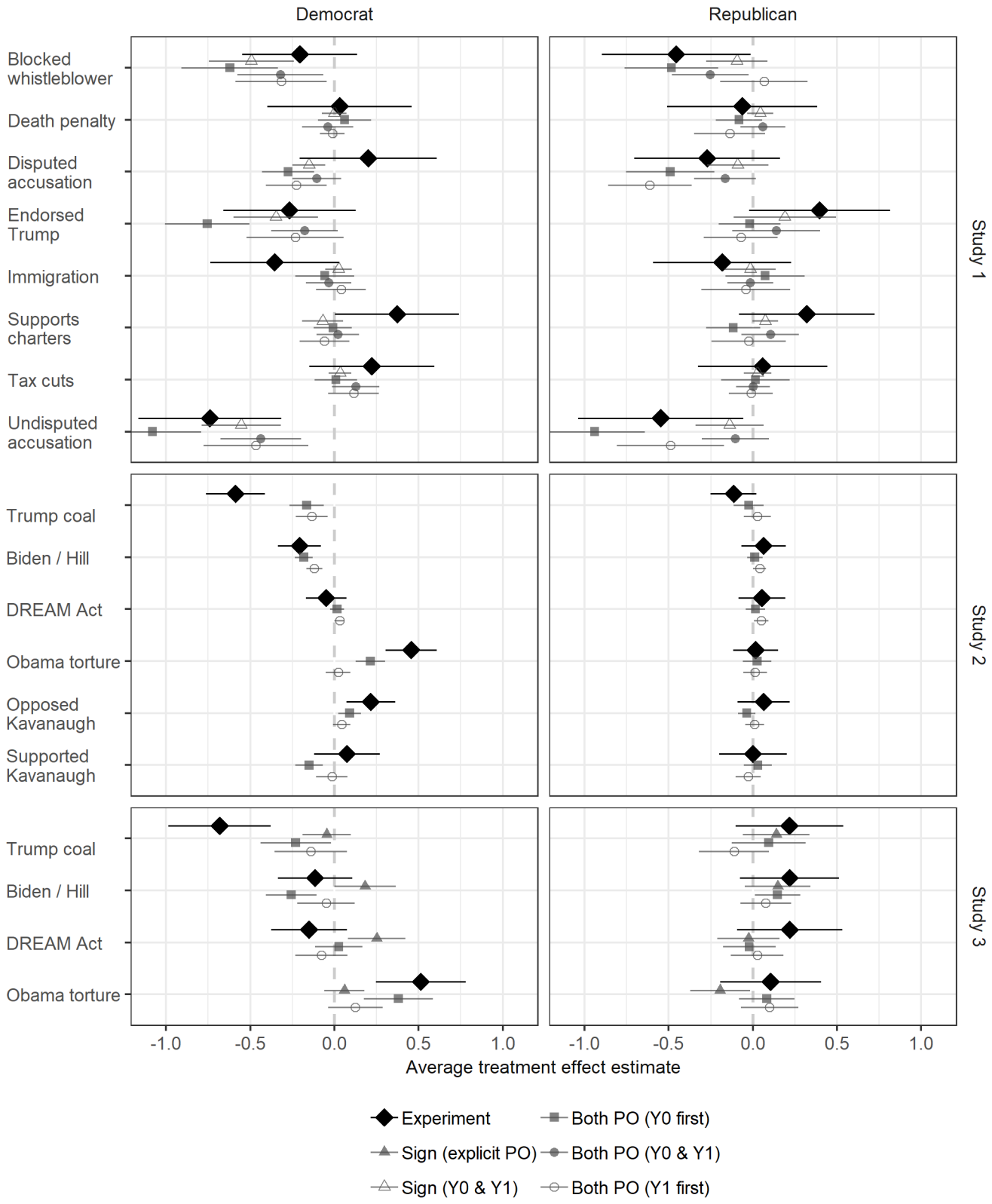
[5]We computed RMSE according to the formula $\sum_m (\hat{\tau}_{\text{EXPER}} - \hat{\tau}_{\text{SELF}})^2$, where m indexes mini-experiments and $\hat{\tau}$ is a treatment effect estimate.

distinguished from one another statistically. However, we can make a few observations that may be useful to future efforts to elicit accurate self reports. First, asking respondents to consider both PO at the same time may be a better way to encourage accuracy than asking the two PO in succession. In Study 1, the both PO format that asked $Y_i(0)$ and $Y_i(1)$ at the same time elicited the most accurate self reports. Second, when both PO are asked in succession, it may not make make much difference whether $Y_i(0)$ or $Y_i(1)$ was elicited first. Pooling across all three studies, Republicans provided slightly better guesses under the both PO (Y0 first) format while Democrats did slightly better on both PO (Y1 first) Third, explicit encouragement to think about attitude change may make self-reports less accurate. In Study 3, the sign (explicit PO) format performed worse than the both PO formats, attaining the highest RMSE and yielding the least accurate point estimate in six of eight party-mini-experiment combinations (Figure 4).

Table 4: Root Mean Squared Error of Self-Reported ATE Estimates

Study	Format	Democrat	Republican
1	Both PO (Y0 & Y1)	0.67	0.64
1	Both PO (Y0 first)	1.02	0.80
1	Both PO (Y1 first)	0.79	0.87
1	Sign (Y0 & Y1)	0.79	0.69
2	Both PO (Y0 first)	0.79	0.78
2	Both PO (Y1 first)	0.73	1.02
3	Both PO (Y0 first)	0.52	0.28
3	Both PO (Y1 first)	0.67	0.41
3	Sign (explicit PO)	0.93	0.40
All	Both PO (Y0 first)	1.39	1.15
All	Both PO (Y1 first)	1.27	1.40

Figure 4: Experiment versus Self Reports (two-PO formats only)



Results III: Changing the Story

As noted above, direct comparisons between experiments and self-reports are limited by the intersection of (1) the fundamental problem of causal inference and (2) the fact that standard self-report questions, which only ask τ_i's sign, cannot be placed on the same scale as an average treatment effect. Above, our analysis dealt with this issue by refraining from comparing questions that were measured using different scales. In exchange for sound empirical practice, this forced us to avoid direct comparisons between the standard sign (no anchor) format and the experiment. To fill this gap in our analysis, this section compares the substantive conclusions one would take away from self-reported change questions to the substantive conclusions suggested by randomized experiments.

To aid our comparisons, Figures 5 and 6 plot, for Democrats and Republicans in all 18 mini-experiments, the average treatment effect estimate and the distribution of the sign (no anchor) format. We describe each mini-experiment as we display its results; for full details of each mini-experiment, see the online appendix.

Study 1

Of the eight mini-experiments in Study 1, four produced substantively misleading self reports, three produced self reports that appeared entirely consistent with the experiments, and one was a borderline case. The four misleading cases are:

- **Disputed accusation.**[6] Much like the Roy Moore example in the introduction, the self reports paint a picture in which Democrats are repelled by the allegations and Republicans largely stick by their candidate (Figure 1, row 3). Among Democrats, 87% said the accusations made them less likely to support Cornish, compared with just 2% more likely. Among Republicans, 24% said less likely and 19% said more likely. By comparison, the experimental point estimate is positive among Democrats and negative among Republicans, with an effect on the difference in attitudes that comes close to statistical significance in the opposite direction as that suggested by the self reports ($\hat{\bar{\tau}}_D - \hat{\bar{\tau}}_R = 0.47$, $p = 0.11$). This is inconsistent with the self reports'

[6] For a more detailed explanation of this mini-experiment, see Table 2 and the surrounding text.

implication that Republicans stuck by their candidate and Democrats' support fell off dramatically.

- **Immigrant population.** Treatment respondents were informed, with an accompanying graphic, that the Census Bureau projects that the percentage of U.S. residents who are immigrants will reach an all-time record by 2020. Republicans reported becoming less supportive of immigration as a consequence of the information, which is consistent with their experimental point estimate. Among Democrats, enough respondents reported that the information made them more supportive of immigration that, on net, over-interpretation of self-reports would suggest no attitude change (Figure 5, row 5). To the contrary, Democrats had a negative, borderline statistically significant treatment effect in the experiment ($\hat{\bar{\tau}}_D = -0.35$, $p = 0.07$).

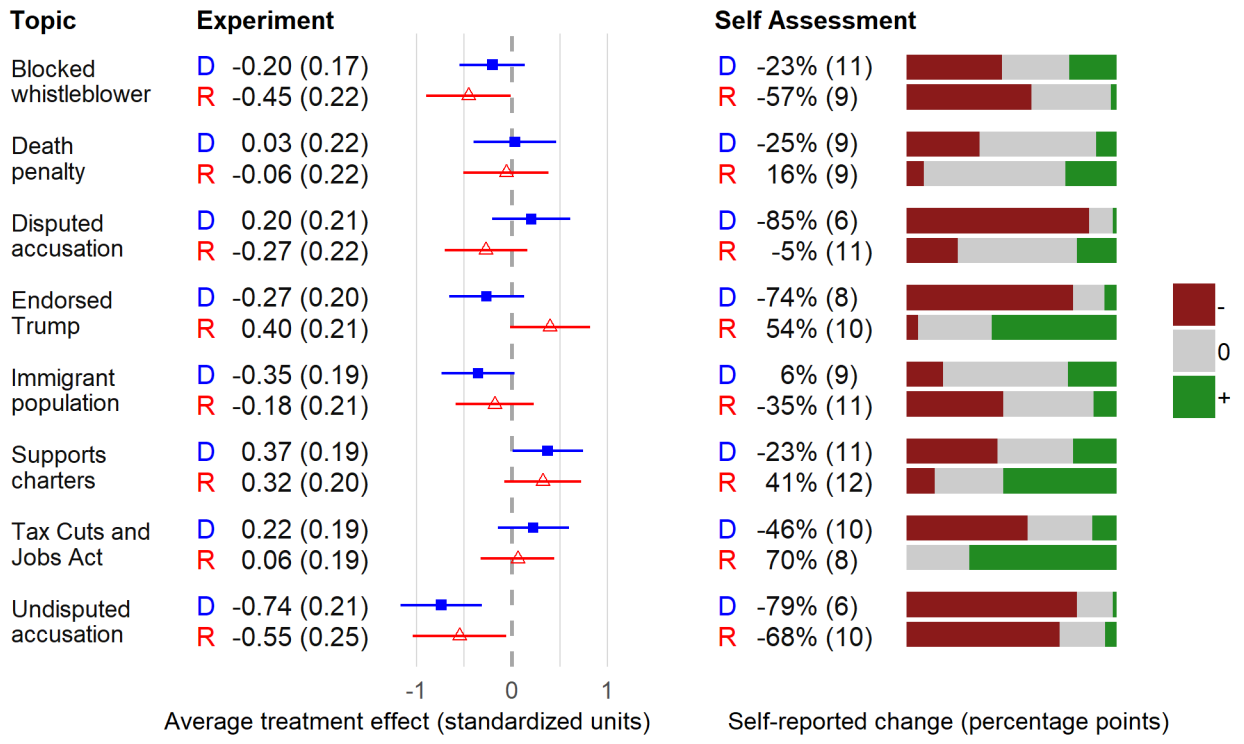
- **Supports charters.** In the experiment, learning that a Republican candidate supports charter schools increases Democrats' support for that candidate ($\hat{\bar{\tau}}_D = 0.37$, $p < .05$), but on the standard self-report measure Democrats are much more likely to report that the Republican candidate's positions make them less supportive. The misleading pattern from the standard format is weakened in the anchored formats and eliminated in the both PO formats. The approximately equal number of Democrats and Republicans reporting positive and negative change is more consistent with the experiment (Figure 5, row 6).

- **Tax Cuts and Jobs Act.** The parties appear to move apart in the standard format, with Republicans overwhelmingly reporting that the Act's contents make them more supportive of it and Democrats reporting the opposite. Moving down the rows in Figure 1, the anchored direction formats reduce self-reports of attitude change, especially when $Y_i(1)$ is anchored. The both PO formats impliy that about the same small shares of people in both parties people became more and less supportive of the Act after learning about it. These patterns are more consistent with the experiment (Figure 5, row 7).

A fifth mini-experiment in Study 1, *death penalty*, was an edge case for misleading self reports. In this mini-experiment, respondents saw text and a graphic depicting the fact that states with and without the death penalty have similar murder rate trends over time. The self reports imply some movement apart by Democrats and Republicans that is not evident in the experiment, but because the self-reports are not overwhelming, we did not feel that the close-to-zero experimental point estimates were plainly inconsistent with the self-reports.

Three of the other three mini-experiments in Study 1 yielded no obvious inconsistencies between self reports and the experiments. Both parties self-reported less support for a

Figure 5: Experiment versus Standard Self Assessment: Study 1

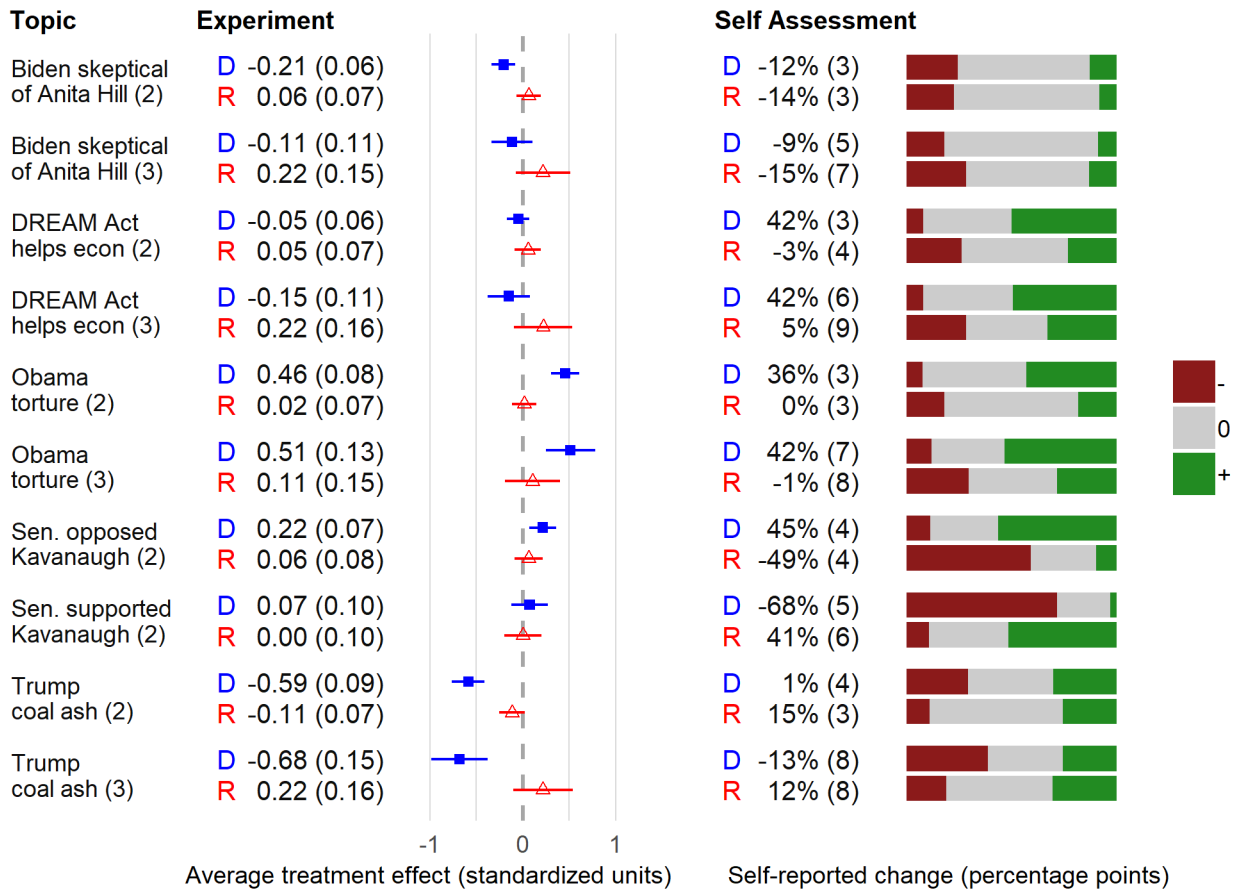


Democrat who blocked whistleblower protections for sexual assault accusers; the experimental point estimates are also negative and one is statistically significant (Figure 5, row 1). A pro-environment Republican's endorsement of President Trump polarized both Democrats and Republicans (Figure 5, row 4). For the case of a Democrat whose seven sexual harrassment accusers were backed by an even more sordid incident in his past, the experiment and self reports both implied large negative effects (Figure 5, row 8).

Studies 2 and 3

Studies 2 and 3 featured ten total mini-experiments, but with overlap between the topic areas. For our analysis of the substantive takeaways, we collapse our generalizations into five topic areas: Biden skeptical of Anita Hill (Studies 2 and 3), DREAM Act helps economy (Studies 2 and 3), Obama executive order on torture (Studies 2 and 3), Senator's position on

Figure 6: Experiment versus Standard Self Assessment: Studies 2 and 3



Brett Kavanaugh's confirmation (two mini-experiments in Study 2), Trump executive order on Coal Ash (Studies 2 and 3). Each produced misleading self reports. In three cases, the misleading self reports are evident from the comparsion to the experiment:

- **Biden skeptical of Anita Hill.**[7] In the *Biden skeptical of Anita Hill* condition, respondents were asked how then-Senate Foreign Relations Committee chair Joe Biden's skepticism of Anita Hill's accusations against Clarence Thomas affected their support for Biden's widely-rumored 2020 presidential bid. We first conducted this mini-experiment in Study 2 and replicated it with binary support measures in Study 3. In both studies, Democrats' and Republicans' self-reports imply highly similar reactions (Figure 6, rows 1-2). Yet in the experiments, Democrats have a negative treatment effect (pooled $\hat{\bar{\tau}}_D = -0.19$, s.e.$= 0.06$, $p = 0.01$) and Democrats and Republicans diverge

[7]This mini-experiment is similar to an Emerson Polling question asked on March 28-30, shortly after a Nevada legislator publicly accused Biden of inappropriate touching. See Emerson Polling, "Nevada 2020: Biden and Sanders lead Democratic Primary Field," March 31, 2019.

(pooled $\hat{\bar{\tau}}_D - \hat{\bar{\tau}}_R = 0.30$, s.e.= 0.09, $p < 0.01$).

- **DREAM Act helps economy.** According to the self-reports, learning that "The Center for American Progress estimates that passing the DREAM Act would cause the U.S. economy to grow by an additional $30,000 per beneficiary" caused about 50 percent of Democrats to become more supportive of the DREAM Act, compared with only a handful that became less supportive. Republicans seemed about evenly divided, leading to a wider gap between the two parties (Figure 6, rows 3-4). The experiments suggest the opposite conclusion: if anything, the information moved Democrats' and Republicans' attitudes closer together (pooled $\hat{\bar{\tau}}_D - \hat{\bar{\tau}}_R = -0.17$, s.e.= 0.09, $p = 0.06$).

- **Senator's position on Kavanaugh.**[8] For the two Kavanaugh mini-experiments, we used each respondent's state to identify the next Senator who was up for re-election and voted along party lines on Brett Kavanaugh's confirmation to the Supreme Court, then personalized the question to include the Senator's name, gender, party, vote, and re-election date. Consequently, respondents' state of residence determined their eligibility for either the *Senator supported Kavanaugh* or *Senator opposed Kavanaugh* conditions.

  The self-reports to both Kavanaugh mini-experiments imply massive polarization in respondents' support for their incumbent Senators (Figure 6, rows 7-8), but Democrats' and Republicans' attitudes do not appear to diverge in the experiment. The experiments confirm Democrats' greater self-reported support for Senators who opposed Kavanaugh, but in the other three cases (Democrats whose senators supported Kavanaugh, Republicans in both support and oppose), overwhelming self reports of change change are paired with experimental point estimates that have the opposite sign or fall right at zero.

For the two remaining topics in Studies 2 and 3, an executive order by President Obama against the use of torture and an executive order by President Trump on coal ash regulations, we do not observe any clear inconsistencies between the experiments and the self-reports (Figure 6, rows 5-6, 9-10). However, in the previous section, we found evidence consistent with response substitution: Democrats, and to a lesser extent Republicans, asked to consider their missing potential outcome appeared to be influenced by the potential outcome they had just reported. These misleading patterns in the respondents' consideration of the separate potential outcomes happened to cancel out in a way that makes the experiment

[8]This mini-experiment is similar to six questions asked by Fox News around the time of Kavanaugh's confirmation hearings. See Table 1.

look consistent with the self reports. This reiterates the superior measurement properties of our alternative formats: self report formats that collect more information enable a more complete vetting of whether the self reports are accurate.

Discussion

> *White House adviser David Axelrod questioned the USA TODAY survey's methodology, saying those who report being more sympathetic to the protesters now were likely to have been on that side from the start. "There is a media fetish about these things," Axelrod said of the protests, "but I don't think this has changed much" when it comes to public opinion.*
>
> — *USA Today*, April 9, 2009

After angry protestors descended on legislators' town hall meetings during the health reform debate that produced the Affordable Care Act, self-reported attitude change questions helped feed a narrative that the protests had galvanized public opinion against reform. David Axelrod's pushback against this interpretation captures our own misgivings about survey questions that ask respondents to assess the causal effects of political events on their attitudes, opinions, and beliefs. It is of course possible that learning about protests caused respondents to become more supportive of the protesters' cause—public opinion change is after all a main goal of many protests. But standard survey questions that ask subjects whether and in which direction their opinion changed are a very poor way to measure such effects. Axelrod's explanation for the bias is similar to ours: respondents often report the level of their opinion rather than the change in it.

The first goal of our paper was to conduct the first formal test of the widespread suspicion that this standard format is badly biased. We believe we have conclusively affirmed this hypothesis. Relative to experimental benchmarks (and our read of the alternative formats), the standard question wildly overstates attitude change. We believe that we have unpacked

the main reason that the standard question is biased, response substitution. Instead of reporting their best guess of the causal effect of the event on their attitudes, subjects appear to substitute in their absolute level of opinion. If true, this means that respondents misreport levels as changes, which can lead to major inferential errors on the part of survey analysts. This question format should be abandoned altogether, both in polls for public consumption and in academic survey research. It systematically misleads readers and, as in the case of the Roy Moore example with which we opened the paper, gives them an unwarranted opportunity to vilify their political opponents.

Our second goal was to propose alternative survey question formats that would be less biased. We believe that our alternatives strictly dominate the standard question, but we acknowledge that they remain biased compared with unbiased experimental benchmarks. Our main approach was to help survey respondents think about changes by helping them to think counterfactualy. We want respondents to consider what their response *would have been* if they did not know the treatment information. It is still hard to know how one would have responded in that alternative world, but at least the respondents are induced to consider the correct attitude when forming a response. We tried a series of formats that differ in their approach to inducing counterfactual thinking; all of them help, but we do not have enough evidence to declare a winner.

We think a promising format is the "both POs" format, which combines a standard experiment with a counterfactual question. Essentially, subjects assigned to the control condition report their untreated outcome first and subjects assigned to the treatment condition report their treated outcome first. In a second question, all subjects report their guess of the other potential outcome. A comparison of the answers, by treatment group to the first question represents an unbiased estimate of the treatment effect; the counterfactual guesses are a bonus that survey researchers can use or not use depending on whether the diagnostic checks we suggest are successful.

Assessing the causal effects of political events on political attitudes is a major goal of pollsters and academics alike. Politicians anticipate the causal effect of their activities on support for their policies and candidacies. Shoddy survey research that asks directly asks respondents whether their support increased or decreased produces misleading inferences. We believe that our alternative formats represent a step in the right direction and hope that future research will improve upon them.

References

Babbie, Earl R. 2011. *The Practice of Social Research*. SAGE Publications.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics* 119(1):249–275.

Coppock, Alexander and Oliver A. McClellan. 2019. "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents." *Research & Politics* 6(1):1–14.

Fowler, Floyd J. 1995. *Improving Survey Questions*.

Fowler, Floyd J. 2014. *Survey Research Methods*. 5 ed. SAGE Publications.

Gal, David and Derek D Rucker. 2011. "Answering the Unasked Question: Response Substitution in Consumer Surveys." *Journal of Marketing Research (JMR)* 48(1):185–195.

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

Graham, Matthew H. 2018. "Self-Awareness of Political Knowledge." *Political Behavior* .

Graham, Matthew H and Milan W Svolik. 2019. "Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States." *Unpublished Manuscript* .

Groves, Robert M., Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Elanor Singer and Robert Torangeau. 2009. *Survey Methodology*. Wiley.

Guess, Andrew and Alexander Coppock. 2018. "Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments." *British Journal of Political Science* forthcoming.

Guess, Andrew M. 2015. "Measure for measure: An experimental test of online political media exposure." *Political Analysis* 23(1):59–75.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2015. "Causal Inference in Conjoint Analysis: Understanding Multi-Dimensional Choices via Stated Preference Experiments." *Poltical Analysis* 22(1):1–30.

Holbrook, Allyson L. and Jon A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique." *Public Opinion Quarterly* 74(1):37–67.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Jerit, Jennifer, Jason Barabas, William Pollock, Susan Banducci, Daniel Stevens and Martijn Schoonvelde. 2016. "Manipulated vs. Measured: Using an Experimental Benchmark to Investigate the Performance of Self-Reported Media Exposure." *Communication Methods and Measures* 10(2-3):99–114.

Lord, Charles G., Lee Ross and Mark R. Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* 37(11):2098–2109.

Neyman, Jerzey. 1923. "On the application of probability theory to agricultural experiments." *Annals of Agricultural Sciences* 10:1–51.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66(5):688–701.

Sudman, Seymour and Norman M. Bradburn. 1982. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey-Bass.

Torangeau, Robert and Lance J. Rips. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Vavreck, Lynn. 2007. "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports." *Quarterly Journal of Political Science* (4):325–343.

Yair, Omer and Gregory A Huber. 2018. "How robust is evidence of perceptual partisan bias in survey responses? A new approach for studying expressive responding." *Unpublished Manuscript* .